

# Re-Identification with Consistent Attentive Siamese Networks

Meng Zheng<sup>1</sup>, Srikrishna Karanam<sup>2</sup>, Ziyang Wu<sup>2</sup>, and Richard J. Radke<sup>1</sup>

<sup>1</sup>Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy NY

<sup>2</sup>Siemens Corporate Technology, Princeton NJ

zhengm3@rpi.edu, {first.last}@siemens.com, rjradke@ecse.rpi.edu

## Abstract

We propose a new deep architecture for person re-identification (re-id). While re-id has seen much recent progress, spatial localization and view-invariant representation learning for robust cross-view matching remain key, unsolved problems. We address these questions by means of a new attention-driven Siamese learning architecture, called the Consistent Attentive Siamese Network. Our key innovations compared to existing, competing methods include (a) a flexible framework design that produces attention with only identity labels as supervision, (b) explicit mechanisms to enforce attention consistency among images of the same person, and (c) a new Siamese framework that integrates attention and attention consistency, producing principled supervisory signals as well as the first mechanism that can explain the reasoning behind the Siamese framework's predictions. We conduct extensive evaluations on the CUHK03-NP, DukeMTMC-ReID, and Market-1501 datasets and report competitive performance.

## 1. Introduction

Given an image or a set of images of a person of interest in a “probe” camera view, person re-identification (re-id) attempts to retrieve this person of interest among a set of “gallery” candidates in another camera view. Due to its broad appeal in several video analytics applications such as surveillance, re-id has seen explosive growth in the computer vision community [16, 44, 45].

While we have seen tremendous progress in re-id [4, 5, 29, 32, 33, 35, 37, 38], there are several problems that still hinder the reliable, real-world use of person re-id. Probe and gallery camera views in real-world applications typically have large viewpoint variations, causing substantial view misalignment between probe and gallery images of the same person. Illumination differences between the locations where the cameras are installed, as well as occlusions in the captured data, add to re-id's challenges. Ideally, we want a method that can reliably spatially localize the

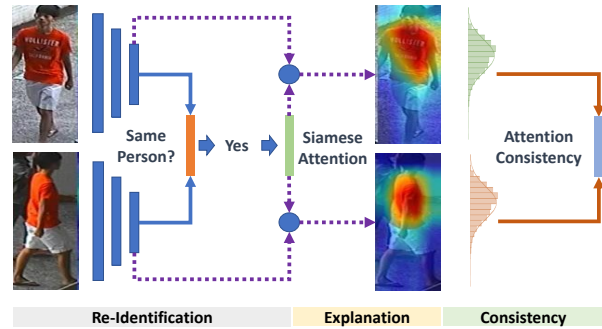


Figure 1: We present the first framework for re-id that provides mechanisms to make attention and attention consistency end-to-end trainable in a Siamese learning architecture, resulting in a technique for robust cross-view matching as well as explaining the reasoning for why the model predicts that the two images belong to the same person.

person of interest in the image, while also providing a robust representation of the localized part in order to match accurately to the gallery of candidates. This suggests we consider the spatial localization and feature representation problems jointly and formulate the learning objective in a way that can facilitate end-to-end learning.

Attention is a powerful concept for understanding and interpreting neural network decisions [10, 26, 28, 49], providing ways to generate attentive regions given image-level labels and trained models, and to perform spatial localization. Unlike its use as a weight matrix in some existing work [2, 3, 36], here we refer to attention computed by means of class-specific gradient backpropagation [28, 49]. Some recent extensions [19] take this a step forward by training models with attention providing end-to-end supervision, resulting in improved spatial localization. These methods were not designed for the re-id problem and consequently did not have to consider localization and invariant representation learning jointly. While there have been some attempts at joint learning with these two objectives [18, 22, 25, 39], these methods do not explicitly enforce any sort of attention consistency between images of the same person. Intuitively, given same-person images from different views,

there typically exist some common regions that are important for matching, which should be reflected in how attention is modeled and used for supervision.

Furthermore, such attention consistency should lead to consistent feature representations for the two different images, leading to invariant representations for robust cross-view matching. These considerations naturally suggest the design of a Siamese framework that jointly learns consistent attention regions for images of the same person while also producing robust, invariant feature representations. While one recent paper approached these problems jointly [39], this method requires specially-designed architectures for attention modeling and considers the attention in each image independently, ignoring the intuition that attentive regions across images of the same person have to be consistent. It also does not have an explicit mechanism to explain the reasoning behind the model’s prediction. To this end, we design and propose a new deep architecture for re-id, which we call the Consistent Attentive Siamese Network (CASN), addressing all the key questions and considerations discussed above (Figure 1). Specifically, we design a novel two-branch architecture that (a) produces attentive regions during training without requiring any additional supervision other than identity labels or any specially-designed architecture for modeling attention, (b) explicitly enforces these attentive regions to be consistent for the same person, (c) uses attention and attention consistency as an explicit and principled part of the learning process, and (d) learns to produce robust representations for cross-view matching.

To summarize, our key contributions include:

- We present a technique that makes spatial localization of the person of interest a principled part of the learning process, providing supervision only by means of person identity labels. This makes spatial localization end-to-end trainable and automatically discovers complete attentive regions.
- We present a new scheme that enforces attention consistency as part of the learning process, providing supervision that facilitates end-to-end learning of consistent attentive regions of images of the same person.
- We present the first learning architecture that integrates **attention consistency and Siamese learning** in a joint learning framework.
- We present the first Siamese attention mechanism that jointly models consistent attention across similar images, resulting in a powerful method that can help explain the reasoning behind the network’s prediction.

## 2. Related Work

Traditional person re-id algorithms involved hand-crafted feature design followed by supervised distance met-

ric learning. See Karanam *et al.* [16] and Zheng *et al.* [45] for detailed experimental and algorithmic studies.

Recent developments in deep learning [12, 13] have influenced the design of re-id algorithms as well, with deep re-id algorithms achieving impressive performance on challenging datasets [5, 33, 35]. However, naive training of re-id models without being spatial-localization-aware will not result in satisfactory performance due to cross-view misalignment, occlusions, and clutter. To get around these issues, several recent methods adopt some form of localized representation learning. Zhao *et al.* [43] decomposed person images into different part regions and learned region-specific representations followed by an aggregation scheme to produce the overall image representation. Li *et al.* [18] proposed to first learn and localize part body features by means of spatial transformer networks [15], followed by a combination of local and global features to learn a classification network. Su *et al.* [32] used human pose information as a supervisory signal to learn normalized human part representations as part of an identification network. However, these and several other recent methods [19] consider the spatial localization problem in itself and produce representations and localizations that are not cross-view consistent. On the other hand, our approach tackles spatial localization and representation learning in a holistic, joint framework while enforcing consistency, which is key to re-id.

**Attention** has been used in re-id to tackle localization and misalignment problems. Liu *et al.* [25] proposed the HydraPlus-Net architecture that learns to discover low- and semantic-level attentive features for richer image representations. Li *et al.* [22] designed a scheme to simultaneously learn “hard” region-level and “soft” pixel-level attentive features for a multi-granular feature representation. Li *et al.* [20] learned multiple, predefined attention models and showed that each model corresponds to a specific body part, the outputs of which are then aggregated by means of a temporal attention model. These methods typically have inflexible region-specific attention models as part of the overall framework to learn important regions in the image, and more importantly, do not have an explicit mechanism to enforce attention consistency. Our approach is markedly different from these and other methods [31, 41] in this category in that we only need image-level labels to learn attention, while also enforcing attention consistency by making it a principled part of the learning process.

Consistency is an important aspect of re-id to account for cross-view differences. While this has been studied under the term “equivariance” in some prior work [17], for re-id, it has been reflected in Siamese-like designs that attempt to learn invariant feature representations [7, 9, 21, 29, 42]. These models learn features and distance metrics jointly and do not address the spatial localization problem directly, typically formulating a local parts-based approach to solve

the problem. In scenarios involving occlusion and clutter, this may not be an optimal solution, with attention leading to better spatial localization. To this end, our method, as opposed to these approaches, exploits attention during the learning process while also learning consistent spatial localization and invariant feature representations jointly.

### 3. The Consistent Attentive Siamese Network

In this section, we introduce our proposed attention-based deep architecture for person re-id, the Consistent Attentive Siamese Network (CASN), summarized in Figure 2. CASN includes an identification module and a Siamese module that provide for a powerful, flexible approach to deal with viewpoint variations, occlusions, and background clutter. The identification module (Section 3.1), with its explicit attention guidance as supervision given only identity labels, helps find reliable and accurate spatial localization for the person of interest in the image and performs identity (ID) prediction. The Siamese module (Section 3.2) provides the network with supervisory signals from attention consistency, ensuring that we obtain spatially consistent attention regions for images of the same person, as well as learning view-invariant feature representations for robust gallery matching. In the following, we describe each of these two modules in more detail, leading up to the overall design of the CASN.

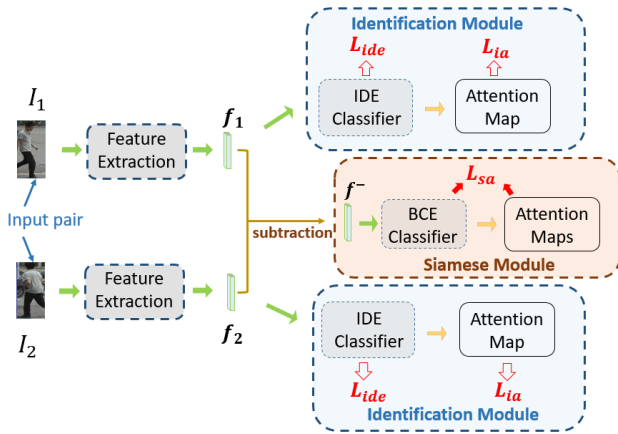


Figure 2: The Consistent Attentive Siamese Network.

#### 3.1. The Identification Module

We first introduce the architecture of the identification module of the CASN. We begin by describing the baseline architecture for training an identification (IDE) model [45], followed by the overall identification module that integrates attention guidance into the IDE architecture.

##### 3.1.1 The IDE Baseline Architecture

The IDE baseline is based on the ResNet50 architecture [12], following the work in [45] and recent papers that adopt ResNet50 [20, 35, 37]. Convolutional layers from *conv1* through *conv5* are pretrained on ImageNet [11], following which an IDE classifier comprised of two fully-connected layers produces the identity prediction for the input image. The identification baseline is visually summarized in Figure 3. Note that while Figure 3 shows the IDE architecture [45], this can be easily swapped with any other baseline architecture that can give the feature vector  $f$ . For instance, to use the part-based convolutional baseline (PCB) architecture [35], one would simply swap the “Feature Extraction” block in Figure 3 with PCB’s backbone prior to obtaining  $f$ . PCB is a modification of IDE that replaces the global average pooling operation in IDE with spatial pooling for discriminative part-informed feature learning. The baseline model is learned by optimizing the identification loss, which essentially maximizes the likelihood of predicting the correct class (identity) label for each training image. Formally, given  $N$  training images  $\{I_n\}_{n=1}^N$  belonging to  $C$  different identities, with each image having an identity label  $\{c_n\}_{n=1}^N \in \{1, \dots, C\}$ , we optimize the following multi-class cross-entropy loss:

$$L_{ide} = - \sum_{n=1}^N \log \frac{\exp(y_{c_n})}{\sum_j \exp(y_j)} \quad (1)$$

where  $y_{c_n}$  is the prediction of class  $c_n$  from the IDE classifier for input image  $I_n$ .

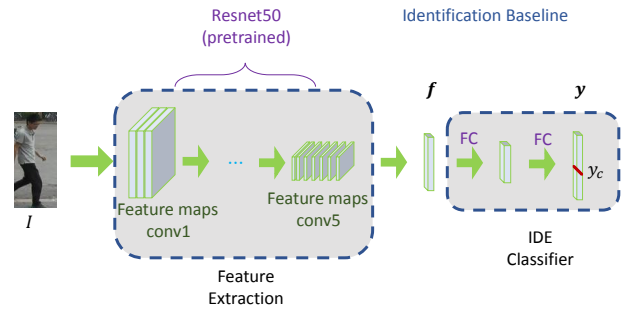


Figure 3: The baseline.  $f$  is the feature vector after Resnet50 *conv5*,  $y$  is the ID prediction vector with dimensionality equal to the total number of training identities, and  $y_c$  is the prediction score of ID label  $c$  for the input image. Note that the “Feature extraction” block here can come from any baseline architecture, e.g., IDE or PCB [35].

##### 3.1.2 Identification Attention

Spatial localization of the person of interest is a key first step for a re-id algorithm, which should be reflected in the end-to-end learning process. While much recent work has

focused on generating attention regions given image-level labels [10, 26, 28, 49], we need to make attention an explicit part of the learning process itself, which can then guide the network to better localize the person of interest.

To this end, we adopt the framework of Li *et al.* [19] and introduce attention learning as part of our identification module, helping the network generate spatially attentive regions in person images without needing any extra information as supervision other than identity labels, which are already available.



Figure 4: An attention map with identification loss (left) and identification loss with attention learning (right).

CAM attention map example shown in Figure 4 (left) for an image from Market1501 [44]. The gray pants of the person attract the most attention, but the blue jacket is also useful information that is ignored in the attention map on the left. To obtain more complete attention maps and focus on the foreground subject, we use the notion of attention learning. Specifically, given  $I_n$  and  $c_n$ , we compute its attention map  $M_n$  and mask out the most discriminative regions in  $I_n$  (corresponding to high responses in  $M_n$ ) by means of the soft-masking operation  $\Sigma(\cdot)$  to get  $\bar{I}_n = I_n * (1 - \Sigma(M_n))$ , where  $*$  is pixel-wise multiplication and  $\Sigma(\cdot) = \text{sigmoid}(\alpha(M_n - \beta))$ . This produces an  $\bar{I}_n$  that excludes all high-response image pixels. If  $M_n$  perfectly spatially localizes the person of interest,  $\bar{I}_n$  will contain no pixels contributing to the corresponding identity prediction  $\bar{y}_{c_n}$ . We use this notion to provide supervision to the identification module to produce more complete spatial localization. Specifically, we define the identification attention loss  $L_{ia}$  for the identification module as the prediction score of masked input image  $\bar{I}_n$ :

$$L_{ia} = \bar{y}_{c_n} \quad (2)$$

A comparison of the attention maps retrieved from a model trained only with the identification loss and one with identification loss and attention learning is shown in Figure 4, where we see more foreground subject coverage with

attention learning on the right. To summarize, in the identification module, we first use the IDE baseline architecture to obtain identity predictions. Attention maps are computed with Grad-CAM and refined using the identification attention objective on masked images that exclude high-attention regions to perform more complete spatial localization.

### 3.1.3 Discussion

While the IDE architecture can provide a good baseline feature representation for matching [16, 37, 45] and our proposed identification module discussed above can further lead to reasonable spatial localization by design, several problems still remain unaddressed. First, the identification module has no mechanism to ensure we obtain consistent attention regions for different images of the same person. This can be inferred from the design itself, which lacks any guiding principle to result in attention consistency. Intuitively, this is key to robust re-id since there are typically common regions in different images of the same person that need to be brought out as important during matching. Second, the identification module has no mechanism to learn invariant identity-aware representations across different camera views. Furthermore, attention consistency should correspond to consistent feature representations, suggesting it should inform representation learning. Finally, the attention component of the identification module is not particularly suitable during inference since we do not know the identity of a test image to compute its attention map. While a workaround to this problem would be to use the top-k predictions to compute attention, this clearly would be a sub-optimal solution.

The problems with the identification module lead us to the design of the Siamese module of the CASN, which attempts to address these issues in a principled manner.

## 3.2. The Siamese Module

In this section, we introduce the Siamese module to complement the identification module of the proposed CASN. Given a pair of input images, we first consider a binary classification problem (Section 3.2.1), whose objective function is then used to formulate a Siamese attention mechanism (Section 3.2.2) to enforce attention consistency and consistency-aware invariant representation learning.

### 3.2.1 Binary Classification

Given a pair of input images, we construct a binary classification objective for predicting whether or not the pair belongs to the same class. Given feature vectors  $\mathbf{f}_1$  and  $\mathbf{f}_2$  for the images  $I_1$  and  $I_2$  in the input pair (see Figure 3), we compute the difference  $\mathbf{f}^- = \mathbf{f}_1 - \mathbf{f}_2$ , which forms the input for a classifier that uses the binary cross-entropy objective (BCE) to get the class prediction for the current



input pair. Note that since we set out to compute attention in the spirit of GradCAM [28], we needed a classification objective to compute Siamese attention described next, and we chose BCE for this purpose. The BCE classifier is structurally similar to the IDE classifier in Section 3.1.1, with two fully connected layers. The output prediction vector  $z$  of the BCE classifier is a 2-dimensional vector, which indicates whether or not the input pair belongs to the same identity. The BCE classification objective that is optimized is defined, for a batch of  $P$  input pairs, as:

$$L_{bce} = - \sum_p \log \left( \frac{\exp(z_{c_p})}{\exp(z_0) + \exp(z_1)} \right) \quad (3)$$

$$c_p \in \{0, 1\}, p = 1, \dots, P$$

where  $z_{c_p}$  is the same ( $c_p = 1$ ) or different ( $c_p = 0$ ) identity prediction of the BCE classifier for input pair  $p$ .

### 3.2.2 The Siamese Attention Mechanism

As discussed previously, identification attention alone does not ensure attention consistency and identity-aware invariant representations. To this end, we propose a new Siamese attention mechanism with explicit guidance towards attention consistency. Consider two images  $I_1$  and  $I_2$  of the same identity and the corresponding BCE classifier prediction  $z_1$ . We first localize the attentive regions in the two images that contribute to this BCE prediction. To this end, we compute the gradient of the prediction score with respect to the feature vector  $\mathbf{f}^-$ , i.e.,  $\frac{\partial z_1}{\partial \mathbf{f}^-}$ . We then find the features in  $\mathbf{f}^-$  that have a positive influence on the final BCE prediction by means of an indicator vector  $\alpha$  constructed as:

$$\alpha_i = \begin{cases} 1, & \text{if } \frac{\partial z_1}{\partial f_i^-} > 0 \\ 0, & \text{otherwise} \end{cases}, i = \{0, \dots, \dim(\mathbf{f}^-)\} \quad (4)$$

Based on the indicator vector  $\alpha$ , the importance scores for the input feature vectors  $\mathbf{f}_1$  and  $\mathbf{f}_2$  can be calculated as the dot products of  $\alpha$  and the feature vectors:  $s_1 = (\alpha, \mathbf{f}_1)$  and  $s_2 = (\alpha, \mathbf{f}_2)$ . In the same spirit as Grad-CAM [28], gradients backpropagated from  $s_1$  and  $s_2$  are first globally average-pooled to find the channel importance weights  $\alpha_1^k = \text{GAP} \left( \frac{\partial s_1}{\partial A_1^k} \right)$  and  $\alpha_2^k = \text{GAP} \left( \frac{\partial s_2}{\partial A_2^k} \right)$ , where  $A_1$  and  $A_2$  are feature maps of image  $I_1$  and  $I_2$  at the last convolutional layer. The attention maps can then be computed as  $M_1 = \text{ReLU} \left( \sum_k \alpha_1^k A_1^k \right)$  and  $M_2 = \text{ReLU} \left( \sum_k \alpha_2^k A_2^k \right)$ .

Visualizations of the attention maps, extracted from the BCE loss, are shown in Figure 5. For images of the same person, we want the attention maps  $M_1$  and  $M_2$  to provide consistent importance to corresponding regions in the images. For instance, as we can see in Figure 5(b), the attention map in Image 1 focuses on the full body of the person while the one in Image 2 mostly focuses on the lower

part. To provide an explicit attention-consistency-aware supervisory signal and guide the network to discover consistent cross-view importance regions, we introduce the notion of spatial attention constraints based on the attention maps derived from the BCE classification objective.

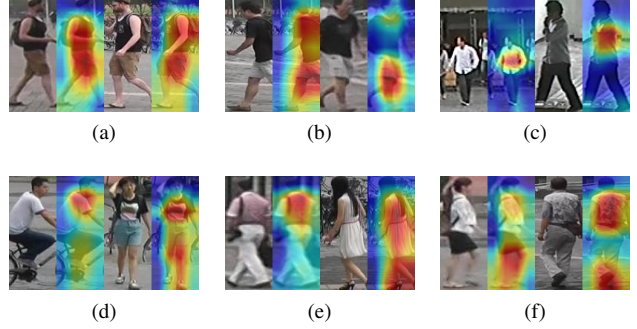


Figure 5: Demonstration of attention maps from BCE loss. (a-c): positive pairs, (d-f): negative pairs.

Given the attention maps  $M_1$  and  $M_2$ , we first apply the max-pooling operation to compute the highest response across each horizontal row of pixels, giving us the two importance vectors  $M_{m1}$  and  $M_{m2}$ . To enforce attention consistency, we explicitly constrain them to be as close as possible. To avoid alignment issues as in Figure 5(c), we find the first and the last element of the vertical vector larger than a certain threshold  $t$  in  $M_{m1}$  and  $M_{m2}$ , and then resize the remaining elements to be of the same dimensions. We define the Siamese attention loss that enforces attention consistency as:

$$L_{sa} = L_{bce} + \alpha \|M_{m1}^* - M_{m2}^*\|_2 \quad (5)$$

where  $L_{bce}$  is defined in Equation 3,  $M_{m1}^*$  and  $M_{m2}^*$  are resized vectors of  $M_{m1}$  and  $M_{m2}$  after alignment,  $\|M_{m1}^* - M_{m2}^*\|_2$  is the  $l_2$  distance between  $M_{m1}^*$  and  $M_{m2}^*$ , and  $\alpha$  is a weight parameter controlling the importance of the BCE loss vis-a-vis the spatial attention constraints.

A visual summary of our proposed Siamese attention mechanism is shown in Figure 6. For input pairs belonging to the same identity, attention maps are retrieved from the BCE classifier predictions, following which they are max-pooled to gather localization statistics for enforcing spatial attention consistency.

### 3.3. Overall Design of the CASN

With the identification and Siamese modules discussed in the previous sections, we now present our overall framework that integrates these two modules. Our proposed CASN, depicted in Figure 2, is a two-branch architecture. During training, we pass as input a pair of images belonging either to the same or different identity. After feature

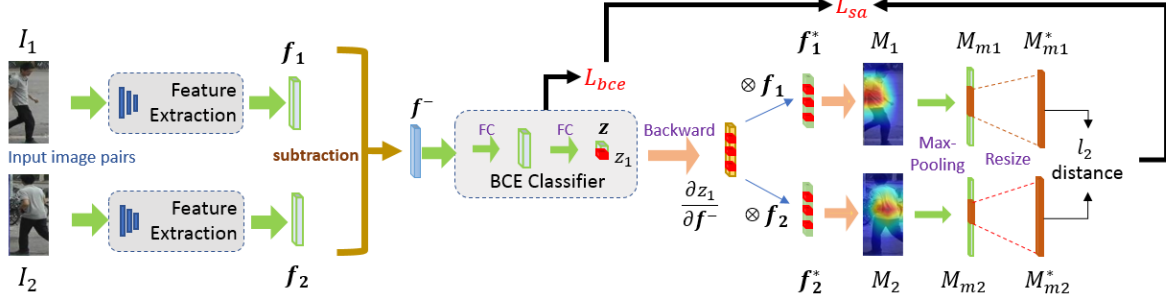


Figure 6: Demonstration of the Siamese Attention Mechanism. Yellow arrows denote backward operation and green arrows denote forward operation. The BCE loss  $L_{bce}$  and spatial constraints are added as Siamese Attention loss  $L_{sa}$ . Note that the “Feature extraction” block here can come from any baseline architecture, e.g., IDE or PCB.

extraction (see Figure 3), the feature vectors are input to the identification module and Siamese module separately. In the identification module, the feature vectors are first passed to the IDE classifier for identity classification, following which an attention map for the input image in the current branch is retrieved from its identity label. The identification attention loss then guides the identification module to discover complete attention regions for the input image. The Siamese module takes as input the element-wise subtraction of the feature vectors from two branches, which is then input to the BCE Classifier to retrieve the image-pair attention maps from the BCE loss. Given this, we enforce the spatial constraint objective to ensure spatial consistency of attentive regions across the two images in the input pair.

We optimize our proposed CASN for all the objectives described here jointly, with the overall CASN training objective given as:

$$L = L_{ide} + \lambda_1 L_{ia} + \lambda_2 L_{sa} \quad (6)$$

where  $L_{ide}$  is the IDE classification loss,  $L_{ia}$  is the identification attention loss, and  $L_{sa}$  is the Siamese attention loss. Note that the feature extraction blocks across the two branches in Figure 2 share weights. The proposed CASN addresses all problems discussed previously in a principled fashion, allowing us to (a) generate attention maps with attention consistency, (b) learn identity-aware invariant representations by design, and (c) use attention maps during inference for identities not seen during training. Furthermore, compared to existing attention mechanisms employed in person re-id, our framework is flexible by design in that it can be used in conjunction with any base architecture or baseline re-id algorithm. For instance, in Section 4, we show performance improvements with both the IDE [40] and the PCB [35] baselines. Furthermore, we only need identity labels during training (which are used by competing algorithms as well), but crucially, do not need any specially designed architecture sub-modules to make attention a part of the learning process.

## 4. Experiments and Results

**Datasets.** We use Market-1501 [44], CUHK03-NP [21, 48], and DukeMTMC-ReID [27, 46]. Market-1501 [44] collects person images from 6 camera views, containing 12,936 training images with 751 different identities. Gallery and query sets have 19,732 and 3,368 images respectively with 750 different identities. CUHK03-NP is a new training-testing split protocol for CUHK03 [21], first proposed in [48], splitting the training and testing sets into 767 and 700 identities. DukeMTMC-ReID [46] is an image-based re-id dataset generated from DukeMTMC [27] that randomly splits training and testing sets equally into 702 identities.

**Implementation Details.** We resize all images to  $288 \times 144$ , use SGD with momentum of 0.9, learning rate of 0.03, and a total of 40 epochs, with the learning rate decreased by a factor of 10 at epoch 30. The parameter  $\alpha$  in Equation 5 is set to 0.2, and  $\lambda_1$  and  $\lambda_2$  in Equation 6 are set to 0.5 and 0.05 respectively. For the PCB baseline, we follow the same protocol as in [35] and resize images to  $384 \times 128$ . We set the batch size to 16, use two NVIDIA GTX-1080Ti GPUs, and implement all code in Pytorch [1].

**Evaluation Protocol.** After training, we use query and gallery as pair inputs to obtain attention maps from BCE classifier predictions. The  $l_2$  distance of the attention maps (Equation 5 in Section 3.2.2) and  $l_2$  distance of the feature vectors are normalized and summed for final ranking. We report the rank-1 Cumulative Match Characteristic (CMC) and mean average precision (mAP) results.

### 4.1. Comparison to the State of the Art

In Tables 1 and 2, we compare the performance of our method with several recently proposed algorithms applied to the CUHK03-NP, DukeMTMC-ReID, and Market-1501 datasets. Note that all our results are evaluations without re-ranking [48] and the PCB [35] architecture as the backend.

**CUHK03-NP.** We report experimental results on both detected and labeled person images. The new train-test split, containing only around 7,300 training images, is much

more prone to overfitting when compared to the other datasets. However, results show that our method surpasses the state of the art substantially for rank-1 (+4.7%, +5.7%) on detected and labeled sets respectively, demonstrating the strong generalization ability of the CASN. More crucially, compared to a recently proposed attention-based method, HA-CNN [22], our CASN achieves 29.8% and 25.8% rank-1 and mAP improvements (on detected sets) respectively.

Table 1: CUHK03-NP (detected and labeled).

	Detected		Labeled	
	R-1	mAP	R-1	mAP
BoW+XQDA [44]	6.4%	6.4%	7.9%	7.3%
LOMO+XQDA [23]	12.8%	11.5%	14.8%	13.6%
IDE [45]	21.3%	19.7%	22.2%	21.0%
PAN [47]	36.3%	34.0%	36.9%	35.0%
DPFL [8]	40.7%	37.0%	43.0%	40.5%
HA-CNN [22]	41.7%	38.6%	44.4%	41.0%
MLFN [4]	52.8%	47.8%	54.7%	49.2%
DaRe+RE [38]	63.3%	59.0%	66.1%	61.6%
PCB+RPP [35]	63.7%	57.5%	-	-
MGN [37]	66.8%	66.0%	68.0%	67.4%
CASN (IDE)	57.4%	50.7%	58.9%	52.2%
<b>CASN (PCB)</b>	<b>71.5%</b>	<b>64.4%</b>	<b>73.7%</b>	<b>68.0%</b>

Table 2: DukeMTMC-ReID and Market-1501 (SQ).

	DukeMTMC-ReID		Market-1501	
	R-1	mAP	R-1	mAP
BoW+KISSME [44]	25.1%	12.2%	44.4%	20.8%
LOMO+XQDA [23]	30.8%	17.0%	43.8%	22.2%
SVDNet [34]	76.7%	56.8%	82.3%	62.1%
HA-CNN [22]	80.5%	63.8%	91.2%	75.7%
DuATM [29]	81.8%	64.6%	91.4%	76.6%
PCB+RPP [35]	83.3%	69.2%	93.8%	81.6%
DNN-CRF [6]	84.9%	69.5%	-	-
<b>MGN [37]</b>	<b>88.7%</b>	<b>78.4%</b>	<b>95.7%</b>	<b>86.9%</b>
CASN (IDE)	84.5%	67.0%	92.0%	78.0%
CASN (PCB)	87.7%	73.7%	94.4%	82.8%

**DukeMTMC-ReID.** We report competitive results in Table 2. Again, compared to recently proposed attention-based methods, HA-CNN [22] and DuATM [29], our CASN achieves 7.2% and 5.9% rank-1 accuracy improvements and 9.9% and 9.1% mAP improvements respectively.

**Market-1501.** We report competitive results with CASN in Table 2. However, compared to recently proposed attention-based methods, e.g., HA-CNN [22] and DuATM [29] (shown in the table), and CAN [24] (R-1: 60.3%, mAP: 35.9%), HPN [25] (R-1: 76.9%), MSCAN [18] (R-1: 80.3%, mAP: 57.5%) our method produces much higher results with both rank-1 and mAP.

As can be noted from these results, the proposed CASN

substantially outperforms existing attention-based methods for re-id. More importantly, unlike these competing attention-based methods, CASN does not require any specially designed deep architecture for modeling attention, relying only on identity labels for supervision. This allows the CASN to be highly flexible for use in conjunction with any baseline CNN architecture, such as VGGNet [30], DenseNet [13], or SqueezeNet [14]. For instance, with DenseNet and the IDE baseline, CASN achieves a rank-1 and mAP performance of 57.2% and 52.0% respectively on CUHK03-NP (detected), which are close to CASN’s results with ResNet50 and IDE, discussed next.

## 4.2. Ablation Study and Discussion

In this section, we further study the role of the identification attention and Siamese attention mechanisms individually, and how they influence the performance of the CASN. In Table 3, we report evaluation results of our proposed model on CUHK03-NP (detected), DukeMTMC-ReID and Market-1501, starting from baseline IDE and PCB architectures and working up to the full CASN model. From Table 3, we can see clear performance improvements over the baseline with individual attention modules. For instance on CUHK03-NP, IDE+IA improves the rank-1 and mAP performance of baseline IDE by 9.0% and 9.2% whereas IDE+SA improves the rank-1 accuracy by 9.4% and 10.2% respectively. This provides evidence for our initial hypothesis that spatial localization, via end-to-end trainable attention mechanisms, should be an important and integral part of the framework design. Furthermore, adding both attention modules improves performance as measured by both rank-1 accuracy and mAP, demonstrating the importance of using both identification and Siamese modules.

Table 3: Ablation study. IA: Identification Attention, SA: Siamese Attention, SQ: Single-Query.

Loss type	CUHK03-NP		DukeMTMC-ReID		Market-1501 (SQ)	
	R-1	mAP	R-1	mAP	R-1	mAP
IDE [35]	43.8%	38.9%	73.2%	52.8%	85.3%	68.5%
IDE + IA	54.8%	48.1%	83.2%	66.0%	91.0%	76.9%
IDE + SA	55.2%	49.1%	83.5%	66.0%	91.6%	77.7%
<b>CASN(IDE)</b>	<b>57.4%</b>	<b>50.7%</b>	<b>84.5%</b>	<b>67.0%</b>	<b>92.0%</b>	<b>78.0%</b>
PCB [35]	61.3%	54.2%	81.7%	66.1%	92.4%	77.3%
PCB + IA	68.5%	62.4%	87.3%	73.4%	93.9%	81.8%
PCB + SA	69.9%	64.2%	86.8%	73.5%	94.1%	82.6%
<b>CASN(PCB)</b>	<b>71.5%</b>	<b>64.4%</b>	<b>87.7%</b>	<b>73.7%</b>	<b>94.4%</b>	<b>82.8%</b>

Comparisons of the attention maps acquired from the models trained with BCE loss and BCE loss with Siamese Attention loss are shown in Figure 7(a-b). Clearly, with the proposed Siamese attention mechanism, we obtain more consistent attention maps of the same person image pair in Figure 7(b) compared to Figure 7(a). Furthermore, we also demonstrate these attention maps for the testing image pairs

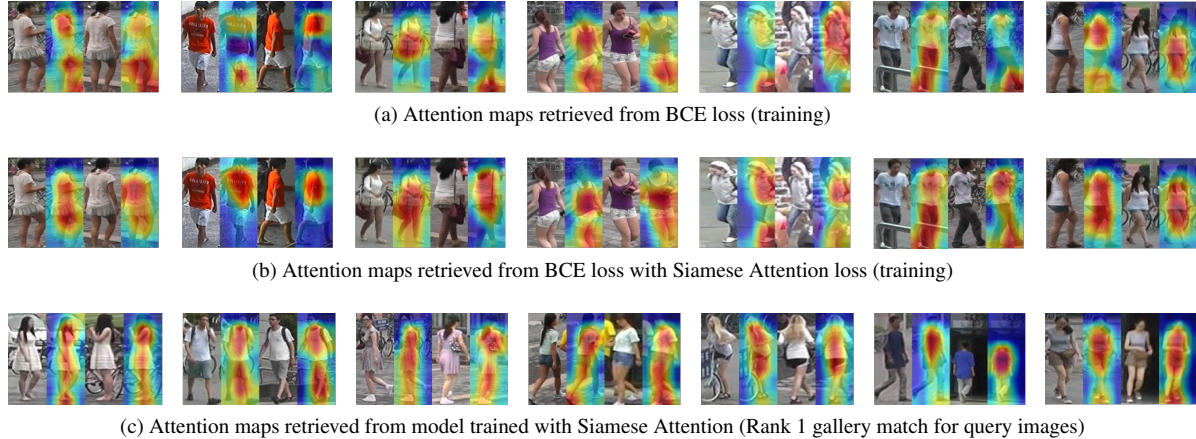


Figure 7: Demonstrating the efficacy of the proposed Siamese attention by means of attention maps for same-person images.

in Figure 7(c), where we again see attention consistency among the query and retrieved gallery images. These examples demonstrate the effectiveness of our proposed Siamese attention mechanism, and also provide a powerful interpretability tool. With such attention maps, we can now explain why our Siamese network predicts a certain input image pair to be similar or dissimilar, leading to intuitive explanations for person re-id. In more detail, Figure 8(a) shows two query images (one on each row), along with their rank-1 (left column) and ground-truth matches (right column). Each rank-1 match is a wrong match (failure case) while the ground-truth has a lower rank, and we can understand the reasoning from our attention maps. For instance, on the first row, we see reasonable attention consistency between the query and rank-1 (notice both show women in dresses), explaining why the wrong match was ranked 1, unlike the ground-truth, where we see attention focused on different regions, leading to lower rank (rank 3 in this example). In Fig 8(b), we demonstrate the efficacy of our proposed Siamese attention (two examples, one on each row). The left column shows {query, ground-truth} and the ground truth’s rank without Siamese attention. The right column shows these results with Siamese attention. We can see that Siamese attention results in better attention consistency, which is also reflected in the improved rank.

## 5. Conclusions

We proposed the first learning architecture that integrates attention consistency modeling and Siamese representation learning in a joint learning framework, called the Consistent Attentive Siamese Network (CASN), for person re-id. Our framework provides for principled supervisory signals that guide our model towards discovering consistent attentive regions for same-identity images while also learning identity-aware invariant representations for cross-view matching. We conducted extensive evaluations on three popular person

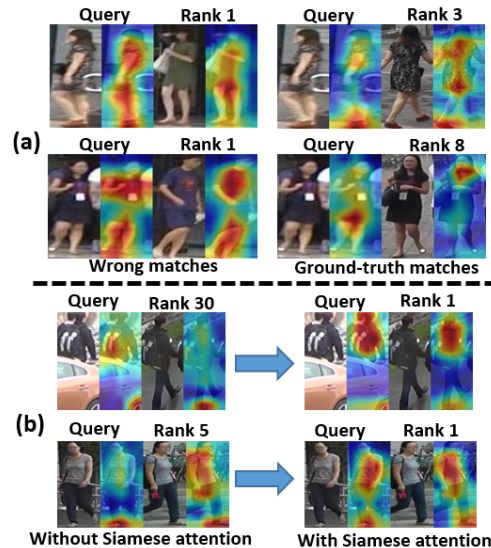


Figure 8: (a) Our attention maps can explain wrong (high rank, e.g., rank 1) and ground-truth matches (low rank, e.g., rank 3). (b) Siamese attention gives rank improvements, providing reasoning with attention consistency.

re-id datasets and demonstrated competitive performance. While computing attention as in Section 3.2.2 is specific to standing poses that are common in existing benchmarks, our framework is extensible to enforce different kinds of consistency given data- or domain-specific priors for real-world generalizability.

**Acknowledgements** This material is based upon work supported by the U.S. Department of Homeland Security under Award Number 2013-ST-061-ED0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.



## References

- [1] Pytorch. <https://pytorch.org/>.
- [2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, 2016.
- [3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.
- [4] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018.
- [5] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *ECCV*, 2018.
- [6] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep CRF for person re-identification. In *CVPR*, 2018.
- [7] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *CVPR*, 2017.
- [8] Yanbei Chen, Xiatian Zhu, , and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *ICCVW*, 2017.
- [9] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *CVPR*, 2016.
- [10] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Jan 2017.
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [13] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.
- [14] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. *arXiv:1602.07360*, 2016.
- [15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*. 2015.
- [16] Srikrishna Karanam, Mengran Gou, Ziyang Wu, Angels Rates-Borras, Octavia Camps, and Richard J. Radke. A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:523–536, Mar. 2019.
- [17] Karel Lenc and Andrea Vedaldi. Learning covariant feature detectors. In *ECCVW*, 2016.
- [18] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.
- [19] Kunpeng Li, Ziyang Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *CVPR*, 2018.
- [20] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, 2018.
- [21] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deep-reid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.
- [22] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.
- [23] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.
- [24] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, July 2017.
- [25] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017.
- [26] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015.
- [27] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016.
- [28] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.
- [29] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex ChiChung Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018.
- [30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [31] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, pages 1179–1188, 2018.
- [32] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.
- [33] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018.
- [34] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. SVDNet for pedestrian retrieval. In *ICCV*, 2017.
- [35] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018.
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.
- [37] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In *ACM MM*, 2018.
- [38] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kil-

- ian Q. Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, 2018.
- [39] Lin Wu, Yang Wang, Junbin Gao, and Xue Li. Where-and-when to look: Deep Siamese attention networks for video-based person re-identification. *IEEE Transactions on Multimedia*, 2018.
  - [40] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.
  - [41] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, 2018.
  - [42] Hongsheng Li Yantao Shen, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, 2018.
  - [43] Liming Zhao, Xi Li, Jingdong Wang, and Yueting Zhuang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017.
  - [44] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.
  - [45] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. *ArXiv e-prints*, 2016.
  - [46] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *ICCV*, 2017.
  - [47] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.
  - [48] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.
  - [49] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.