Unsupervised Attention Based Instance Discriminative Learning for Person Re-Identification

Kshitij Nikhal
University of Nebraska-Lincoln
knikhal2@huskers.unl.edu

Benjamin S. Riggan
University of Nebraska-Lincoln
briggan2@unl.edu

Abstract

Recent advances in person re-identification have demonstrated enhanced discriminability, especially with supervised learning or transfer learning. However, since the data requirements—including the degree of data curations—are becoming increasingly complex and laborious, there is a critical need for unsupervised methods that are robust to large intra-class variations, such as changes in perspective, illumination, articulated motion, resolution, etc. Therefore, we propose an unsupervised framework for person re-identification which is trained in an end-to-end manner without any pre-training. Our proposed framework leverages a new attention mechanism that combines group convolutions to (1) enhance spatial attention at multiple scales and (2) reduce the number of trainable parameters by 59.6%. Additionally, our framework jointly optimizes the network with agglomerative clustering and instance learning to tackle hard samples. We perform extensive analysis using the Market1501 and DukeMTMC-reID datasets to demonstrate that our method consistently outperforms the state-of-the-art methods (with and without pre-trained weights).

1. Introduction

Person re-identification (Re-ID) is the process of detecting identical persons to a query subject in frames from distributed cameras with non-overlapping field of views (FOVs). Given a query image [42], a video sequence [27] or text description [35], Re-ID is used to track individuals in public spaces, such as airports, universities, malls, cities, etc. For example, Re-ID may be used to rapidly build a timeline for a person's whereabouts prior to a critical event (or crime), such as a mass shooting, bombing or other public threat, providing beneficial tools for local and federal law enforcement or military/government intelligence, surveillance and reconnaissance (ISR).

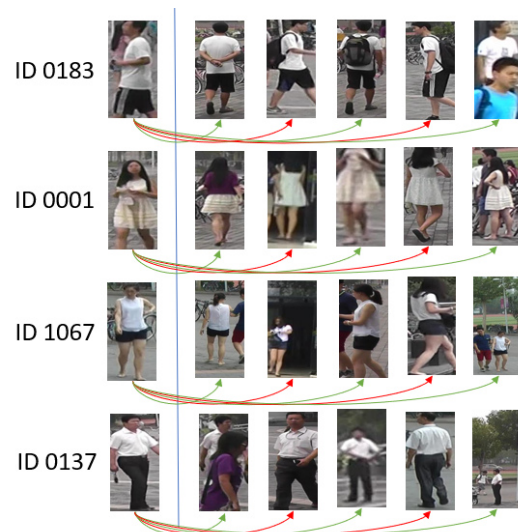The primary challenge with Re-ID is detecting persons

Figure 1: Example hard positive (green arrows) and hard negative (red arrows) examples for Re-ID.

that match query subjects under significant variations in viewpoint, resolution, compression, illumination, and occlusion. Also, temporal variations such as changes in clothing, hair style, or accessories present additional challenges.

Despite initial success using hand crafted features [5, 34] and metric learning [10, 17], current state-of-the-art [41] Re-ID techniques are based on convolutional neural networks (CNNs), which, if supervised, require large-scale annotated (i.e., labeled) datasets to learn a robust embedding subspace. Comprehensive surveys on person Re-ID using hand crafted systems and recent deep learning techniques are presented in [7, 44] and [29, 36], respectively.

Annotating large-scale datasets for Re-ID, especially methods requiring multiple bounding boxes for each person, is very labor intensive, time consuming and cost prohibitive. Therefore, in this paper, we propose a new unsupervised framework for Re-ID.

Unsupervised methods commonly exploit pre-trained models or transfer learning, where parameters are assumed

to generalize between two independent tasks, in order to improve the accuracy. For example, [18] uses the Resnet50 [8] architecture, pre-trained on ImageNet [1], before fine-tuning the network in an unsupervised manner and [30] combines one-shot and unsupervised learning. Although pre-training and transfer learning have been empirically shown to significantly enhance the performance of neural networks, it is not amenable for adapting parameters between significantly different domains or architectures.

In this paper, we propose a new framework that exploits "soft" attention maps (section 4.1), which are produced by efficiently and effectively combining group convolutions with channel-wise attention to enhance unsupervised spatial attention and to alleviate over-fitting by minimizing the number of trainable parameters.

Additionally, instance discriminative learning and hierarchical clustering are jointly used to improve unsupervised learning with and without pre-training. Similar to supervised attention models, this joint optimization enables our attention mechanism to learn image regions that are most discriminative according to clustering metrics in an unsupervised manner (i.e., without labels). We assign pseudo-labels to each training sample and learn to be robust to representations of the same instance under various perturbations (section 4.2). In successive stages, we tackle the hard positives, as shown in Figure 1, by gradually merging samples together using similarity scores (section 4.3).

Our proposed framework achieves enhanced performance for unsupervised Re-ID (with and without pre-training) using the Market1501 [43] and DukeMTMC-reID [47] datasets, beating the current state-of-the-art.

The contributions of this paper include:
- a new grouped attention module (GAM),
- the joint optimization of an instance discrimination loss (IDL) and agglomerative clustering loss (ACL),
- extensive analysis for Re-ID using Market1501 and DukeMTMC-reID datasets,
- ablation studies that analyze attention maps, embedding dimensionality, and number of filter groups.

2. Related Work

Supervised Person Re-ID: Supervised methods for Re-ID have been successful, in part, due to ubiquitous graphics processing units (GPUs), machine learning application programming interfaces (APIs), and large-scale datasets with annotations. However, supervised methods must ensure they are not over-fitting to any particular dataset.

Recently, AlignedReID [41] achieved impressive performance on the Market1501 [43] dataset with a rank-1 accuracy of 94.4% by jointly learning the global features with the local features. Zhou et al. [48] proposed a new efficient architecture that achieves 94.8% with 2.2 million parameters compared to the ResNet50 [8] architecture with around

24 million parameters, reducing the possibility of overfitting. Zheng et al. [46] use joint generative and discriminative learning and achieves an accuracy of 94.8% on the Market1501 dataset. In this paper, we aim to minimize the gap between supervised and fully unsupervised performance.

Unsupervised Person Re-ID: Both dictionary learning and metric learning have been applied to unsupervised person Re-ID. Kodirov et al. [12] use a graph regularised dictionary learning algorithm with a robust L1-norm term to learn cross-view discriminative information. Yu et al. [38] learn a specific projection for each camera view based on asymmetric clustering. However, these features are not as discriminative as deeply learned features. Wu et al. [30] focus on one-shot learning for video-based Re-ID to exploit unlabeled tracklets by progressively adding hard samples. Yu et al. [39] learn a soft multi-label for each unlabeled person by comparing (and representing) the unlabeled person with a set of known reference persons from a labeled dataset. Xiao et al. [31] propose an online instance matching loss function that maintains a lookup table of features from all the labeled identities, and compare distances between mini-batch samples and all the registered entries. Lin et al. [18] use bottom up clustering (BUC) to optimize a CNN and balances cluster volume using a diversity term. However, several methods that claim to be unsupervised use some form of supervision like labeled supplementary datasets [39], pre-trained ImageNet weights [18], or one labeled tracklet for each identity [30]. In our work, we present experiments with and without pre-trained weights initialization.

Attention Models: Attention-based methods [14, 33] involve an attention mechanism to extract additional discriminative features. In comparison with pixel-level masks, attention regions can be regarded as an automatically learned high-level masks. This is analogous to a segmentation model that does not require any annotations and which is learned automatically. Li et al. [15] proposed the Harmonious Attention CNN (HA-CNN) model that combines the learning of soft pixel and hard regional attentions along with simultaneous optimization of feature representations. Zheng et al. [45] proposed an attention-driven siamese network called the Consistent Attentive Siamese Network that uses identity labels as supervision. Wang et al. [25] proposed Residual Attention Network which uses an encoder decoder style attention module. Woo et al. [28] proposed the Convolutional Block Attention Module (CBAM) that infers attention maps along channel and spatial dimensions sequentially. Our attention mechanism is motivated by CBAM [28], except that our proposed framework is more efficient, using substantially fewer parameters.

Unsupervised representation: A classical approach for unsupervised representation learning is to perform clustering (e.g. k-means). Another popular method is to use auto-
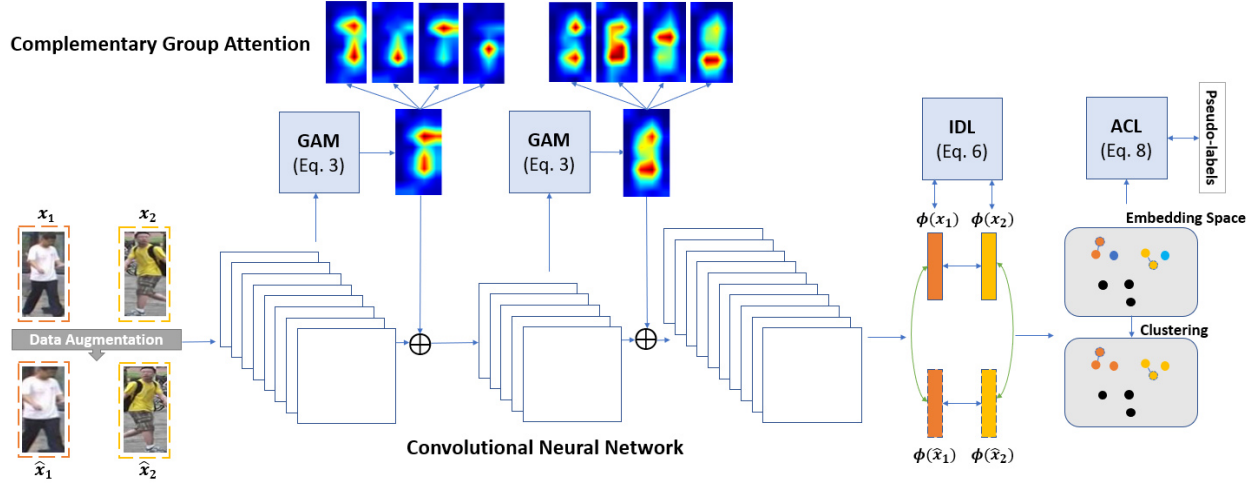
Figure 2: Our framework augments a network with Grouped Attention Modules (GAMs) at multiple scales and trains the network in a fully unsupervised manner using the instance discriminative loss (IDL) and agglomerative clustering loss (ACL). The complementary group attention maps helps generate better attention by learning filter relationships in a more structured way (e.g., the third filter group focuses on the top part only)

.

encoders [24] to compress images into a latent representations, such that images may be optimally reconstructed. Dosovitskiy *et al.* [3] propose to discriminate among a set of surrogate classes, where the surrogate classes are formed by applying a variety of transformations to randomly sampled image patches. In [37], a softmax embedding variant is used where augmented samples should be classified as the same instance while other instances in the batch are considered as negative samples.

Unlike previous methods, our approach combines agglomerative clustering with an attention based CNN and pseudo-supervision in which neither labels nor pre-trained weights are used. Our approach differs from BUC by using a novel attention mechanism and an efficient architecture to produce discriminative representations associated with persons in the field of view. Moreover, our combination of IDL and ACL alleviates the need to use conventional pre-trained architectures. Thus, our proposed approach is capable of being trained in a fully unsupervised manner.

3. Preliminaries

Given an unlabeled training set $X = \{x_1, x_2, ...x_n\}$, the goal is to learn $\phi(x_i; \theta)$—a mapping parameterized by θ used to extract features from an image x_i. This mapping is then applied to the gallery set $X^g = \{x_1^g, x_2^g, ...x_{ng}^g\}$ and query set $X^q = \{x_1^q, x_2^q, ...x_{nq}^q\}$. The gallery set can be considered the test set or the total collection of detections in the database. Representations of the query images, $\phi(x_i^q; \theta)$ are used to search the gallery set to retrieve the most similar matches to x_i^q according to Euclidean distance between the query and gallery embeddings, $d(x_q, x_g) =$

$\|\phi(x_q; \theta) - \phi(x_g; \theta)\|$, where a smaller distance implies increased similarity between the images. Ideally, the top-k (for k equal to 1, 5, or 10) matches returned will correspond to the same identity as that from the query image.

4. Methodology

Our hybrid framework for fully unsupervised Re-ID (Figure 2) consists of three main components: Grouped Attention Module (GAM), Instance Discriminative Loss (IDL), and Agglomerative Clustering Loss (ACL).

Our GAM, which uses reduced filter group co-dependence [23] in conjunction with compact channel and spatial attention, yields more accurate (and efficient) attention maps. Similar to [11, 13], our GAM consists of convolutional layers with filter groups that consistently learn complementary attention maps. Thus, more discriminative representations are generated for enhancing clustering capabilities by focusing on persons' appearances rather than background/clutter, which is observed by the increased accuracy between GAM and CBAM in section 5.3.

IDL and ACL have complementary objectives in regards to unsupervised Re-ID. IDL plays an important role by making the network invariant to different cross-camera views by maximizing the similarity measure between the original images and augmentations (e.g., zoomed/flipped). Unlike IDL, ACL aims to iteratively merged clusters and optimize similarity between image representations (enhanced by GAM and IDL) and associated clusters. Therefore, we jointly optimize both losses to more effectively learn discriminative representations for unsupervised person Re-ID.
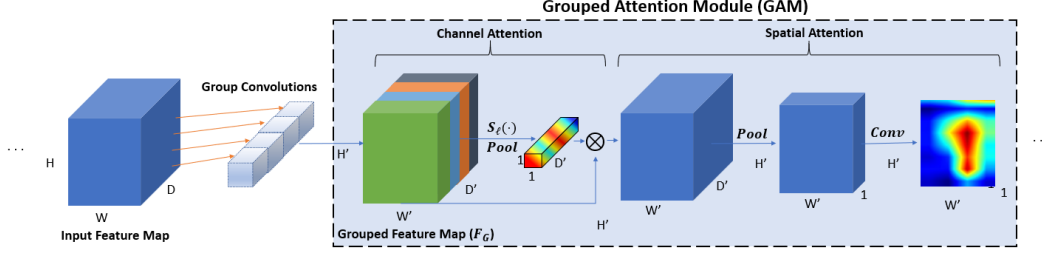
Figure 3: Overview of our Group Attention Module (GAM), combining group-based channel attention and spatial attention.

Each component is discussed in subsections 4.1–4.3.

4.1. Grouped Attention Module (GAM)

The proposed attention mechanism works by summarizing activations of previous layers to emphasize "important" regions for subsequent layers. The attention map is inferred from an intermediate feature map to generate attention aware features. The attention module is split into two attention maps: channel and spatial attention (Figure 3).

Channel Attention: This module exploits the inter-channel relationship of the features, focusing on the important features in the input image. The intermediate feature map is aggregated to remove spatial information and retain the channel information only. It is aggregated by using a sub-network to infer channel-wise attention. In our approach, we use average and max pooling layers to remove the spatial information followed by a linear layer. Let F be an intermediate feature map of dimension $\mathbb{R}^{C \times H \times W}$ that is input to the attention module. The channel attention is

$$A_c(F) = \sigma(S_\ell(Pool(F))), \qquad (1)$$

where S_ℓ is a fully connected layer and σ is the sigmoid activation function.

The intermediate convolutional layers, producing the input F in Eq. 1, are replaced with group convolutions, whose feature maps are denoted as F_G. This significantly reduces the parameters while improving the attention maps (section 5.5). We use 4 filter groups which reduces the parameters from 26 million parameters to 10.5 million parameters, hence making the network efficient by reducing 59.6% of the parameters compared to the Resnet50 [8]. So the channel attention can now be stated as,

$$A_c(F_G) = \sigma(S_\ell(Pool(F_G))). \qquad (2)$$

Spatial Attention Module: The spatial attention module is used to learn the pixels that contribute the most toward the network's inference. For every spatial index (u, v) for F_G, A_c and $F_G(u, v)$ are multiplied in an element-wise manner. For compactness, we denote the procedure as $A_c \otimes F_G$. This new channel attention enhanced feature map is passed to the spatial attention module. Average and max pooling

is applied along the channel axis, which is shown to be effective in highlighting informative regions [40]. Then, we apply a 7×7 convolutional layer to generate a spatial attention map that encodes whether pixels are emphasized or deemphasized.

$$A_s(A_c(F_G)) = \sigma((Conv(Pool(A_c(F_G) \times F_G)))), \quad (3)$$

where $A_s(\cdot) \in \mathbb{R}^{1 \times H \times W}$.

We refer to this sequential attention inference with group convolutions as the grouped attention module (GAM). Despite not being explicitly trained in a supervised manner, our network learns to focus on discriminative image regions. Specifically, as shown in section 5, the attention maps and feature representations are improved.

4.2. Instance Discriminative Loss (IDL)

Ideally, the objective is to retrieve genuine correspondences to the query subject, which implies that we desire genuine and imposter similarity/dissimilarity score distributions that are maximally separable. Therefore, our proposed framework, including GAMs, is trained such that two feature embeddings from imagery acquired across multiple camera views for the same subject are sufficiently close in terms of Euclidean distance and embeddings from different subjects produce higher dissimilarity scores.

Cross-camera views can be imitated to a certain extent with data augmentations. We use various augmentation techniques like random crops, zoom, horizontal flips and occlusions to approximate cross-view variations, and we use pseudo-supervision to classify instances (with unknown identities) to be the same as their augmentations.

Thus, we minimize the difference between an instance and its augmented version while maximizing the difference between other instances. This can be modeled as a binary classification problem. Particularly, for sample x_i, the augmented sample $\hat{x}_i$ is classified as instance i while other samples, x_j for $j \neq i$ are not classified as instance i. The probability of an augmentation, $\hat{x}_i$ being classified the same as image x_i is

$$P(i|\hat{x}_i) = \frac{\exp(\phi(x_i; \theta)^T \phi(\hat{x}_i; \theta)/\tau)}{\sum_k \exp(\phi(x_k; \theta)^T \phi(\hat{x}_i; \theta)/\tau)}, \qquad (4)$$

where τ is a temperature parameter that controls the softness of the probability distribution [9]. Since the embeddings are L2 normalized, maximizing the numerator in Eq. 4 implies increasing the cosine similarity between $\phi(x_i; \theta)$ and $\phi(\hat{x}_i; \theta)$. Thus, maximizing Eq. 4 encourages instances to be robust to cross-camera views using data augmentations. The probability of x_j being classified as instance i is

$$P(i|x_j) = \frac{\exp(\phi(x_i;\theta)^T \phi(x_j;\theta)/\tau)}{\sum_k \exp(\phi(x_k;\theta)^T \phi(x_j;\theta)/\tau)}. \quad (5)$$

Therefore, we want to maximize Eq. 4 and minimize Eq. 5 which is equivalent to minimizing the negative log likelihood

$$J_{idl} = -\sum_i \log P(i|\hat{x}_i) - \sum_i \sum_{j \neq i} \log(1 - P(i|x_j)). \quad (6)$$

4.3. Agglomerative Clustering Loss (ACL)

To improve the discriminative capabilities, our framework incorporates agglomerative clustering, which is a form of hierarchical clustering that successively merges instances into clusters in a bottom up manner. Each cluster center M_β for $\beta \in \{1 \ldots |M|\}$ is used to form a memory bank, M, where $|M|$ denotes the size of the memory bank (i.e., the number of clusters). Initially, $|M| = n$, meaning that all training instances x_i for $i = 1 \ldots n$ are their own singleton clusters. However, as instances merge (i.e, $|M| < n$), non-singleton clusters are formed.

Let β_i denote the cluster label corresponding to x_i for $i = 1 \ldots n$. The probability that image x_i belongs to a cluster β_i is given by

$$P(\beta_i|x_i) = \frac{\exp(M_{\beta_i}^T \phi(x_i;\theta)/\tau)}{\sum_k \exp(M_{\beta_k}^T \phi(x_i;\theta)/\tau)}. \quad (7)$$

At each learning iteration, the representation $\phi(x_i; \theta)$ and the parameters θ are optimized by stochastic gradient descent. Then, the optimized representation is used to update the memory bank M. Therefore, the objective function is to minimize the negative log likelihood as,

$$J_{acl} = -\sum_{i=1}^n \log P(\beta_i|x_i). \quad (8)$$

Successively, each singleton cluster is then merged together according to a dissimilarity metric using the feature representation. The number of clusters N_C to merge is a hyper-parameter and can be dynamically changed during training. We use the Euclidean distance metric:

$$d_0(\phi(x_i), \phi(x_j)) = \sqrt{\phi_{ij}^T \phi_{ij}}, \quad (9)$$

where we dropped the parameter, θ, for compactness and $\phi_{ij} = \phi(x_i) - \phi(x_j)$.

To avoid mode collapse, we use a balancing term in the similarity measure as described in [18]

$$d(\phi(x_i), \phi(x_j)) = d_0 + \lambda(|M_i| + |M_j|) \quad (10)$$

where λ is the weighting term to balance the impact of the balancing term. This helps in merging small clusters while also merging large clusters that are very similar.

Our proposed framework uses both the IDL (Eq. 6) and ACL (Eq. 8) to train a network using GAMs. The total loss is expressed as

$$J_{total} = J_{idl} + J_{acl}. \quad (11)$$

5. Experiments

In this section, we describe the extensive experimental analysis and summarize the results using two established Re-ID benchmark datasets. First, we describe the datasets and implementation details for purposes of reproducibility. Then, we provide quantitative and qualitative analysis, comparing our proposed unsupervised framework with recent unsupervised methods (with and without pre-training). Lastly, we present ablation studies that analyze attention maps, embedding dimensions, and number of filter groups.

5.1. Datasets

For experimental analysis, we employed two widely used image-based Re-ID datasets: the Market1501 and DukeMTMC-reID.

The **Market1501** [43] dataset uses imagery from six cameras in front of the Tsinghua University campus. This dataset contains 32,668 bounding boxes of 1,501 identities. The training set consists of 12,936 images of 751 identities and the testing set contains 19,732 of 750 identities. The dataset uses the Deformable Part Model [6] as opposed to hand drawn bounding boxes. This introduces misalignment, part missing and false positives which reflects a realistic setting. The dataset is collected in an open system, where each identity has multiple images under each camera.

The **DukeMTMC-reID** [47] dataset is a subset of the DukeMTMC [22] dataset for image-based Re-ID which is similar to the format of the Market1501 dataset. There are 1,404 identities that appear in more than two cameras and 408 distractors who appear only in one camera. The dataset is split by randomly selecting 702 identities as the training set and 702 identities as the testing set. In the testing set, one image is picked for each identity as the query image while the others are put in the gallery set. As a result, a

Table 1: Market1501 results using pretrained weights.

Method	Label	rank-1	rank-5	rank-10	mAP
BOW [43]	ImageNet	35.8%	52.4%	60.3%	14.8%
OIM [31]	ImageNet	38.0%	58.0%	66.3%	14.0%
UMDL [19]	Transfer	34.5%	52.6%	59.6%	12.4%
PUL [4]	Transfer	44.7%	59.1%	65.6%	20.1%
EUG [30]	OneShot	49.8%	66.4%	72.7%	22.5%
SPGAN [2]	Transfer	58.1%	76.0%	82.7%	26.7%
TJ-AIDL [26]	Transfer	58.2%	-	-	26.5%
BUC [18]	ImageNet	61.9%	73.5%	78.2%	29.6%
IDL*	ImageNet	39.7%	56.8%	64.7%	15.3%
IDL+ACL*	ImageNet	56.9%	74.8%	80.6%	30.1%
IDL+ACL+CBAM*	ImageNet	63.3%	**81.3%**	**86.3%**	35.0%
ACL+GAM *	ImageNet	57.3%	75.3%	81.9%	28.3%
IDL+ACL+GAM*	ImageNet	**63.9%**	78.8%	85.1%	**35.7%**

* Our approach / implementation.

Table 2: Market1501 results (no pretrained weights).

Method	rank-1	rank-5	rank-10	mAP
BUC* [18]	10.7%	21.7%	27.8%	3.1%
gBiCov [32]	8.28%	-	-	2.23%
HistLBP [20]	9.62%	-	-	2.72%
LOMO [16]	26.07%	-	-	7.75%
BOW+MultiQ [43]	42.64%	-	-	18.68%
IDL*	25.5%	42.4%	51.0%	9.1%
IDL+ACL*	48.2%	68.0%	76.2%	24.3%
IDL+ACL+CBAM*	48.5%	67.2%	74.3%	25.6%
IDL+ACL+A-CBAM *	45.2%	63.1%	70.6%	21.7%
ACL+GAM *	44.6%	63.5%	71.6%	20.3%
IDL+ACL+GAM*	**53.6%**	**71.9%**	**79.4%**	**29.5%**

* Our approach / implementation. A-CBAM - Adjusted CBAM
i.e. parameters are controlled to match GAM.

total of 16,552 images are available for training and 19,889 images are available for testing.

Evaluation: We followed the evaluation protocols described for the respective datasets, where the query and gallery samples were captured by different cameras. Then, given a query image sequence, all gallery items were assigned a similarity score and were ranked according to their similarity with the query. The search process was performed in a cross-camera mode, i.e., relevant images captured in the same camera as the query were not considered.

For quantitative analysis, we used the rank-k accuracy (for $k = 1, 5$, and 10) and mean average precision (mAP) to compare our proposed framework (and components) with other methods. The rank-k accuracy is derived from the cumulative matching characteristic (CMC) curve, where a match is considered to be correct if a true match (according to the ground truth set) is returned within the top-k most similar matches. The average precision (AP) is computed as the area under the precision-recall curve, and the mAP is the average over all queries. A key difference between rank-k accuracy and mAP is that rank-k considers the most similar match scores and mAP considers the separability between genuine and imposter score distributions.

5.2. Implementation Details

The CNN used in our experiments to evaluate the proposed framework was based on the Resnet50 architecture. Resnet50 is composed of 16 bottleneck blocks: $Bottleneck(x) = Relu(\mathcal{F}(x)+x)$, where $\mathcal{F}(x) = Conv \circ BN \circ Conv \circ BN \circ Relu \circ Conv \circ BN \circ Relu(x)$ and $Conv$, BN, and $Relu$ represent convolution, batch normalization, and rectified linear unit layers, respectively. However, with our GAM, the bottleneck layers are $Relu(GAM \circ \mathcal{F}(x)+x)$ and all $Conv$ layers in $\mathcal{F}(x)$ are replaced with group convolutions using 4 filter groups.

All proposed methods are optimized using stochastic

gradient descent (SGD) with weight decay of 0.9 and with a batch size of 32. Initially, the effective learning rate is set to 0.01 when layers are initialized with pre-trained weights and 0.1 otherwise. After 25 epochs, the learning rate is reduced by a factor of 10. The merging percent of clusters is set to 4%. All input images were resized to 256×128 (H $\times$ W) and augmented using random horizontal flips, random contrast change, random zoom and random crop. The temperature parameter τ in Eqs. 4, 5 and 7 is set to 0.1.

In the subset of experiments in which we initialized using pre-trained weights, we pre-trained our modified Resnet50 architecture on ImageNet to provide a fair comparison with other architectures that were either pre-trained on ImageNet in a similar fashion or used some alternative discriminative pre-training. However, for completeness, we also compare our proposed framework with other methods without any discriminative pre-training.

Our code is available at https://git.unl.edu/ece-unl-images-lab/group-attention-module-person-re-id.

5.3. Quantitative Analysis

In the following experiments, we compare our proposed framework with bottom up clustering (BUC) [18] and several recent methods. We provide the results of fair comparisons for two scenario: with and without pre-trained weights.

Using the Market1501 dataset, Table 1 shows the rank-1, rank-5, and rank-10 Re-ID accuracy and mAP under the unsupervised scenario with pre-trained weights. Our proposed framework, using IDL and ACL to train the Resent50 architecture with our GAM, achieved the best rank-1 accuracy and mAP. Whereas, our approach using CBAM achieved the best rank-5 and rank-10 accuracy. Note that, in this case, we improved upon the state-of-the-art rank-1 accuracy and mAP reported in [18] by 2.0% and 6.1%, respectively.

Table 2 shows the rank-1, rank-5, and rank-10 Re-ID

Table 3: DukeMTMC-reID results with pretrained weights.

Method	Label	rank-1	rank-5	rank-10	mAP
BOW [43]	ImageNet	17.1%	28.8%	34.9%	8.3%
OIM [31]	ImageNet	24.5%	38.8%	46.0%	11.3%
UMDL [19]	Transfer	18.5%	31.4%	37.6%	7.3%
PUL [4]	Transfer	30.4%	46.4%	50.7%	16.4%
EUG [30]	OneShot	45.2%	59.2%	63.4%	24.5%
SPGAN [2]	Transfer	46.9%	62.6%	68.5%	26.4%
TJ-AIDL [26]	Transfer	44.3%	-	-	23.0%
BUC [18]	ImageNet	40.4%	52.5%	58.2%	22.1%
IDL*	ImageNet	25.3%	40.4%	46.9%	11.2%
IDL+ACL*	ImageNet	43.9%	59.6%	66.1%	23.7%
IDL+ACL+CBAM*	ImageNet	46.0%	62.4%	69.1%	26.0%
ACL+GAM*	ImageNet	42.7%	59.6%	66.0%	23.6%
IDL+ACL+GAM*	ImageNet	**47.2%**	**63.8%**	**69.8%**	**28.1%**

* Our approach / implementation.

Table 4: DukeMTMC-reID results (no pretrained weights).

Method	rank-1	rank-5	rank-10	mAP
BUC* [18]	5.1%	9.6%	11.7%	1.6%
BOW [43]	17.1%	28.8%	34.9%	8.3%
UMDL [19]	18.5%	31.4%	37.6%	7.3%
IDL*	10.6%	22.2%	28.7%	4.3%
IDL+ACL*	30.6%	47.4%	53.7%	17.0%
IDL+ACL+CBAM*	34.1%	**53.3%**	60.4%	21.2%
IDL+ACL+A-CBAM *	30.4%	48.1%	55.9%	17.7%
ACL+GAM*	28.7%	45.6%	53.2%	15.6%
IDL+ACL+GAM*	**36.2%**	53.2%	**60.9%**	**22.9%**

* Our approach / implementation. A-CBAM - Adjusted CBAM i.e. parameters are controlled to match GAM.

accuracy and mAP for the Market1501 dataset under the challenging fully unsupervised scenario (i.e., no pre-trained weights are used). The table shows that our proposed framework achieved a 36.3% and 26.4% improvement over [18] in rank-1 accuracy and mAP. Notice that the difference in performance between Tables 1 and 2 for BUC [18] is 51.2% for rank-1 accuracy and 26.5% for mAP. Whereas, the difference for our proposed framework is significantly reduced, achieving a separation of 10.3% in rank-1 accuracy and 6.2% in mAP. Thus, our approach minimizes the gap between unsupervised learning with and without discriminative pre-training.

Similarly, for the DukeMTMC-reID dataset, we obtained the best performance among all the compared methods with rank-1 of 47.2% and mAP of 28.1% (Table 3). For the fully unsupervised scenario shown in Table 4, we achieved a rank-1 performance of 36.2% and mAP of 22.9% which surpasses the state-of-the-art by 31.1% and 21.3%, respectively. Moreover, our methods also outperform all methods that leverage supplementary datasets or one shot labeled examples.

5.4. Qualitative Analysis

To evaluate our algorithm, t-distributed stochastic neighbour embedding (t-SNE) [21] is used to visualizes representations from BUC [18] and our proposed framework using the same data and perplexity to compare cluster quality. We pick 18 specific subjects that have hard positives and hard negatives. Figure 4 shows that our approach has better separability and structure of clusters compared with [18], which agrees with the quantitative results from (section 5.3).

5.5. Ablation Studies

Next, we summarize the results of a few ablation studies that consider the effects of GAM, embedding dimensionality, and number filter groups.

Attentions Maps: We compare the activation maps at the final residual block for BUC [18], CBAM [28], and our proposed GAM. Figure 5 shows that the lacks of attention modeling prevents BUC from capturing the important details which should contribute to the representation. CBAM mitigates this problem but still has distracting activations associated with clutter rather than Re-IDs (*e.g.* top of images). Our proposed GAM is able to minimize activations due to clutter and provide enhanced attention for Re-ID.

Embedding Size: We compare the feature embedding size on the DukeMTMC-reID dataset. The rank-1 accuracy seems to increase as embedding size increases (to a point). However, we need compact representations especially in applications like Re-ID. Although Table 5 shows that a 4096 dimensional embeddings achieves the best performances, either 1024 or 2048 dimensional embeddings may better balance performance and efficiency. Making the embedding size even larger than 4096 dimensions had no significant effect on the accuracy.

Table 5: Results on the DukeMTMC-reID dataset with varying embedding sizes without pre-trained weights

Feature Size	rank-1	rank-5	rank-10	mAP
512	35.2%	52.5%	58.8%	22.4%
1024	35.6%	53.3%	61.9%	22.8%
2048	36.2%	53.2%	60.9%	22.9%
4096	**38.9%**	**56.5%**	**63.1%**	**24.4%**

Number of Filter Groups: Table 6 compares fully unsupervised Re-ID performance when varying the number of filter groups for GAM on the Market1501 dataset. Increasing the number of filter groups led to better performance while significantly reducing parameters. Notice that group convolutions with 8 filter groups seemed to achieve the best rank-1 accuracy and mAP. However, increasing beyond 8 filter groups degraded performance, which is likely due to
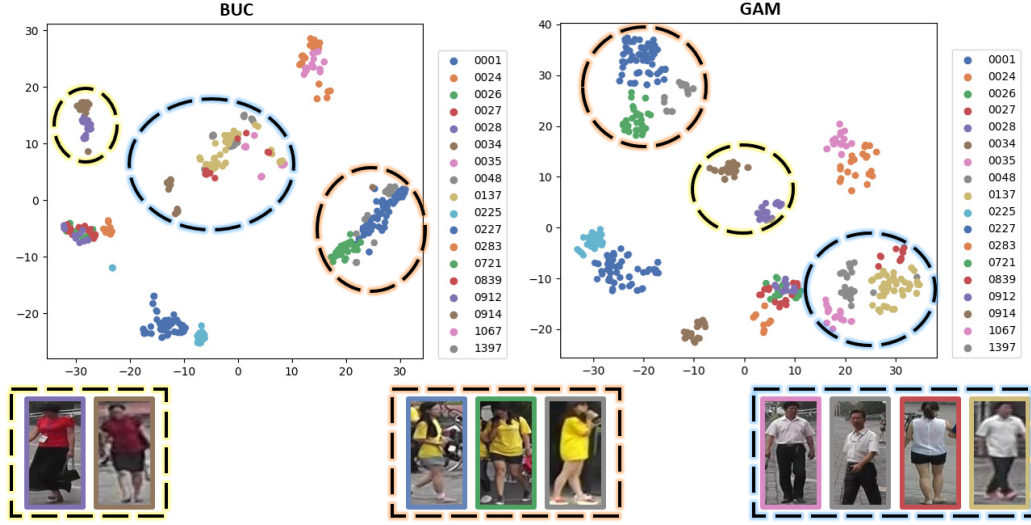
Figure 4: Examples of similar subjects that are closely clustered using t-SNE visualization are circled to show how our framework provides increased separability between different subjects over BUC [18].
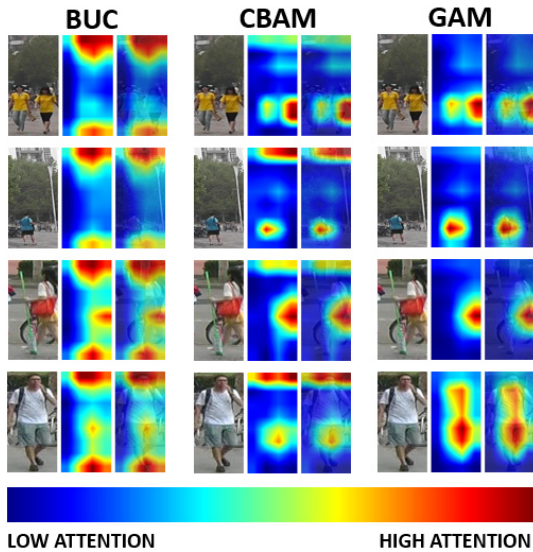

Figure 5: Comparison between BUC, CBAM, and GAM attention maps, with original image (left), attention map (middle), superimposed attention map and image (right).

eliminating too many parameters.

Table 6: Results on the Market1501 dataset with varying number of filter groups without pre-trained weights

Filter Groups	rank-1	rank-5	rank-10	mAP
2	51.1%	69.2%	76.0%	26.3%
4	53.6%	71.9%	**79.4%**	29.5%
8	**55.0%**	**72.5%**	79.2%	**30.0%**
16	54.4%	71.2%	78.9%	28.6%

6. Conclusion

Unsupervised person Re-ID is especially difficult under significant variations in viewpoint, resolution, compression, illumination, and occlusion. In this paper, we studied the effects of group based attention, instance discrimination, and agglomerative clustering, where we demonstrated that our framework composed of all three provided state-of-the-art Re-ID performance on the Market1501 and DukeMTMC-reID datasets. Specifically, we showed that whether or not the network is pre-trained on ImageNet, our proposed framework provided the best Re-ID performance. Most importantly, we showed that by providing a more discriminative unsupervised method without initializing with pre-trained weights, we significantly reduced the gap between supervised and unsupervised methods from greater than 80% down to around 40% rank-1 accuracy for the Market1501 and DukeMTMC-reID datasets. The impacts of closing this gap are that (1) novel architectures (not pre-trained on ImageNet or similar datasets) may be more efficiently developed, without having to pre-train every network modification, and (2) new large-scale data collections may be simplified by reducing the amount of laborious, time consuming, and cost prohibitive data annotation process.

References

[1] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[2] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification, 2017.

[3] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks, 2014.

[4] Hehe Fan, Liang Zheng, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning, 2017.

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

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[7] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person Re-Identification*. Springer Publishing Company, Incorporated, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[10] Martin Hirzer, Peter M. Roth, Martin Köstinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 780–793, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[11] Yani Ioannou, Duncan Robertson, Roberto Cipolla, and Antonio Criminisi. Deep roots: Improving cnn efficiency with hierarchical filter groups, 2016.

[12] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person re-identification by unsupervised ℓ_1 graph learning. volume 9905, pages 178–195, 10 2016.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification, 2018.

[15] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification, 2018.

[16] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan 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*, pages 2197–2206, 2015.

[17] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3685–3693, 2015.

[18] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019.

[19] Jianming Lv, Weihang Chen, Qing Li, and Can Yang. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns, 2018.

[20] Bingpeng Ma, Yu Su, and Frederic Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, 32(6-7):379–390, 2014.

[21] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[22] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[23] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[24] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, Dec. 2010.

[25] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification, 2017.

[26] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification, 2018.

[27] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 688–703, Cham, 2014. Springer International Publishing.

[28] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module, 2018.

[29] Di Wu, Si-Jia Zheng, Xiao-Ping Zhang, Chang-An Yuan, Fei Cheng, Yang Zhao, Yong-Jun Lin, Zhong-Qiu Zhao, Yong-Li Jiang, and De-Shuang Huang. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, 337:354 – 371, 2019.

[30] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018.

[31] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search, 2016.

[32] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *European conference on computer vision*, pages 1–16. Springer, 2014.

[33] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification, 2018.

[34] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z. Li. Salient color names for person re-identification. In David Fleet, Tomas Pajdla, Bernt Schiele,

and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 536–551, Cham, 2014. Springer International Publishing.

[35] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, Jun Chen, and Jun Liu. Specific person retrieval via incomplete text description. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, page 547–550, New York, NY, USA, 2015. Association for Computing Machinery.

[36] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook, 2020.

[37] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature, 2019.

[38] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Unsupervised person re-identification by deep asymmetric metric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):956–973, Apr 2020.

[39] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, 2016.

[41] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification, 2017.

[42] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[43] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[44] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. *ArXiv*, abs/1610.02984, 2016.

[45] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J. Radke. Re-identification with consistent attentive siamese networks, 2018.

[46] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification, 2019.

[47] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[48] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification, 2019.