Learning Global Average Attention Pooling (GAAP) on Resnet50 Backbone for Person Re-identification Problem

Syamala Kanchimani, Maloji Suman, P. V. V. Kishore
Department of Electronics and Communication Engineering,
Koneru Lakshmaiah Education Foundation,
Vaddeswaram, India

***Abstract*—Person re-identification has been an extremely challenging task in computer vision which has been seen as a success with deep learning approaches. Despite successful models, there are gaps in the form of unbalanced labels, poor resolution, uncertain bounding box annotations, occlusions, and unlabelled datasets. Previous methods applied deep learning approaches based on feature representation, metric learning, and ranking optimization. In this work, we propose Global Average Attention Pooling (GAAP) on Resnet50 applied on four benchmark Re-ID datasets for classification tasks. We also perform an extensive evaluation on the proposed Attention module with different deep learning pipelines as backbone architecture. The four benchmark person Re-ID datasets used is Market-1501, RAiD, Partial-iLIDS, and RPIfield. We computed cumulative matching characteristics (CMC) and mean Average Precision (mAP) as the performance evaluation parameters of the proposed against the state of the art. The results obtained have shown that the added attention layer has improved the overall recognition precision over the baselines.**

***Keywords*—Person re-identification; attention network; ResNet50; global average attention**

I. Introduction

The goal of person re-identification (Pe-reID) is to identify and fetch a random person across mutually exclusive camera sources [1]. The objective of Pe-reID models is to determine whether a given query person has reappeared in the frame at a different point in time or in any other camera source at the same point in time [2]. The given query of the person can be an image [3], video [4] and also in text format describing some attribute of the query person [5]. The application range of Pe-reID spans surveillance systems with intelligence that can provide automated feedback on people's movements in real-time.

The Pe-reID pipelines are made up of 5 different tasks. They are arranged chronologically as data collection, bounding box creation, training data annotations, model design and person re-identification. According to [6], the above steps are being considered as a closed world Pe-reID system where the data is structured effectively during preparation. Whereas the open-world Pe-reID system will operate on raw datasets with no annotations and labels. Specifically, this work uses the closed world approach to the Pe-reID problem. The closed world setting is based on the following conditions:

1) Single modality video or image data has been used.
2) The annotations are fixed with persons in bounding boxes with same area identities.
3) The query person is extracted from the training data.
4) Finally, there is enough training data from annotations for supervised learning of person re-identification.

The above processes require expertise and domain knowledge for transforming a video surveillance security problem into a challenging person re-identification problem.

Pe-reID is still considered a super constrained problem. This is due to multiple challenges such as background clutter, low image resolution, poor image quality, partial occlusions, uneven bounding box annotations, human-object interactions [7], etc. Preliminary investigations on Pe-reID problems focused on the hand-crafted feature extraction methods such as Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP). Similar Works considered body reconstruction models and distance metric learning [8]. In the last decade, the growth of AI-based approaches has captured the Pe-reID problem and has subsequently proved its potential with exceptionally good recognition accuracies across datasets [9]. However, the results obtained were nowhere close to the requirements of a real-time deployment pipeline.

In this work, we propose to redesign the regular deep learning approaches with additional layers to learn focused attention. The proposed attention model is called the Global Average Attention Pooling (GAAP) network. The GAAP learns by averaging the features extracted from previous convolutional layers in query, key and value channels. The output of the GAAP network is used to select features that contribute to making correct decisions about a given query person. The attention block has been created based on the non-local attention technique from [2] and the global average pooling is initiated on the attention features to generate a maximally discriminating learnable feature representation.

The proposed GAAP layers or block is integrated with the existing benchmark deep networks such as VGG, ResNet and Inception Net. All these models are trained from scratch on four different types of Pe-reID image datasets. The evaluation metrics used are Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP). The proposed GAAP block is integrated at the last layer of the backbone networks as against the previous works where it was added in the primary layers. Experiments are designed to test and evaluate

the GAAP block's performance in the identification of people in the test data.

The proposed GAAP integrated framework differs in three major areas from the existing baselines:

1) The addition of an attention-based GAAP layer after the convolutional layers in the backbone network ensures a maximally discriminating of learnable features for the dense layers.
2) The global average pooling is performed to mobilize feature space obtained from a large set of convolutional filters into a singular feature representation.
3) An attempt is made to validate the proposed GAAP integrated deep learning framework with cross dataset testing.

To develop the proposed GAAP block for Pe-reID problem, we propose the following objectives:

1) To restructure the datasets into training, testing and validation sets as input to the backbone networks with an integrated GAAP block.
2) To train the deep learning pipelines with an integrated GAAP layer and evaluate their performance.
3) To construct experiments for validating the proposed model.

The rest of the paper is organized as follows. The next section presents previous research performed on person re-identification problems. The third section gives an elaborated discussion on the construction, training and testing of the integrated GAAP network with backbone architectures. Experiments were built for validation and the obtained results are analyzed expensively for characterizing the model pipelines for Pe-reID. Finally, Section 5 concludes on the attained research outcomes with future direction.

II. Literature Review

Remarkably, the closed world Pe-reID problem [6] has been acknowledged previously through the following compositions: 1) Feature representational learning, 2) metric learning and 3) ranking optimization. In this section, we discuss the above approaches and their underlying research findings with their capabilities for providing better Pe-reID solutions. Feature representation learning has been implemented with three derivatives such as global, local and auxiliary. The global features represent the whole person's image for learning [10]. Contrastingly, the local feature learning approaches use parts of the image as input to the extraction algorithm [11]. However, in auxiliary feature representation, data generation models such as Generative Adversarial Networks (GAN) are used to learn variations in the existing datasets. The features of the entire person image are learned in the global feature representation model. These models used powerful deep learning architectures as a classifier for the identification of persons. A set of highly discriminating features were captured with single image representation and cross image representation on triplet loss embedding [12]. The other most popular Pe-reID models treated it as a multi-class classification problem [13] and multi-scale representation problem [14]. Though global feature representation learning has leveraged its full potential for giving good accuracies, it suffered from overfitting problems.

The overfitting problem occurred in global feature representation as the network learned mostly the background information rather than focusing on the person of interest. The global features also have an image misalignment problem that is induced because of the multiple views and person orientations in the training images. Part-based local feature representation has been proposed to overcome the misalignment and overfitting problems in global features. Two mechanisms were formulated in the form of pose-based [15] and rough horizontal-based [16] body part detection for training. In the automated body part detection models, the full body and part features were fused together for classification. Especially, part base local feature representation models such as multi-channel aggression [13], multi-scale context-aware convolutions [14], multi-stage feature decomposition [2], and bilinear pooling [3] have shown expedient performance. Moreover, the performance has been enhanced further by pose-driven [17] and pose-guided matching [18] methods. However, in horizontally divided part-based models, the part-based conventional baseline [18] has served as a building platform for part-based Siamese long short-term memory (LSTM) networks [3]. Other highly accurate models such as Interaction and Aggregated (IA) [19], and second-order nonlocal attention [3] have used reinforced feature learning approaches. The local feature representation learning approaches are limited by the use of noisy pose estimations and large background clutter. Some of the problems associated with both global and part-based feature learning models were maneuvered efficiently by using additional attributes in the training data. These additional attributes are generative datasets [3], semantic representations [20], viewpoint data [17] and data augmentation [18]. The above auxiliary features are found to provide additional data samples for training, which greatly enhanced their ability to identify persons. However, these auxiliaries are computationally expensive and required an additional pre-processing stage for input pairing.

Apart from image-based Pe-reID methods, some recent works have used video-based inputs to relocate a person in the multi-view video frames. Though the video representation has more information in the form of both spatial and temporal data, they fail to capture them accurately due to the unpredictable nature of the persons appearing in the video sequences. Predominantly, recurrent neural networks (RNN) were proposed to capture the temporal information [21] with a temporal pooling layer at the end of RNN. Mixed attributes of spatial and temporal information using sequential fusion are used to enumerate the frame-level feature representations for improved recognition [22]. A varying length video sequence is considered challenging in most video-based applications. In [19], long video sequences are divided into tiny snippets and are ranked in descending order to learn the compact embedding from the top – K segments.

In most of the works on Pe-reID, the backbone architecture is similar to that of the standard ones used for image classification tasks such as VGG-16 and ResNet50. Few works on Pe-reID have modified the ResNet backbone by introducing size 1 in the last convolutional layer or by adding adaptive average pooling in the last pooling layer [23]. However, a tremendous amount of design time can be curtailed by adopting AutoML models for Pe-reID as shown in [20]. The primary objective of all the above-discussed models is to improve

the identification accuracy of the person. One such model which has improved the performance of the Pe-reID deep learning methods is deep metric learning (DML) [24]. The DML uses a metric loss function to calculate the distance between the features from within the class and between classes during training for generating a maximally discriminant feature vector for classification in the dense layers. The identity loss has been widely studied in multiple Pe-reId methods than any other models. The other type of DML model that has indeed improved the performance of the Pe-reID is triplet loss embedding, which starts by computing the distance between the positive class pairs and negative class pairs. The learning is initiated by maximizing the distance between the negative pairs and minimizing it between the positive pairs. The only shortcoming is during the pre-processing stage where the pairing process is performed between the samples from within the class and across classes. Moreover, this pairing complexity increases with the increase in the number of samples per class or an increase in the number of classes itself. Similar to the above DML model, ranking optimization has been shown to improve retrieval efficiency during the testing phase [25]. Very recently, attention-based models [2] have been shown to further strengthen the efficiency of Pe-reID models.

In this work, we propose a global average attention pooling (GAAP) layer at the end of the convolutional layers in ResNet backbone architectures and evaluate its performance against state-of-the-art models. We evaluate the importance of the proposed GAAP against various attention models and across two popular backbone architectures VGG and ResNet. Finally, we conclude by reasoning the significance of the GAAP layer in Pe-reID implementation through experimentation.

III. Methodology: Rank View Triplet Loss Embedding

Learning in Pe-reID is accomplished with D data samples $X_{Pe-reID} = \{x_i, y_i\}\ \forall\ i = 1\ to\ D$ with the goal of finding a mapping function between input x_i and their corresponding labels y_i. The objective of the Pe-reID deep learning neural networks is to learn a mapping function $\theta : X_{Pe-reID} \mapsto F$ that transforms the combined feature space $X_{Pe-reID}$ into F, in which the samples are highly discriminative. Given a set of test images T_{Pe_reID}, the learned mapping function θ will try to project the test images into constituent labels. The testing images and training images in this case are totally nonoverlapping. The primary challenge in the above model is in the learning process of the mapping function θ which in the previous works has been a simple image classifier. The mapping function in case of Pe-reID has to adapt to varying and insufficient training samples per class which results in inconsistent loss parameters during the training process. To regularize the loss function during training we propose to induce an attention layer with global averaging pooling as an architectural upgrade to the existing 50-layer ResNet model. In this subsection, we present the complete Global Average Attention Pooling (GAAP) network and deconstruct the entire pipeline implementation in tensorflow2.3.

A. GAAP Architecture

Global Average Attention Pooling (GAAP) is a network built on top of the existing ResNet50 model with an additional 4 layers. The first three layers in GAAP are convolutional layers, and the last one is a pooling layer. The overall GAAP model for Pe-reID has been illustrated in Fig. 1. A multiscale architecture such as ResNet50 is being used as the backbone network as it has become the state-of-the-art model in many previous works [14]. The proposed attention model has effectively shown to fuse the low level and high-level features to isolate the features of importance within a class label. The attention features are further averaged globally to regenerate a highly discriminative feature map for a particular class label. The model is end–to–end trained on the classification loss function only, which is categorical cross-entropy.

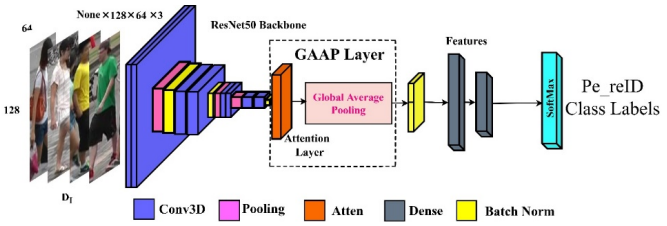

Fig. 1. Global Average Attention Pooling Network Illustration for Person Re-Identification.

The mark of a good person reidentification model is to retrieve accurately the query person image that closely matches the sample images in the class label. However, the previous models used different loss functions to attain close matches between the quey and training images. The two most commonly used loss functions are contrastive and triplet loss. The implementation of these loss functions requires enormous computation resources for training. The proposed GAAP attention framework adds only four layers to the existing network and therefore occupies uses comparatively lesser computational resources. Moreover, the total parameters of the GAAP architecture are lesser than the metric learning models.

The model in Fig. 1 takes input images in training data and transform them into features. The backbone network is Resnet with 50 layers with skip connections.The appearance images $A_n\,(x)\ \forall\ n = 1\ to\ N$ is divided into a length of N samples per class, the appearance sequence is a multidimensional tensor represented as $\mathrm{A} \in R^{r \times c \times 3}$. Here, (r, c) are RGB image height and width in three color channels. Since the GPU capacity is 8GB, the images are standardized to 128×64×3 across all datasets. This becomes the input to the RGB appearance stream $S_A \rightarrow A_n\,(None, x, y, c)$. The S_A stream is made from backbone ResNet consisting of multiple convolutional, maximum pooling with rectified linear activations and batch normalization layers. There is no padding in convolutional layers. This S_A stream will extract features from A using the trainable parameters Θ_{S_A} by optimizing the loss function L_{S_A} on the entire dataset

$$\Theta_{S_A} = \arg\min_{\Theta_{S_A}} L_{S_A}\left(\Theta_{S_A} : \mathrm{A}\,(x), y\right) \tag{1}$$

Here y denotes the class labels. The S_A stream is optimized using the categorical cross-entropy loss L_{S_A} defined as

$$L_{S_A} = -\sum_{i=1}^{C} \left(y_i \times \log\left(y_i\right) + \left(1 - y_i\right) \times \log\left(1 - y_i\right)\right) \tag{2}$$

The trained model M(Θ_{S_A}) will output at the end of i^{th} convolutional layer with j^{th} appearance feature map by using the expression

$$F_A^{ij}(x,y) = f_a\left(\sum_p \sum_{n=0}^{r-1}\sum_{m=0}^{c-1}\left(W_{ijp}^{nm} * \mathrm{A}_{(i-1)p}(x+n, y+m)\right) + b_{ij}\right) \quad (3)$$

Where, A is the person image and f_a is the activation function. W_{ijp}^{nm} are the weights at position (n,m) associated with p^{th} feature map in the $(i-1)^{th}$ layer of the CNN ResNet50 network. The parameter b_{ij} is the bias associated with each of the neurons. Eq'n (3) depicts the convolutional operation between the images and the weight matrix, which is updated sequentially during training of the network. The output RGB appearance features has the dimension $F_{\mathrm{A}} \in R^{r_j \times c_j \times 3 \times C}$. Here, C is the channels or filter kernels applied in j^{th} convolutional layer. These features are further processed using the attention layers before being applied to the dense layers for classification.

B. Attention Layers and Global Average Pooling (GAAP Attention Module)

The proposed attention layers are shown in Fig. 2. The proposed model is inspired by self-attention in [11]. It consists of four 1×1 convolutional layers with stride 1 and one residual connection to preserve the original feature encodings. The dot product enhances the features that are important and discards the others that are least useful in the decision process. This allows the features to concentrate on the areas of the pixels that are highly discriminative in nature. The difference between self-attention in [12] and the proposed in Fig. 2 is that the latter takes input from different features within the class for computing the attention maps. Contrastingly, the self-attention model uses the same features of a single sample to calculate the attention map. The attention map in out proposed model is calculated between the $F_A^i(x,y)$of i^{th} feature of an image A_i in a class and the $F_A^j(x,y)$ of the j^{th} feature of the same image in the class. These features are obtained from the learned backbone network. This enables the network to learn similar appearances across the same image with different features computed using the learned filters in the feature mapping network. The proposed cross-feature attention (CFA) is defined as

$$CFA(f_i) = soft\max\left(\frac{\alpha Q_i.K_i^T + (1-\alpha)\,Q_j K_j^T}{\sqrt{d_k}}\right).V_i \quad (4)$$

Where $i,j \in 1,...,J$, with J is the number of filters in the convolutional layers and $\alpha = 0.5$ is the set hyperparameter for all the layers. The (Q,K,V) are the query, key and value as three convolutional layers in the Fig. 2.

The dot product enhances the features that are important and discards the others that are The weighted sum of features are obtained from all possible positions using the following learning model.

$$a_i = W_a \times \theta\left(f_A^i\right) + f_A^i \quad (5)$$

Where, a_i is the attention maps obtained from the learned W_a with parameters θ of the network. The $+f_A^i$ is the residual connection. Finally, to capture the domain specific features, we apply global average pooling instead of maximum pooling used regularly. The global average pooling of attention features is formulated as

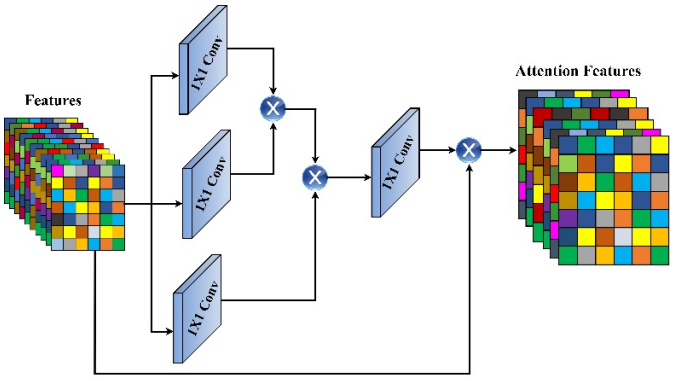

Fig. 2. Attention Layers and its Architecture.

$$f_{ga} = [f_1, f_2, ..., f_K]^T = \left(\frac{1}{|F_k|}\sum_{f_i \in F_k} f_i\right) \quad (6)$$

Where, F_K is the total number of features in the feature maps with K features. f_k represents feature maps which are learned by the backbone and attention layers in the PereID pipeline using the backpropagation algorithm. Finally, the backbone Resnet50 is presented in Fig. 3.

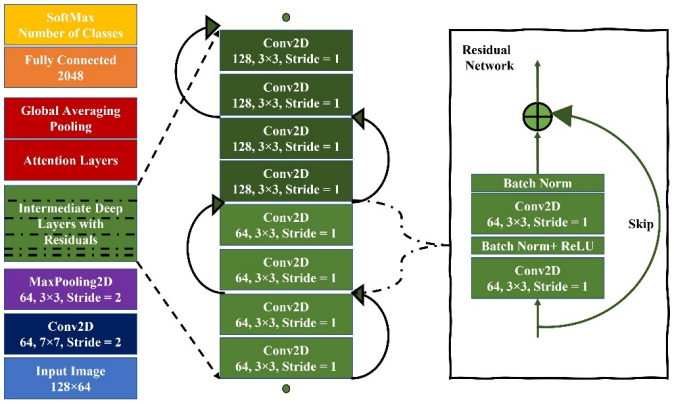

Fig. 3. The ResNet50 Architecture used as Backbone Network in the Proposed GAAP Model

Finally, the combined loss of the entire GAAP network is computed as

$$L_{GAAP} = \sum_{i=1}^{itr} L_{S_A}\left(\Theta\left(A_i(x);\theta\right), y_i\right) \quad (7)$$

Where, Θ is the trainable parameters of Resnet50 and θ are the learnable parameters of the attention network. The final feature representation is learned by minimizing the categorical cross entropy function L_{S_A} for classification. During testing only $A_i(x)$ are used for inferencing the trained model.

IV. Results and Discussion

The four benchmark person Re-ID datasets used in this work are Market-1501, RAiD, Partial-iLIDS and RPIfield. We evaluated the performance of the proposed method using the following parameters, computed cumulative matching characteristics (CMC) and mean Average Precision (mAP). This section gives details about the benchmark person re-identification datasets used for evaluation, the model configuration set for training and testing with in-depth assessment of attention framework.

A. Datasets for Pe-reID

This work has conducted extensive experimentation on four popular benchmark datasets, Market-1501 [1], RAiD [26], Partial-iLIDS [16] and RPIfield [27]. Market-1501 is the largest person re-identification image dataset containing 1501 identities captured with six camera angles and 32,668 bounding box persons that are annotated with deformable part model pedestrian detector. An average of 3.6 images are obtained per person per viewpoint. The training and testing sets have 750 and 751 classes respectively with 3368 additional query images. In this work, we train the model with 750 image classes and test with 750 classes. The training set is split into 15% validation. The image resolution is 128×64. Few training samples of the Market – 1501 are shown in Fig. 4. RAiD is developed in 2014 which has multiple person trajectories recorded using four static camera views. The data is primarily focused on persons on sidewalks and crosswalks. The images in the dataset appear cleaner in the background when compared to the other datasets used in this work. The RAiD dataset has 43 classes with 6290 image samples that are split into 0.7:0.15:0.15 for training, validation and testing. The image resolution is 128×64. Partial iLIDS has occluded person re-identification samples from 476 images from 119 classes. It contains four camera views with a varying resolution of the hand cropped images from the surveillance video data. However, the proposed work has set the resolution of all the images in the dataset as $128\times64\times3$. The occlusions in the images are due to another person or luggage. The RPIfield is constituted in 2018 with 112 class identities with 12 non-overlapping camera viewpoints having 601581 samples is being shown in Fig. 4. The images are annotated using fast pyramid features for bounding box detection which is the reason for multiple resolutions across the dataset. However, the proposed work has normalized the use of image resolution to 128×64 across all the datasets and subsequently across the models used in this work.

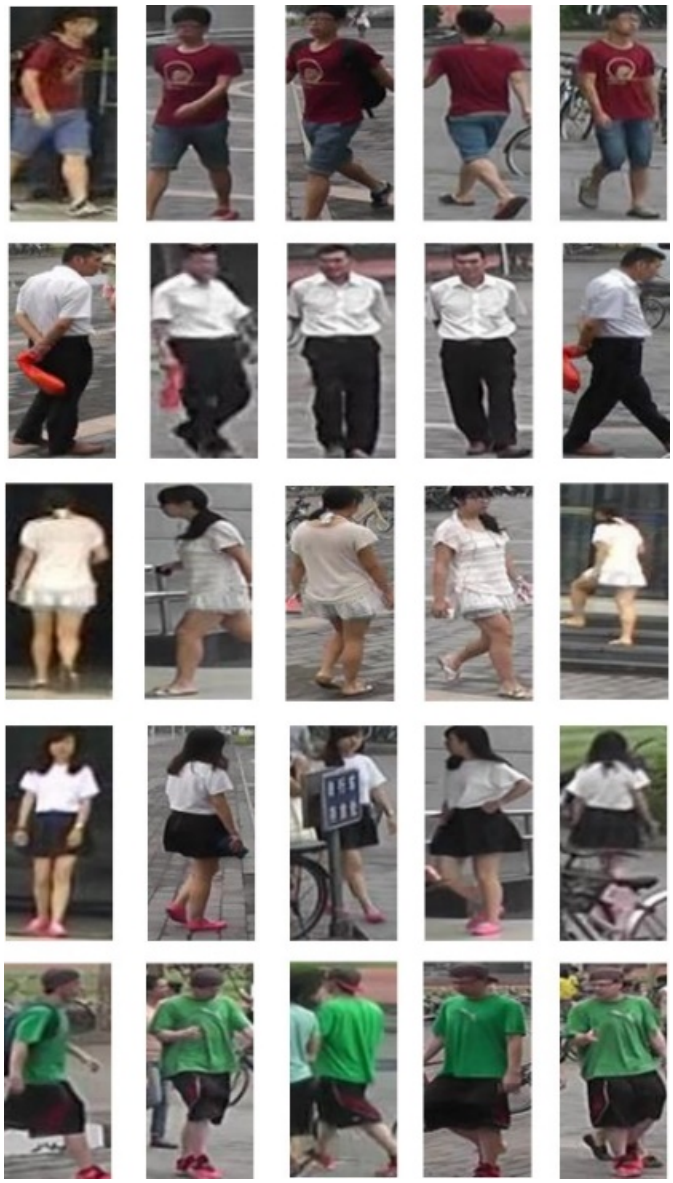

Fig. 4. Samples from Market – 1501 Dataset.

B. Model Configuration

For feature extraction we selected three Resent models as the backbone for feature extraction. The first was tiny ResNet-18 with the attention layers added before the dense layers. This model was used to train the model in a lightweight configuration for real time implementation. The feature extraction process was handled with the help of 8 convolution layers in ResNet-18. Similarly, we applied ResNet-34 with 32 and ResNet-50 with 48 convolutional layers each for feature extraction respectively. The ResNet-50 is deep with added 1×1 convolutions to preserve the input features and decrease the dimensionality of the feature vector. We also included the popular VGG – 16 and – 19 models to analyse the real time deployability as these models are highly recommended in this regard. To evaluate our proposed ResNet with attention layers, we adopted categorical cross entropy loss for optimization with Adam optimizer. This has resulted in providing level playing comparisons with the previous works. In the next subsection, evaluation protocols for the proposed model are being formulated.

C. Model Evaluation Protocols

All the backbone networks and the associated layers are trained from scratch on the benchmark datasets used in this work. Weights and biases were initialized using the standard zero mean 0.01 variance gaussian distribution function at the start of every training session. The network learns by updating weights and biases by optimizing the gradient losses that are backpropagated in reverse. The other training initialized hy-

perparameters would be learning rate, activations, momentum factor, frame dimensions, number of epochs, learning rate decay and minimum allowable loss of the trained classifier. The training set in each class is unbalanced in all the datasets and no attempt has been made to normalize the sample images in each class. However, data augmentation has been initiated on each image to increase the size of the dataset. Four types of augmentation were applied in the form of horizontal and vertical shifts, zoom and crop as shown in Fig. 5. The batch size was selected as 32 which means that there will be 32 images per training batch in each episode. The GAAP model with ResNet – 50 backbone was initialized with a learning rate of 0.0001 with a decay of 10% whenever the validation loss became constant for more than 4 epochs. The other backbones of ResNet were initialized with a higher learning rate to compensate for the lesser depth in features. The momentum factor has been considered as 0.8 across all networks and databases. The average number of training epochs were set at 25. Since the structure of the datasets has been unbalanced in the sample size, the image resolution is kept constant at 128×64 along with all other training initializers to maintain balanced evaluation.

Unseen training samples were used for testing the proposed GAAP network. The SoftMax outputs provide a statistical measure of the probability distribution of the test person image that closely matches the labels in the trained class. Here, we evaluate the global average pooling network with attached attention layers and their impact on the overall performance of the network across all datasets. Finally, we compare the proposed GAAP model with other Pe-reID methods and also perform a detailed ablation study to analyse the behaviour of the model under various test loads. All the models were trained and test on 8GB Nvidia A-4000 series with 16GB memory. The implementation has been done in TensorFlow and Keras packages.

D. Evaluation of the Proposed GAAP Model

The evaluation of the proposed method is conducted by calculation of cumulative matching characteristics (CMC) and mean Average Precision (mAP) across the training dataset. We computed single- and five-fold cross testing on all the datasets. We designed five backbone architectures to evaluate the models performance in identification of a person under various circumstances. Table I provides the results on all benchmark datasets with five backbone networks: VGG-16, VGG-19, ResNet-18, ResNet-34 and ResNet-50. The larger versions of ResNet such as ResNet-101 and ResNet-150 were not trained due to the GPU hardware insufficiency.

All the backbone networks were trained on exactly similar protocols as discussed in previous sections. The above table shows that VGG has failed to take advantage of the attention layers attached after the feature extraction convolutional layers. The reason for underperformance by VGG when compared to ResNet is the missing residual connections in the former model. The residual connections make the ResNet models to avoid overfitting and vanishing or exploding gradients problems. As the networks get deeper, the deep layers may sometimes get zero gradients as input which contributes to faulty decision making on the class labels. The success of ResNet – 50 is attributed to the fact that network is made deeper by adding 1×1 convolutions that help increase the feature quality and reduce the dimensionality.

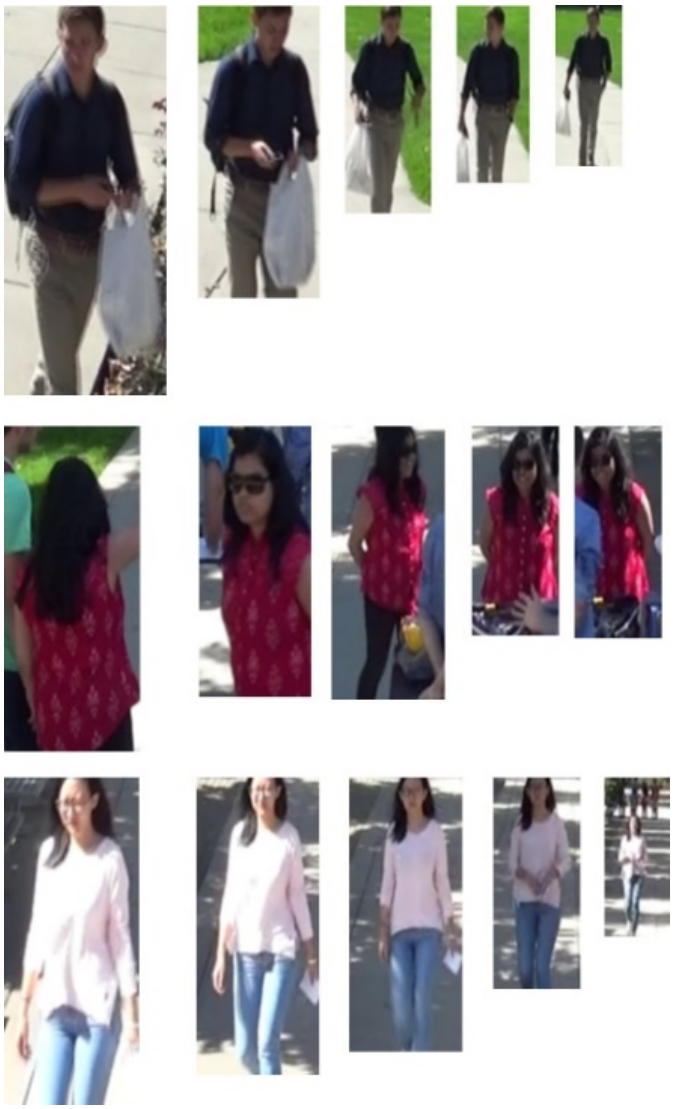

Fig. 5. Samples from RPIfield Pe-reID Dataset.

The results are as expected, and the overall recognition rate mRA with ResNet – 50 was averaged around 91.2% after 5-fold repetition. This is better than the other Pe-reID recognition frameworks in table I by an average of around 10%. The reason lies in the residual connections in ResNet - 50 and added attention module that highlighted the relationships between within-class samples to drive the appearance of the person in multiple cameras. RPIfield dataset has been shown to have maximum test accuracy due to the presence of large training data in all the considered datasets. In the next section, we evaluate the importance of the attention module.

E. Evaluation of Attention Layers

Table II shows the computed parameters on the test data with attention layers and without attention layers. The results show that the there is a 30% increase in network confidence for recognition with the proposed GAAP architecture when

TABLE I. EVALUATION OF THE GAAP MODEL ON BENCHMARK DATASETS WITH DIFFERENT BACKBONE NETWORKS

Backbone Networks Trained in GAAP	Pe-reID Datasets	mRA		CMC	
		1 - Fold	5 - Fold	1 - Fold	5 - Fold
VGG - 16	Market-1501	0.731	0.776	0.727	0.746
	RAiD	0.743	0.788	0.732	0.752
	Partial-iLIDS	0.705	0.696	0.713	0.701
	RPIfield	0.741	0.81	0.764	0.775
VGG - 19	Market-1501	0.716	0.721	0.741	0.713
	RAiD	0.803	0.848	0.799	0.818
	Partial-iLIDS	0.815	0.86	0.804	0.824
	RPIfield	0.777	0.768	0.785	0.773
ResNet - 18	Market-1501	0.813	0.882	0.836	0.847
	RAiD	0.788	0.793	0.813	0.785
	Partial-iLIDS	0.84	0.885	0.836	0.855
	RPIfield	0.852	0.897	0.841	0.861
ResNet - 34	Market-1501	0.814	0.805	0.822	0.81
	RAiD	0.85	0.919	0.873	0.884
	Partial-iLIDS	0.825	0.83	0.85	0.822
	RPIfield	0.83	0.875	0.819	0.839
ResNet - 50	Market-1501	0.842	0.873	0.8	0.828
	RAiD	0.828	0.897	0.851	0.862
	Partial-iLIDS	0.803	0.808	0.828	0.8
	RPIfield	0.867	0.912	0.863	0.882

compared to the traditional models. All the models are trained using the same initial conditions as discussed in the Section 4.3. The big jump in the proposed model is due to the ability of the network to train on the features that are important for classification. The use of attention layers guarantees highly discriminative features within a class label. In the next section, we evaluate the global average pooing of attention features against the traditional maximum pooling model used in previous works.

F. Evaluation of Global Average Pooling

Before evaluating the global average pooling layers in GAAP architecture with backbone CNN models, we present the attention maps obtained on Market – 1501 dataset with ResNet – 50 backbone CNN model in Fig. 6. The figures provide a visual confirmation of the concentration of features used for training the dense layers in the GAAP pipeline for recognition of persons in Market – 1501 dataset. We observed similar kind of results across all other datasets used in this work.

In the following Table III, we computed the performance parameters CMC and mRA for the proposed global average pooling and maximum pooling of attention features after the convolutional layers in all the backbone networks used in this work. The results show that the global average pooling results in an 10 $\pm$2% increase in performance of the backbone network for recognition of persons when compared to the traditional maximum pooling model. In the traditional maximum pooling model, the largest values in the feature space on the batch size within a class label are pooled together for training on the dense layers. This procedure generates outlier features that are not concentrated on the person or object of interest in the entire feature space at all times. Hence, it is intuitive that the final feature space for dense layers may possibly miss some of the prominent features necessary for correct identification. This can be avoided by considering an averaging feature space across a batch size within a class label. Consequently, the global averaged features have shown to exhibit the characteristics of all prominent regions of interest across a batch size given a generalized representation of the

Fig. 6. Data Augmentation Applied on Market – 1501 Dataset.

person across the class label. This is observable in the Fig. 7 visualization of attention regions projected on the original images of Market – 1501 dataset. Finally, we compare our proposed GAAP with ResNet – 50 backbone with the state – of – the – art methods for Pe-reID on the benchmark datasets in the following section.

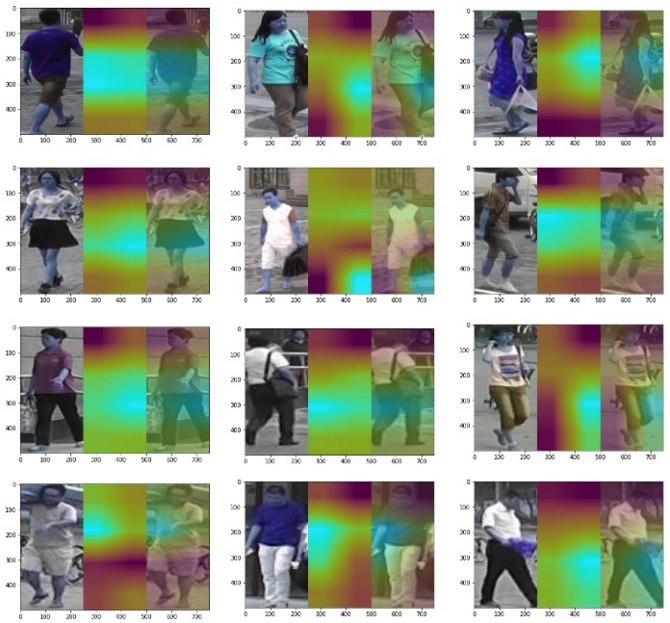

Fig. 7. Attention Maps Obtained from the Proposed GAAP Architecture on ResNet – 50 Backbone CNN Model.

G. Comparison with the State-of-the-Art Pe-reID Methods

This section draws comparisons of different Pe-reID methods against the proposed GAAP architecture. As can be observed from the above analysis that the ResNet – 50 backbone

TABLE II. PERFORMANCE EVALUATION OF THE SELECTED BACKBONE NETWORKS ON PE-REID DATASETS WITH AND WITHOUT ATTENTION LAYERS

Classifiers	Datasets	With Attention Layers (GAAP)				Without Attention Layers			
		mRA		CMC		mRA		CMC	
		1 - Fold	5 - Fold	1 - Fold	5 - Fold	1 - Fold	5 - Fold	1 - Fold	5 - Fold
VGG - 16	Market-1501	0.731	0.776	0.727	0.746	0.504	0.489	0.497	0.499
	RAiD	0.743	0.788	0.732	0.752	0.546	0.544	0.569	0.539
	Partial-iLIDS	0.705	0.696	0.713	0.701	0.476	0.474	0.445	0.465
	RPIfield	0.741	0.81	0.764	0.775	0.54	0.546	0.574	0.564
VGG - 19	Market-1501	0.716	0.721	0.741	0.713	0.521	0.508	0.51	0.504
	RAiD	0.803	0.848	0.799	0.818	0.576	0.561	0.569	0.571
	Partial-iLIDS	0.815	0.86	0.804	0.824	0.618	0.616	0.641	0.611
	RPIfield	0.777	0.768	0.785	0.773	0.548	0.546	0.517	0.537
ResNet - 18	Market-1501	0.813	0.882	0.836	0.847	0.612	0.618	0.646	0.636
	RAiD	0.788	0.793	0.813	0.785	0.593	0.58	0.582	0.576
	Partial-iLIDS	0.84	0.885	0.836	0.855	0.613	0.598	0.606	0.608
	RPIfield	0.852	0.897	0.841	0.861	0.655	0.653	0.678	0.648
ResNet - 34	Market-1501	0.814	0.805	0.822	0.81	0.585	0.583	0.554	0.574
	RAiD	0.85	0.919	0.873	0.884	0.649	0.655	0.683	0.673
	Partial-iLIDS	0.825	0.83	0.85	0.822	0.63	0.617	0.619	0.613
	RPIfield	0.83	0.875	0.819	0.839	0.633	0.631	0.656	0.626
ResNet - 50	Market-1501	0.842	0.873	0.8	0.828	0.563	0.561	0.532	0.552
	RAiD	0.828	0.897	0.851	0.862	0.627	0.633	0.661	0.651
	Partial-iLIDS	0.803	0.808	0.828	0.8	0.608	0.595	0.597	0.591
	RPIfield	0.867	0.912	0.863	0.882	0.64	0.625	0.633	0.635

TABLE III. COMPARATIVE ANALYSIS OF GLOBAL AVERAGE POOLING AND THE TRADITIONAL MAXIMUM POOLING OF ATTENTION FEATURES FOR PE-REID TASKS

Classifiers	Datasets	With Global Average of Attention Features (GAAP)				With Maximum Pooling of Attention Features			
		mRA		CMC		mRA		CMC	
		1 - Fold	5 - Fold	1 - Fold	5 - Fold	1 - Fold	5 - Fold	1 - Fold	5 - Fold
VGG - 16	Market-1501	0.731	0.776	0.727	0.746	0.628	0.613	0.621	0.623
	RAiD	0.743	0.788	0.732	0.752	0.67	0.668	0.693	0.663
	Partial-iLIDS	0.705	0.696	0.713	0.701	0.6	0.598	0.569	0.589
	RPIfield	0.741	0.81	0.764	0.775	0.664	0.67	0.698	0.688
VGG - 19	Market-1501	0.716	0.721	0.741	0.713	0.645	0.632	0.634	0.628
	RAiD	0.803	0.848	0.799	0.818	0.7	0.685	0.693	0.695
	Partial-iLIDS	0.815	0.86	0.804	0.824	0.742	0.74	0.765	0.735
	RPIfield	0.777	0.768	0.785	0.773	0.672	0.67	0.641	0.661
ResNet - 18	Market-1501	0.813	0.882	0.836	0.847	0.736	0.742	0.77	0.76
	RAiD	0.788	0.793	0.813	0.785	0.717	0.704	0.706	0.7
	Partial-iLIDS	0.84	0.885	0.836	0.855	0.737	0.722	0.73	0.732
	RPIfield	0.852	0.897	0.841	0.861	0.779	0.777	0.802	0.772
ResNet - 34	Market-1501	0.814	0.805	0.822	0.81	0.709	0.707	0.678	0.698
	RAiD	0.85	0.919	0.873	0.884	0.773	0.779	0.807	0.797
	Partial-iLIDS	0.825	0.83	0.85	0.822	0.754	0.741	0.743	0.737
	RPIfield	0.83	0.875	0.819	0.839	0.757	0.755	0.78	0.75
ResNet - 50	Market-1501	0.842	0.873	0.8	0.828	0.687	0.685	0.656	0.676
	RAiD	0.828	0.897	0.851	0.862	0.751	0.757	0.785	0.775
	Partial-iLIDS	0.803	0.808	0.828	0.8	0.732	0.719	0.721	0.715
	RPIfield	0.867	0.912	0.863	0.882	0.764	0.749	0.757	0.759

has shown better performance when compared to other four models. Table IV records the performance of the models on benchmark datasets. All the models were trained from scratch on the same 8GB GPU with 16GB memory under similar initial conditions, except for the learning rate which has been selected differently to avoid overfitting. The stopping criteria is set as the flat validation error for more than 5 epochs and after two times decrease in learning rate.

The works in table IV are based on supervised and unsupervised methods that have clocked maximum mRA and CMC in the literature. We also compared with attention-based methods like AGW and the results show that the GAAP has indeed performed better than the AGW. The proposed GAAP has attention layers at the end of convolutional networks which enables the model to generate attention features for dense net classifier. However, the past attention-based methods used attention inside the convolutional layers that failed to capture the essential focused features for classification. We also found that the proposed model trains faster than the previous most popular triplet loss embedding with ResNet -50 backbone network. Fig. 8 and 9 shows the training accuracies and loss plots on Market – 1501 dataset for GAAP and DML with triplet loss respectively. Overall, our proposed GAAP model have shown good performance on RIPfiled Pe-reID dataset due to its rich multi view and multi resolution representation of the person images. Finally, the average recognition on all the benchmark datasets is around 84.12 which is 5% more than the previous methods.

V. CONCLUSION

In this work, we present an attention framework-based solution for person reidentification problem. The attention framework is built at the end of the feature extraction network and before the classifier dense network. Subsequently, the attention features are pooled using global averaging across the within class images. The proposed GAAP network is trained with ResNet – 50 as a backbone architecture for feature extraction. Consequently, extensive experimentation on four bench-

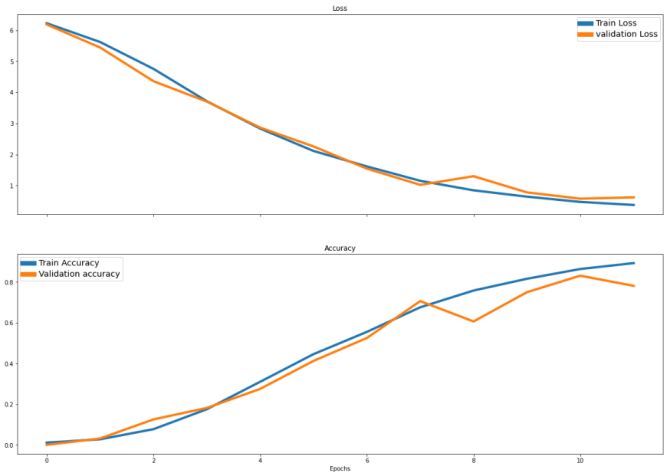

Fig. 8. Training Performance of GAAP with ResNet – 50 Backbone on Market – 1501 Dataset.

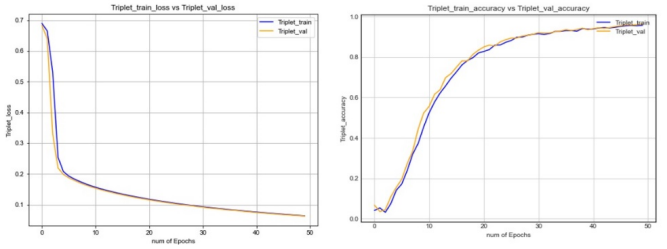

Fig. 9. Training Performance of DML with Triplet Loss Embedding with ResNet – 50 Backbone on Market – 1501 Dataset.

mark person Pe-reID datasets has shown that the proposed model performs better than state-of-the-art. Interestingly, the proposed model generated an average identification accuracy of around 84.12. Also, the proposed GAAP model trains in less time and achieves a competitive average validation accuracy on the benchmark datasets. However, the improvement of performance is achieved on large datasets with heterogeneous properties.

TABLE IV. COMPARISON OF PREVIOUS STATE – OF – THE – ART PE-REID METHODS AGAINST THE PROPOSED GAAP MODEL

Methods	Market - 1501		RAiD		Partial-iLIDS		RPIfield	
	mRA	CMC	mRA	CMC	mRA	CMC	mRA	CMC
PCB [28]	0.812	0.808	0.849	0.949	0.712	0.708	0.749	0.849
MGN [21]	0.832	0.821	0.852	0.952	0.732	0.721	0.752	0.852
HAN [3]	0.842	0.856	0.824	0.924	0.742	0.756	0.724	0.824
BDB [22]	0.803	0.817	0.785	0.885	0.703	0.717	0.685	0.785
IANet [19]	0.817	0.831	0.798	0.898	0.717	0.731	0.698	0.798
BoT [28]	0.824	0.838	0.812	0.912	0.724	0.738	0.712	0.812
AGW [2]	0.838	0.852	0.787	0.887	0.738	0.752	0.687	0.787
FPR [29]	0.823	0.837	0.798	0.898	0.723	0.737	0.698	0.798
PGFA [15]	0.813	0.827	0.843	0.943	0.713	0.727	0.743	0.843
HOReID [7]	0.824	0.838	0.891	0.991	0.724	0.738	0.791	0.891
PVPM [17]	0.822	0.836	0.802	0.902	0.722	0.736	0.702	0.802
GML [25]	0.813	0.827	0.813	0.913	0.713	0.727	0.713	0.813
HCT [30]	0.592	0.606	0.653	0.753	0.492	0.506	0.553	0.653
UDAML[24]	0.654	0.668	0.729	0.829	0.554	0.568	0.629	0.729
TLE [12]	0.713	0.727	0.758	0.858	0.613	0.627	0.658	0.758
MEB-Net [18]	0.752	0.766	0.765	0.865	0.652	0.666	0.665	0.765
PLF [16]	0.723	0.737	0.733	0.833	0.623	0.637	0.633	0.733
GAAP (OURS)	0.842	0.873	0.828	0.897	0.803	0.808	0.867	0.912

ACKNOWLEDGMENT

Author thanks the management of Koneru Lakshmiah Education Foundation and R&D department for funding the project.

REFERENCES

[1] 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*, 2015, pp. 1116–1124.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[3] L. Chen, H. Yang, Q. Xu, and Z. Gao, "Harmonious attention network for person re-identification via complementarity between groups and individuals," *Neurocomputing*, vol. 453, pp. 766–776, 2021.

[4] D. Wu, S.-J. Zheng, X.-P. Zhang, C.-A. Yuan, F. Cheng, Y. Zhao, Y.-J. Lin, Z.-Q. Zhao, Y.-L. Jiang, and D.-S. Huang, "Deep learning-based methods for person re-identification: A comprehensive review," *Neurocomputing*, vol. 337, pp. 354–371, 2019.

[5] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8933–8940.

[6] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 1092–1108, 2019.

[7] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6449–6458.

[8] H.-X. Yu, A. Wu, and W.-S. Zheng, "Unsupervised person re-identification by deep asymmetric metric embedding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 4, pp. 956–973, 2018.

[9] H. Zheng, X. Zhong, W. Huang, K. Jiang, W. Liu, and Z. Wang, "Visible-infrared person re-identification: A comprehensive survey and a new setting," *Electronics*, vol. 11, no. 3, p. 454, 2022.

[10] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1367–1376.

[11] 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*, 2017, pp. 384–393.

[12] Z. Tang and J. Huang, "Harmonious multi-branch network for person re-identification with harder triplet loss," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 4, pp. 1–21, 2022.

[13] J. Yu and H. Oh, "Graph-structure based multi-label prediction and classification for unsupervised person re-identification," *Applied Intelligence*, pp. 1–13, 2022.

[14] Y. Li, L. Liu, L. Zhu, and H. Zhang, "Person re-identification based on multi-scale feature learning," *Knowledge-Based Systems*, vol. 228, p. 107281, 2021.

[15] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 542–551.

[16] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 393–402.

[17] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person reid," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 744–11 752.

[18] Y. Zhai, Q. Ye, S. Lu, M. Jia, R. Ji, and Y. Tian, "Multiple expert brainstorming for domain adaptive person re-identification," in *European Conference on Computer Vision*. Springer, 2020, pp. 594–611.

[19] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9317–9326.

[20] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, and S. Satoh, "Learning sparse and identity-preserved hidden attributes for person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 2013–2025, 2019.

[21] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[22] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch dropblock network for person re-identification and beyond," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3691–3701.

[23] H. Kim, H. Kim, B. Ko, J. Shim, and E. Hwang, "Two-stage person re-identification scheme using cross-input neighborhood differences," *The Journal of Supercomputing*, vol. 78, no. 3, pp. 3356–3373, 2022.

[24] R. Pierre and M. Qi, "Unsupervised domain adaption based on metric learning for person re-identification," in *2021 3rd International Conference on Advances in Computer Technology, Information Science and Communication (CTISC)*. IEEE, 2021, pp. 417–421.

[25] J. Meng, W.-S. Zheng, J.-H. Lai, and L. Wang, "Deep graph metric learning for weakly supervised person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[26] A. Chakraborty, A. Das, and A. K. Roy-Chowdhury, "Network consistent data association," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1859–1871, 2015.

[27] M. Zheng, S. Karanam, and R. J. Radke, "Rpifield: A new dataset for temporally evaluating person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1893–1895.

[28] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[29] L. He, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8450–8459.

[30] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 657–13 665.