IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 July 2023, accepted 11 August 2023, date of publication 17 August 2023, date of current version 23 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3305924

RESEARCH ARTICLE

Dual Attention Network for Unsupervised Domain Adaptive Person Re-Identification

HAIQIN CHEN, HONGYUAN WANG, ZONGYUAN DING, AND PENGHUI LI
School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou 213000, China

Corresponding author: Hongyuan Wang (hywang@cczu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61976028, and in part by the Jiangsu Province Postgraduate Research and Practice Innovation Program under Grant KYCX22_3061.

ABSTRACT Clustering-based unsupervised domain adaptive person re-identification methods reduce much of the annotation cost. However, many pseudo-labels with inaccurate labels are produced when fine-tuning the target domain as a result of the shortcomings of the less than ideal network model and less than excellent clustering. In order to remove the mislabeled pseudo-labels and improve the purity of the network to obtain the correctly labeled pseudo-labels, this paper proposes a dual attention network framework. Specifically, in order to make the network learn to focus on the focal object, this paper introduces the CBAM attention mechanism, which can focus well on the information on the image channel and space, and then identify the correct and incorrect pseudo-labels to improve the overall performance of the network model. In addition, in order to allow the network to extract richer semantic information, this paper introduces Non-local blocks, which can directly capture the remote dependencies between image features by computing the interaction between any two locations. Accordingly, we conducte extensive experiments on four common unsupervised domain adaptive person re-identification tasks, namely, DukeMTMC→Market-1501, Market-1501→DukeMTMC, DukeMTMC→MSMT17, and Market-1501→MSMT17, in which the mAP/R1 and baseline compared to 3.6%/1.4%, 1.8%/2.5%, 4.7%/5.2%, and 3.4%/3.0%, respectively.

INDEX TERMS Clustering, unsupervised domain adaptive, CBAM, non-local block.

I. INTRODUCTION

In order to identify a certain individual from several cameras, person re-identification is used. [1], [2]. The duty is crucial in sectors like metropolitan road traffic and video surveillance [3], and it can be used in situations involving missing children, apprehending criminals, and a variety of other human beings. A great deal of research on deep learning-based person ReID models employs a supervised method, meaning that the training set of data needs to be manually labeled, which takes a lot of time and effort. Unsupervised approaches [4], [5] have gained greater focus and research in order to boost the scalability of models and to be more applicable to real-world applications. These methods have also gradually advanced, with experimental results that are on par with or even better than those of supervised learning techniques. Although the supervised person ReID [6] techniques' great advancements, some practical applications of these techniques are challenging due to their heavy reliance on manual annotation [7]. Additionally, it will be challenging to generalize the trained model to new datasets once the source and target domains diverge in terms of domain. Using UDA person ReID methods, which entail adapting a model developed on a labeled source domain to an unlabeled target domain in order to increase the model's discriminability on the unlabeled target domain, is one way to solve this problem.

Methods based on ranking, domain transfer and clustering are some UDA person ReID techniques [8], [9]. A clustering-based method typically produces higher performance, which generally consists of three stages: (1) pre-training on the labeled source domain; (2) generating pseudo-labels on the unlabeled target domain using clustering algorithms or similarity measures; (3) fine-tuning the model using the target domain image fine-tuning. In this case, the initial step is only carried out a single time, however the remaining two

The associate editor coordinating the review of this manuscript and approving it for publication was Paulo Mendes.

stages are continually carried out in order to enhance each other. Some methods [10] produce global target domain features using just one attribute extractor, which can result in a large number of incorrect pseudo-labels. Different strategies [11] employ a teacher-student structure to progressively learn average-weighted models in order to provide more trustworthy pseudo-labels. Despite producing notable results, these fine-tuning techniques mostly concentrate on obtaining global features that only include broad information about semantics, while disregarding local features that may supply fine-grained detail.

For some UDA person re-identification methods, problems like suboptimal feature embedding, imperfect clustering, and unknown proportion of pseudo-labels with mislabels among the pseudo-labels generated in the target domain can cause misleading feature learning. To address this problem, P2LR [12] proposed a domain-adaptive person re-identification method based on probabilistic uncertainty-guided progressive label refinement. In this paper, based on P2LR, we consider another perspective to improve the network so that the deep neural network can extract richer semantic features, find out more pseudo-labels with mislabeling, and promote the model to be optimized continuously during the training process so as to have better performance.

This paper discusses two key issues: (1) how to instruct the network to put emphasis on particular insights; and (2) how to strengthen the semantic content of the features the network gets. For the first issue, during the training process, the picture data that the network model concentrates on may be inaccurate, incomplete, or too broad in scope. This will ultimately result in the extraction of less-than-optimal features, and using these features for clustering will likely result in a large number of pseudo-labels with noise, deceiving the network's learning process. Therefore, in this paper, a lightweight attention module CBAM [13] is introduced, and with the addition of CBAM, the features cover more parts of the person to be recognized and have a higher chance of eventually discriminating the person. For the second problem, when the network sets up clustering based on global and local features, respectively, it may result in an unmarked sample with a variety of entirely distinct pseudo-labels, which will prevent the model from being capable to easily assign them to the right person during the period of training. Thus, based on this, this paper introduces Non-local [14] blocks, which can be directly incorporated into the global features to provide richer semantic information for the later layers [15].

The following is a brief overview of this paper's major contributions.

(1) This study proposes a dual-attentive network structure and fully tests the suggested approach on three datasets mainly used for person re-identification, with further improvements over baseline results in four typical tasks.

(2) This study offers a lightweight attention module, CBAM, which can focus on more areas of the person and enhance the network's accuracy in recognizing person and enables the network to concentrate on crucial information during training.

(3) For the features that the network captured, to make them represent the person identity more accurately, Non-local operations are introduced in this paper, which will enhance the semantic content of the obtained features.

II. RELATED WORK

A. GENERAL DOMAIN ADAPTATION

Transferring acquired knowledge from a source domain with labels to a target domain is the process of domain adaptation. A ''domain gap'' occurs when there is typically a distinct data distribution between two domains; this degrades the network's performance. The majority of domain adaptation algorithms fall into one of two groups, namely feature-level, and sample-level. For instance, MMD [16] addressed the issue at the feature level by minimizing the inter-domain divergence and maximizing the intra-class density. Asymmetric mapping between domains was introduced by SBADA-GAN [17] to reconstruct the class source target picture from the sample level. According to recent studies, sample-level and feature-level data are equally crucial for unsupervised domain adaptable tasks. Therefore, LPJT [18] proposed combining distribution matching for feature adaptation and landmark selection samples for adaptation. Such experimental results are good, but an abundance of data for training in the target domain makes real-world applications tricky. A fast domain adaptive network, which requires fewer processing resources and yields better accuracy, has been suggested as a solution to this problem. However, the broad domain adaptable technique, which belongs to the same class across domains, is inappropriate for the task since identities in the two domains are distinct in person re-identification. Therefore, it is crucial to design the corresponding algorithm for person re-identification domain adaption.

B. UNSUPERVISED DOMAIN ADAPTIVE PERSON RE-IDENTIFICATION

Three categories can be utilized for categorizing the common unsupervised domain adaptive algorithms. The first class [19] is image-level methods that use GAN to transform the domain of the source image into the desired style for the domain of the target image [20]. For example, PTGAN [21] focused on transferring knowledge. However, the performance of these methods does not compare to that of fully supervised methods. The second class consists of feature-level techniques. For instance, IM [22] explored three different fundamental invariants, including neighborhood, paradigm invariance, and camera. The final class is called clustering-based adaptation, and they all work by using a pre-training phase on the source domain and then transferring the conditions that have been learned to the target domain. Because of the unstable clustering method and excessive domain deviations, the produced pseudo-labels typically incorporate noise, which prevents further model performance improvement. Although MMT [23] is used in UDA person re-identification to fix

this issue by generating soft pseudo-labels using two neural networks, as the training process advances and the two neural networks gradually converge, they will inevitably have a high degree of similarity. Consequently, it is essential to develop various networks that improve complementarity.

C. ATTENTION MECHANISM

For augmented representation learning, attention is frequently used in domains like object detection and image categorization. As an illustration, the SE [24] block updates the channel feature response, and the Convolutional Attention Module (CBAM) [25] further investigates ''what'' and ''where'' using channel attention and spatial attention. Attention modules can be stacked to provide adaptive modules, attention-aware features, and a residual network. The Non-local block utilizes global features as well as explores the connections between various locations on the feature map. Multiple self-attention components are merged to form the multi-head self-attention mechanism. Better feature extraction implies each head that learns attributes via several graphical subspaces to assemble loads reconstructions. The most recent fully supervised algorithms employed in the research of person re-identification, including ConsAtt [26] and ABD-Net [27], applied attention processes to several widely used person re-identification datasets. These studies have employed attention mechanisms to identify features that are either discriminative or essential to raising performance. The research presented in this publication also reveals that the attention mechanism can enhance the variations brought about by wave blocks in addition to the previously indicated features. As a result, by including the attention mechanism, the neural network's extracted features are made to be more complimentary and discriminative, which enhances the performance of the network.

III. PROPOSED METHOD

Adapting a practiced model from the source $D_s = \left\{\left(x_i^s, y_i^s\right) |_{i=1}^{N_s}\right\}$ to the target domain $D = \left\{x_i |_{i=1}^{N}\right\}$ is the aim of UDA person re-identification, where N_s and N signify the quantity of samples in the source domain and the quantity of samples in the target domain, respectively. x_i^s and y_i^s indicate the source domain's samples and labels. Clustering is used to produce the pseudo-labels $\tilde{y}_i$ for the samples x_i in the target domain.

FIGURE 1 shows the flowchart of the general framework proposed in the paper for unsupervised domain adaptive person re-identification. A mutual mean teacher model is first used to build a clustering baseline. On top of the foundational approach, a probabilistic uncertainty-guided asymptotic label refinement method is added to assess the noise level of pseudo-labels and lessen the detrimental effects of noisy samples. In addition, the CBAM attention mechanism [28] and Non-local blocks are introduced to strengthen the network's focus on feature specifics in order to increase its capacity to spot false pseudo-labels. The clustering baseline, probabilistic uncertainty-based modeling, improved network structure, CBAM attention mechanism [29], and Non-local blocks will be introduced in the next sections, respectively.

A. INTRODUCTION OF CLUSTERING BASELINE

In this paper, we build a baseline for domain-adaptive unsupervised person re-identification based on P2LR via clustering. The pre-training of the source domain model, clustering, and fine-tuning of the target domain are the three stages that make up the general clustering methods process. Two relatives networks are set up with the same topology, commenced with two separate randomised seeds, and and trained using identical data during the source pre-training phase, but they are subjected to different data augmentation [30] to lessen interdependence. Identity loss L_{id} and triplet loss L_{tri} are used to better the networks. The pseudo-labels for the target domain data are produced by clustering during the clustering phase. Utilizing the exponential moving average of the two student models from the iteration, two mean teacher models, $\overline{M_1}$ and $\overline{M_2}$, were produced for the target domain fine-tuning phase. Use the predictions of the teacher model as soft labels to oversee the training process of another student model. Along with the identity loss and triplet loss for hard pseudo-labels produced by clustering, each model also used a Kullback-Leibler (KL) divergence loss L_{KL} and a soft triplet loss L_{stri} for soft labels.

The entire loss function for the fine-tuning phase is shown below:

$$L = L_{id} + \lambda_{tri} L_{tri} + \lambda_{KL} L_{KL} + \lambda_{stri} L_{stri} \tag{1}$$

where λ_{tri}, λ_{KL} and λ_{stri} are, respectively, corresponding loss weights.

B. MODELING PROBABILISTIC UNCERTAINTY

The clustering-based UDA person re-identification task generates a significant number of noise-filled pseudo-labels, which may lead to mistakes in the network's training at the step of fine-tuning the target domain and degrade the model's performance. Uncertainty estimation is a simple technique to get rid of erroneous pseudo-labels to lessen the impact of pseudo-label noise. This paper models on the basis of P2LR, the results of which show that samples of pseudo-labels with inaccurate tags have high probability uncertainty while samples of pseudo-labels with accurate tags have low probability uncertainty. Therefore, to evaluate the sample's noise level laterally, this variation in distribution is used as probabilistic uncertainty. Give pseudo-labels $\tilde{y}_i$ to each unlabeled sample x_i in the target domain by clustering. We construct an external classifier $\phi_t\left(\cdot \left| w_t^{cls}\right.\right)$ based on the t-th step clustering, where $w_t^{cls} \in R^{c \times d}$ denotes the external classifier's parameterized weight. It should be noted that the weights are dynamically created and do not require extra training. In the experiments, the classifier's weights are cluster centroids of d dimensions. The classifier as stated in equation (2) is utilized to determine the classification probability distribution among the several identities for the features$f_i \in R^d$ of the sample $(x_i, \tilde{y}_i)$ obtained from the mean

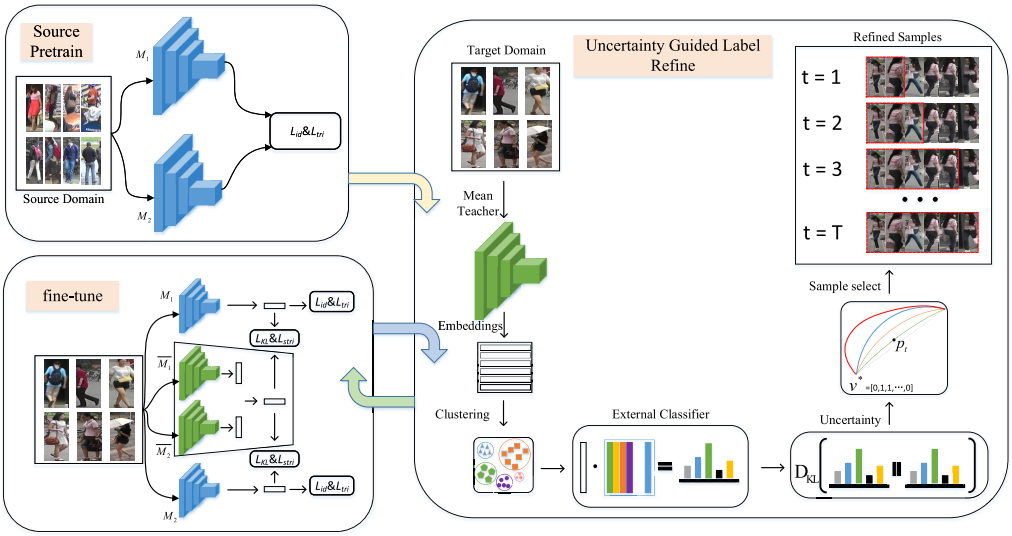

FIGURE 1. The general framework flow chart.

teacher model. α in the equation is the temperature parameter.

$$P(x_i, \tilde{y}_i) = \phi(f_i) = Soft\max\left(\frac{w_t^{cls}}{\|w_t^{cls}\|} \cdot \frac{f_i}{\|f_i\|} \cdot \alpha\right) \quad (2)$$

The modeling of the generalized ideal distribution of the samples in the target domain is motivated on the grounds that the single-pulse distribution of the samples may be derived by ordering the probabilities symmetrically from biggest to smallest centered at y_i^s in the source domain. The smoothed δ distribution linked to high temperatures α is shown to be extremely stable and unaffected by the specific dataset. Other distributions may perform better under specific circumstances when α is reduced (α governs the distribution's diversity), but it is impossible to generalize this. Thus, an ideal distribution $Q(x_i, \tilde{y}_i)$ can be represented as the smoothed δ distribution (Equation(3)).

$$Q(x_i, \tilde{y}_i) = \delta_{smooth}(j - \tilde{y}_i) = \begin{cases} \varepsilon, & if\ j = \tilde{y}_i \\ \dfrac{1-\varepsilon}{c-1}, & otherwise \end{cases} \quad (3)$$

where j represents the identity index, c represents amount of identities (i.e., amount of identities clusters), and ε represents a hyperparameter with a value of 0.99.

The study is a measure of sample uncertainty by estimating the predictive separation of the anticipated probability between the identity and the optimal distribution. Instead of measuring the characteristic inconsistency of the teacher-student model [6], a criterion called probabilistic uncertainty is defined, which calculates the inconsistentness between the predicted distributions $P(x_i, \tilde{y}_i)$ and ideal distributions $Q(x_i, \tilde{y}_i)$. The Kullback-Leibler (KL) divergence is also used in the paper to quantify consistency and define the probabilistic uncertainty as follows:

$$U(x_i, \tilde{y}_i) = D_{KL}(Q(x_i, \tilde{y}_i) \| P(x_i, \tilde{y}_i)) \quad (4)$$

If the pseudo-label produced by clustering has a greater $U(x_i, \tilde{y}_i)$, it should be rejected during the target domain fine-tuning stage since it is more likely to be incorrect.

C. NETWORK STRUCTURE

This paper uses ResNet-50 as the backbone network and feeds the refinement samples as input to the network, which first undergoes convolution, normalization, and maximum pooling operations, then inserts CBAM modules after stage 0 and stage 1, Non-local blocks after stage 2 and stage 3, and then the output is subjected to global average pooling operations after stage 4, and finally the output features are obtained, for which identity loss and triplet loss are performed. The improved network structure diagram is shown in FIGURE 2.

Both the average-pooled feature and the max-pooled feature are used by the CBAM module in this study, and the max-pooled feature, which soft-encodes the most important portion, can make up for the average-pooled feature, which soft-encodes the global statistics. The output of the shared network is merged by aggregate total of elements because the module employs both features and a shared network for these features.

Both convolutional and cyclic operations deal with a local domain by building blocks, Non-local operations are also viewed in this work as a set of components to recognize long-term dependencies. The Non-local operations in this paper assess a location's reaction as a weighted combination of all location attributes. The Non-local operations keep a flexible input size and can be conveniently paired with the convolution operations in ResNet-50, which improves the discriminative power of the network for features up to a point.

D. INTRODUCING CBAM ATTENTION MECHANISM

CBAM can progressively infer a 1D channel attention mapping $M_c \in R^{C\times1\times1}$ and a 2D spatial attention mapping $M_s \in R^{1\times H\times W}$ using a preliminary feature mapping

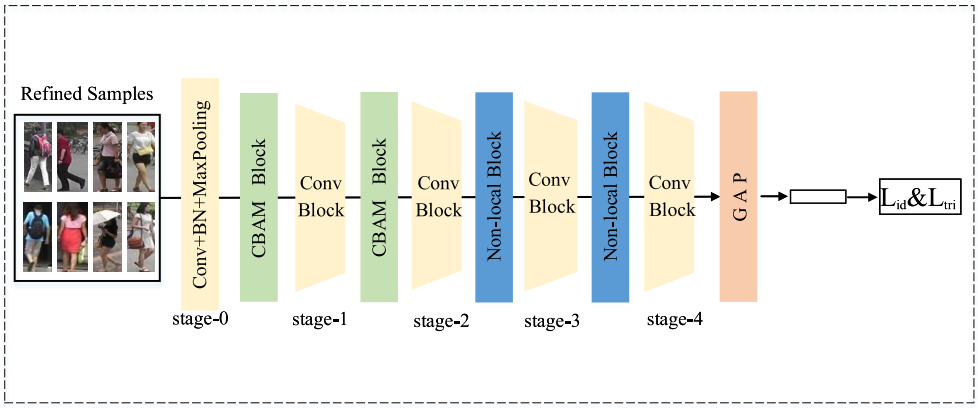

FIGURE 2. Network structure diagram.

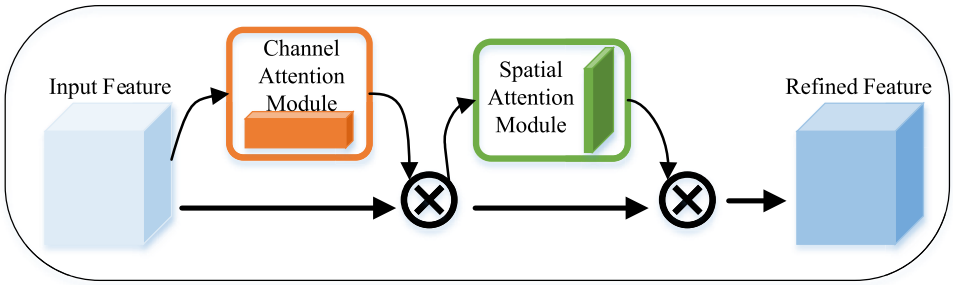

FIGURE 3. Overview diagram of the CBAM module.

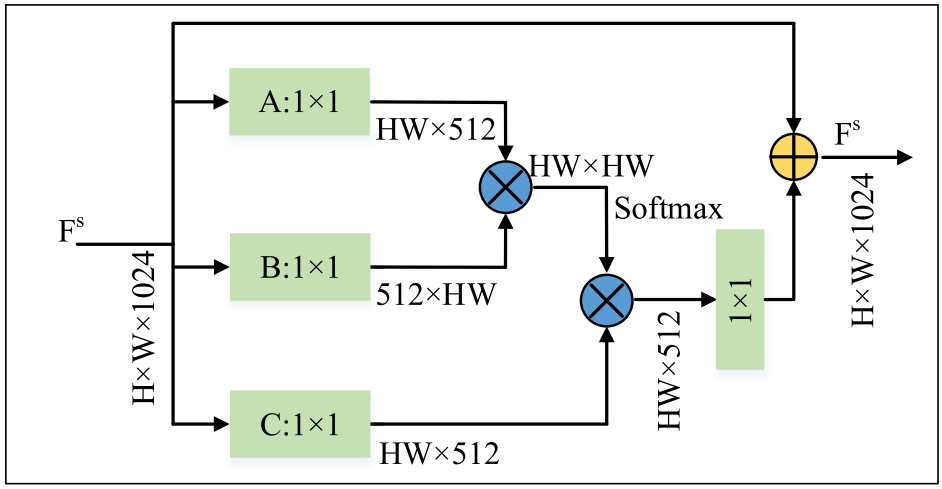

FIGURE 4. Overview diagram of the Non-local block.

$F \in R^{C\times H\times W}$ as input, as shown in FIGURE 3. The whole process of attention can be described as follows:

$$F' = M_c(F) \otimes F$$
$$F'' = M_s(F') \otimes F' \quad (5)$$

where $\otimes$ means element-by-element multiplication. The duplication process governs how mind values spread, with channeled focus values spreading along the spatial dimension and oppositely. F'' is the output of the final refined result.

Channel attention module. The study creates channel focus maps by utilizing the relationships between feature streams. Two distinct traits of the spatial context F^c_{avg} and F^c_{max} are generated by first aggregating the spatial information of the feature graph using the average-pooling and max-pooling operations to represent the average-pooled feature and the max-pooled feature, respectively. The channel attention map $M_c \in R^{C\times 1\times 1}$ is created by transferring both descriptors to a shared network. A multilayer perceptron (MLP) plus a hidden layer make up the shared network. Size of the concealed triggering is set to $R^{\frac{C}{r}\times 1\times 1}$ to eliminate extraneous factors, and the reduction ratio is r. Once the descriptors are

propagated to the common network, then summarization of the components is used to integrate the newly created vector of features. In this regard, the channel attention is established in the following manner:

$$M_c(F) = \sigma\left(MLP\left(AvgPool\left(F\right)\right) + MLP\left(MaxPool\left(F\right)\right)\right) = \left(W_1\left(W_0\left(F^c_{avg}\right)\right)\right) + W_1\left(W_0\left(F^c_{\max}\right)\right) \quad (6)$$

where the sigmoid function is denoted by σ, and $W_0 \in R^{\frac{C}{r}\times C}$ and $W_1 \in R^{C\times\frac{C}{r}}$. Be aware that the MLP's weights W_0 and W_1 are shared across the two inputs, and that weight W_0 comes after the ReLU activation function.

Spatial attention module. For the creation of spatial attention visualizations, the study makes use of the spatial connections that exist between features. Two pooling techniques are used to bring together the channel specifics from a feature mapping and yield two 2D maps: $F^s_{avg} \in R^{1\times H\times W}$ and $F^s_{max} \in R^{1\times H\times W}$. These two mappings represent the average-pooled feature and the max-pooled feature across channels, respectively, which a standard layer of convolution will tie together and bundle to create a 2D attention to space map. Spatial attention is essentially quantified as:

$$M_s(F) = \sigma\left(f^{7\times7}\left(\left[AvgPool\left(F\right); MaxPool\left(F\right)\right]\right)\right) = \sigma\left(f^{7\times7}\left(\left[F^s_{avg}; F^s_{\max}\right]\right)\right) \quad (7)$$

where σ stands for the sigmoid function, $f^{7\times7}$ for the convolution process with a 7×7 filter size.

Using two attentional modules—channel attention and spatial attention, the network can concentrate on ''what'' and ''where'' inside a picture input. The CBAM attention mechanism is obtained by putting the two modules of channel attention and spatial attention into the network in order. The ResNet-50 network used in the paper is made more effective by this module, which may ultimately affect the network to accurately focus on the target person and enhance the person re-identification model's effectiveness.

E. INTRODUCING NON-LOCAL BLOCKS

The Non-local block, which encapsulates Non-local actions, can be integrated into current network patterns. A Non-local block can be described as:

$$z_i = W_z y_i + x_i \quad (8)$$

where i in the equation is the indices of the calculated response output position, y is the output signal, and $+x_i$ denotes the residual connection [31]. Any pre-trained model may have Non-local blocks included in it without affecting the network's original behavior because to this residual connection. FIGURE 4 shows an example diagram of a Non-local block. When employed in advanced subsampling feature mapping, the pairwise calculations of Non-local blocks is compact. A typical convolutional layer in a common network is equivalent to the two-by-two calculation performed through multiplication of matrices. As illustrated in FIGURE 4, the positional embedding on y_i is calculated using the weight matrix W_z in Eq(8), channel counts that are equal to one another on x.

By calculating ties between any two places, irrespective of the gap that separates them, Non-local operations explicitly represent long-term dependencies. Adding a block behind stage2 and stage3 in the ResNet-50 network, the module can perform multiple communications remotely, with information passing back and forth between distant locations in space-time, something that is difficult to achieve with local models. Based on the theory of P2LR, Non-local blocks are added to the network, which can make the network more efficient and can capture long-term dependencies more effectively through Non-local operations, resulting in stronger correlations between the extracted features, which are conducive to improving the discriminative and judgmental power of the network model for person.

IV. EXPERIMENTAL ANALYSIS

A. IMPLEMENTATION DETAILS

In this paper, the network model was designed on the basis of P2LR and educated on four NVIDIA RTX 2080Ti GPUs. A weight decline of 5e-4 in an ADAM optimizer is used to optimize the model, and ResNet-50, pre-trained on ImageNet, is used as the backbone network. We apply standard P-K sampling for person re-identification, with P=4 and K=16 in small batch samples, followed by arbitrary flipping, cropping, and deleting to enhance the data. It should be emphasized that in the source pre-training phase, random erasure is not employed. All person images are adjusted to 256×128. The k-means clustering algorithm is used in this paper, where the number of clusters is set to 500, 700 and 1500 for the Market-1501, DukeMTMC and MSMT17 datasets, respectively. The learning rate is initially set to 3.5e-4 for a total of 80 epochs in the source pre-training phase, and it is decreased by 1/10 at the 40th and 70th epochs. The learning rate in the target fine-tuning phase is set to 3.5e-4.

B. DATASET AND EVALUATION METHODS

The approach in this paper is assessed on three mainstream person re-identification datasets, namely Market-1501 [32], DukeMTMC [33], and MSMT17 [21]. The Market-1501 dataset was recorded by 6 cameras and contains 32,668 annotated photos of 1501 identities; 12,936 of these images were used for training and 19,732 for testing. The 36,411 photos in the DukeMTMC dataset were taken using 8 cameras, and 702 identities were utilized for training and 702 identities for testing. The most significant and difficult person re-identification dataset, MSMT17, was made up of 126,441 photos with 4101 identities, of which 3060 identities were used for testing and 1041 identities were used for training. The experiments in this paper are evaluated using mean accuracy (mAP) and CMC Rank-1/5/10 (R1/R5/R10) accuracy, and no reordering operation is performed throughout the experiments.

TABLE 1. Performance(%) comparison with SOTA UDA Person Re-id methods on DukeMTMC→Market-1501 and Market-1501→DukeMTMC.

Methods	DukeMTMC→Market-1501				Market-1501→DukeMTMC			
	mAP	R1	R5	R10	mAP	R1	R5	R10
ATNet	25.6	55.7	73.2	79.4	24.9	45.1	59.5	64.2
SPGAN+LMP	26.7	55.7	75.8	82.4	26.2	46.4	62.3	68.0
BUC	38.3	66.2	79.6	84.5	27.5	47.4	62.6	68.4
ECN	43.0	75.1	87.6	91.6	40.4	63.3	75.8	80.4
PDA-Net	47.6	75.2	86.3	90.2	45.1	63.2	77.0	82.5
PCB-PAST	54.6	78.4	-	-	54.3	72.4	-	-
SSG	58.3	80.0	90.0	92.4	53.4	73.0	80.6	83.2
ACT	60.6	80.5	-	-	54.5	72.4	-	-
MPLP	60.4	84.4	92.8	95.0	51.4	72.4	82.9	85.0
DAAM	67.8	86.4	-	-	63.9	77.6	-	-
AD-Cluster	68.3	86.7	94.4	96.5	54.1	72.6	82.5	85.5
MMT	71.2	87.7	94.9	96.9	65.1	78.0	88.8	92.5
NRMT	71.7	87.8	94.6	96.5	62.2	77.8	86.9	89.5
B-SNR+GDS-H	72.5	89.3	-	-	59.7	76.7	-	-
MEB-Net	76.0	89.9	96.0	97.5	66.1	79.6	88.3	92.2
UNRN	78.1	91.9	96.1	97.8	69.1	82.0	90.7	93.5
GLT	79.5	92.2	96.5	97.8	69.2	82.0	90.2	92.8
P2LR	81.0	92.6	97.4	98.3	70.8	82.6	90.8	93.7
Ours	**84.6**	**94.0**	**97.8**	**98.8**	**72.6**	**85.1**	**92.2**	**94.3**

TABLE 2. Performance(%) comparison with SOTA UDA Person Re-id methods on Market-1501→MSMT17 and DukeMTMC→MSMT17.

Methods	Market-1501→MSMT17				DukeMTMC→MSMT17			
	mAP	R1	R5	R10	mAP	R1	R5	R10
ECN	8.5	25.3	36.3	42.1	10.2	30.2	41.5	46.8
SSG	13.2	31.6	-	49.6	13.3	32.2	-	51.2
DAAM	20.8	44.5	-	-	21.6	46.7	-	-
NRMT	19.8	43.7	56.5	62.2	20.6	45.2	57.8	63.3
MMT	22.9	49.2	63.1	68.8	23.3	50.1	63.9	69.8
UNRN	25.3	52.4	64.7	69.7	26.2	54.9	67.3	70.6
GLT	26.5	56.6	67.5	72.0	27.7	59.5	70.1	74.2
P2LR	29.0	58.8	71.2	76.0	29.9	60.9	73.1	77.9
Ours	**32.4**	**61.8**	**74.4**	**78.9**	**34.6**	**66.1**	**77.2**	**81.6**

C. COMPARISON WITH STATE-OF-THE-ARTS

On four typical UDA person re-identification tasks, the approach in this study is compared with the state-of-the-art methods. The results are shown in Table 1 and Table 2. DAAM [34] introduced domain alignment constraints and attention modules among the known UDA person re-identification methods. Clustering-based techniques included SSG [35], MMT, MEB-Net [36], and UNRN [9]. SSG evaluated and clusters body parts using both global and local criteria. In this study, a baseline is built using P2LR, and a mutual mean teacher model for UDA person re-identification is introduced. The method presented in this paper greatly increases the accuracy of the UDA person re-identification method when compared to the baseline P2LR.

With a more straightforward and effective framework design than MEB-Net, which created three networks for mutual mean learning, this paper increases the mAP on the DukeMTMC→Market-1501 and Market-1501→DukeMTMC tasks, respectively, by 8.6% and 6.5%. In the target fine-tuning phase, UNRN and GLT [37] made use of source domain data and constructed an external support memory to mine challenging sample pairs. On the MSMT17 dataset, the method in this paper still increases mAP by 8.4% and 7.1% over UNRN on the DukeMTMC→MSMT17 and Market-1501→MSMT17 tasks. It is demonstrated by the experimental findings that the strategy presented in this study executes better on four popular tasks that consist of three datasets, which, in a certain sense, also supports the strategy used in this paper as being beneficial.

D. ABLATION EXPERIMENTS

In this section, a large number of ablation experiments are conducted to demonstrate the accomplishments of each module of the method in this paper. In the ablation study, Market-1501, DukeMTMC and MSMT17 datasets are used, and ResNet-50 is employed as the network infrastructure.

1) EFFECTIVENESS OF THE CBAM ATTENTION MODULE

The CBAM attention module is added behind stage0 and stage1 of the ResNet-50 network on the basis of mitigating the negative effects of pseudo-label noise. The pooling operation for explicit modeling enables for finer attentional judgments as opposed to learnable weighted channel pooling, as seen by the average-pooling operation and the max-pooling operation in the module producing superior accuracy. Taking into account the experimental findings in Table 3, it can be seen that the introduction of the CBAM module improved mAP and R1 by 1.6% and 1.0%, respectively, on the DukeMTMC→Market-1501 task compared to the baseline. On the Market-1501→DukeMTMC task, mAP and R1 were improved by 0.7% and 0.9%, respectively, compared to the baseline.

TABLE 3. Ablation experiments(%).

Methods	DukeMTMC→Market-1501				Market-1501→DukeMTMC			
	mAP	R1	R5	R10	mAP	R1	R5	R10
Base	81.0	92.6	97.4	98.3	70.8	82.6	90.8	93.7
Base+CBAM	**82.6**	**93.6**	**97.6**	**98.4**	**71.5**	**83.5**	**91.4**	**93.6**
Base+Non-local	**84.1**	**93.8**	**98.0**	**98.5**	**72.4**	**84.7**	**91.9**	**94.4**
Base+CBAM+Non-local	**84.6**	**94.0**	**97.8**	**98.8**	**72.6**	**85.1**	**92.2**	**94.3**

As a result, it can be demonstrated that the network using the CBAM method can somewhat increase the baseline's accuracy while also demonstrating the system's strong generalization capabilities on large-scale datasets. This attention module's pooling technique, which creates richer descriptions and spatial attention that successfully complements channel attention, is effective. Additionally, the CBAM module may be easily applied to light-weight networks since it enables efficient feature refinement with only a few parameters and minimal computational overhead.

2) VALIDITY OF NON-LOCAL BLOCKS

Mislabeled samples can have detrimental consequences that can be more successfully mitigated by removing them during training. After stages 2 and 3 of ResNet-50, the Non-local blocks are added to increase the network's ability to discriminate. After several experiments, it was found that the best results were obtained after inserting the Non-local blocks into stage2 and stage3 of ResNet-50, and they were also the most beneficial for network learning. The experimental findings in Table 3 indicate that, when compared to the baseline, mAP and R1 on the DukeMTMC→Market-1501 task improved by 3.1% and 1.2%, respectively. mAP and R1 on the Market-1501→DukeMTMC task improved in comparison to baseline by 1.6% and 2.1%, respectively.

The inclusion of Non-local blocks in the experiments alleviates the long-range messaging problem and improves long-range dependence. Instead of only acquiring global information by stacking many convolutional layers, this module may directly fuse the global information by enhancing the distance dependency, which also adds richer semantic information to the following layers. Additionally, by integrating data regarding attributes associated with the target person and adding this module, the network is able to take into consideration both attention and context, improving the accuracy of the network model for target person recognition. Based on P2LR, adding the operation of Non-local block enhances the capability of the network to extract features and enriches the semantic information. Based on the experimental results in this paper, it can be concluded that the performance's primary contribution improvement of the model is the introduction of the operation of Non-local blocks. Therefore, Non-local blocks are expected to be an important part of multiple network architectures.

V. CONCLUSION

This paper seeks to improve the learning efficiency of deep networks and minimize the damaging effects of pseudo-labels with mislabeling on unsupervised domain-based adaptive person re-identification methods based on clustering. The study finds that the introduction of an attention mechanism can improve the accuracy of the network, which is beneficial to facilitate the research of the UDA person re-identification task. Therefore, this paper proposes a network framework with dual attention, i.e., introducing a CBAM attention mechanism in the shallow layer of the network to enhance the network's attention on the person image channel and spatially, which can more easily help the network to identify the pseudo-labels with mislabeling. In addition, in order to obtain more comprehensive information on feature details, non-local blocks are introduced in the deeper layers of the network to fully obtain both global and local information of the features. Based on the experimental findings, it is evident that, when compared to other methods already in use, the method utilized in this work performs better on each of the four popular UDA person re-identification tasks. Because the MSMT17 dataset is challenging, tests on it take longer to complete, but the accuracy of the results is adequate. However, this poses some difficulties in developing the system, and we will try to investigate more lightweight network structures and better quality algorithms to better handle large datasets in our subsequent work.

REFERENCES

[1] X. Yu, C. Liang, W. Hongyuan, L. Suolan, and Y. Hui, "Unsupervised video-based person re-identification based on the joint global-local metrics," in *Proc. IEEE 7th Int. Conf. Cloud Comput. Intell. Syst. (CCIS)*, Nov. 2021, pp. 176–182.

[2] H. Wang, L. Wu, F. Chen, Z. Ding, Y. Yin, and C. Dai, "Common-covariance based person re-identification model," *Pattern Recognit. Lett.*, vol. 146, pp. 77–82, Jun. 2021.

[3] Y. Liu, K. Wang, L. Liu, H. Lan, and L. Lin, "TCGL: Temporal contrastive graph for self-supervised video representation learning," *IEEE Trans. Image Process.*, vol. 31, pp. 1978–1993, 2022.

[4] W. Fan, L. Yang, and N. Bouguila, "Unsupervised grouped axial data modeling via hierarchical Bayesian nonparametric models with Watson distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9654–9668, Dec. 2022.

[5] H. Sheng, S. Wang, D. Yang, R. Cong, Z. Cui, and R. Chen, "Cross-view recurrence-based self-supervised super-resolution of light field," *IEEE Trans. Circuits Syst. Video Technol.*, early access, May 22, 2023, doi: 10.1109/TCSVT.2023.3278462.

[6] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 480–496.

[7] H. Wang, Z. Ding, J. Zhang, S. Liu, T. Ni, and F. Chen, "Person reidentification by semisupervised dictionary rectification learning with retraining module," *J. Electron. Imag.*, vol. 27, no. 4, 2018, Art. no. 043043.

[8] Y. Dai, J. Liu, Y. Bai, Z. Tong, and L.-Y. Duan, "Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 7815–7829, 2021.

[9] K. Zheng, C. Lan, W. Zeng, Z. Zhang, and Z.-J. Zha, "Exploiting sample uncertainty for domain adaptive person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, 2021, pp. 3538–3546.
[10] M. Yang, J. Zhao, D. Huang, and J. Wang, "Progressive unsupervised domain adaptation for image-based person re-identification," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7730–7736.
[11] X. Liu and S. Zhang, "Graph consistency based mean-teaching for unsupervised domain adaptive person re-identification," 2021, *arXiv:2105.04776*.
[12] J. Han, Y.-L. Li, and S. Wang, "Delving into probabilistic uncertainty for unsupervised domain adaptive person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, 2022, pp. 790–798.
[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.
[14] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[15] T. Ni, X. Gu, H. Wang, Z. Zhang, S. Chen, and C. Jin, "Discriminative deep transfer metric learning for cross-scenario person re-identification," *J. Electron. Imag.*, vol. 27, no. 4, 2018, Art. no. 043026.
[16] J. Li, E. Chen, Z. Ding, L. Zhu, K. Lu, and H. T. Shen, "Maximum density divergence for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3918–3930, Nov. 2021.
[17] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo, "From source to target and back: Symmetric bi-directional adaptive GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8099–8108.
[18] J. Li, M. Jing, K. Lu, L. Zhu, and H. T. Shen, "Locality preserving joint transfer for domain adaptation," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6103–6115, Dec. 2019.
[19] S. Liu, L. Kong, and H. Wang, "Human activities recognition based on skeleton information via sparse representation," *J. Comput. Sci. Eng.*, vol. 12, no. 1, pp. 1–11, Mar. 2018.
[20] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.
[21] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.
[22] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 598–607.
[23] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," 2020, *arXiv:2001.01526*.
[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[25] K. You, G. Qiu, and Y. Gu, "An efficient lightweight neural network using BiLSTM-SCN-CBAM with PCA-ICEEMDAN for diagnosing rolling bearing faults," *Meas. Sci. Technol.*, vol. 34, no. 9, Sep. 2023, Art. no. 094001.
[26] S. Zhou, F. Wang, Z. Huang, and J. Wang, "Discriminative feature learning with consistent attention regularization for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8039–8048.
[27] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "ABD-Net: Attentive but diverse person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8350–8360.
[28] Y. Keshun, Q. Guangqi, and G. Yingkui, "A 3D attention-enhanced hybrid neural network for turbofan engine remaining life prediction using CNN and BiLSTM models," *IEEE Sensors J.*, early access, Jul. 21, 2023, doi: 10.1109/JSEN.2023.3296670.
[29] K. You, G. Qiu, and Y. Gu, "Rolling bearing fault diagnosis using hybrid neural network with principal component analysis," *Sensors*, vol. 22, no. 22, p. 8906, Nov. 2022.
[30] X. Liu, J. He, M. Liu, Z. Yin, L. Yin, and W. Zheng, "A scenario-generic neural machine translation data augmentation method," *Electronics*, vol. 12, no. 10, p. 2320, May 2023.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[32] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[33] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3774–3782.
[34] Y. Huang, P. Peng, Y. Jin, J. Xing, C. Lang, and S. Feng, "Domain adaptive attention model for unsupervised cross-domain person re-identification," 2019, *arXiv:1905.10529*.
[35] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, U. Uiuc, and T. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6111–6120.
[36] Y. Zhai, Q. Ye, S. Lu, M. Jia, R. Ji, and Y. Tian, "Multiple expert brainstorming for domain adaptive person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, Aug. 2020, pp. 594–611.
[37] K. Zheng, W. Liu, L. He, T. Mei, J. Luo, and Z.-J. Zha, "Group-aware label transfer for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5306–5315.

HAIQIN CHEN received the bachelor's degree from the Huaide College, Changzhou University, in 2021, where she is currently pursuing the master's degree. Her research interests include computer vision and person re-identification.

HONGYUAN WANG received the Ph.D. degree in computer science from the Nanjing University of Science and Technology. He is currently a Professor with Changzhou University. His research interests include pattern recognition and intelligence systems. His current interest includes pedestrian trajectory discovery in intelligent video surveillance.

ZONGYUAN DING received the Ph.D. degree from the Nanjing University of Science and Technology, in 2022. He is currently a Lecturer with Changzhou University. His research interests include computer vision and pattern recognition.

PENGHUI LI received the bachelor's degree from the Binjiang College, Nanjing University of Information Science and Technology, in 2022. He is currently pursuing the master's degree with Changzhou University. His research interests include computer vision and person re-identification.