

Received 6 May 2024, accepted 22 May 2024, date of publication 24 May 2024, date of current version 4 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3405169

RESEARCH ARTICLE

Lightweight Person Re-Identification for Edge Computing

WANG JIN^{1,2}, (Member, IEEE), DONG YANBIN¹, AND CHEN HAIMING¹

¹School of Information Science and Technology, Nantong University, Nantong, Jiangsu 226019, China

²School of Computer and Information Engineering, Nantong Institute of Technology, Nantong, Jiangsu 226002, China

Corresponding author: Wang Jin (wj@ntu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China for Young Scholars under Grant 62002179 and in part by the Program for Nantong Basic Science Research Plan under Grant JC22022061.

ABSTRACT In person re-identification, most prevalent models are predominantly designed for cloud computing environments which introduces complexities that limit their effectiveness in edge computing scenarios. Person re-identification systems optimized for edge computing can achieve real-time or near-real-time responses, providing substantial practical value. Addressing this gap, this paper presents the Attention Knowledge-aided Distillation Lightweight Network (ADLN), a network architecture expressly crafted for edge computing. The ADLN enhances inference speed while maintaining accuracy, which is essential for real-time applications. The core innovation of the ADLN lies in its dimension interaction attention mechanism, strategically integrated into the network to boost recognition performance. This mechanism is complemented by a self-distillation approach, transferring attention knowledge from deeper to shallower layers, thereby streamlining the network and accelerating inference. Moreover, the ADLN employs an optimization strategy combining cross-entropy loss, weighted triplet loss regularization, and center loss, effectively reducing intra-class variances. Tested on Market1501 and DukeMTMC-ReID datasets, experiments indicate that the ADLN significantly reduces the model's parameter count and identification latency, while largely maintaining accuracy.


INDEX TERMS Dimensional attention mechanism, edge computing, lightweight network, person re-identification, self-distillation.

I. INTRODUCTION

Person Re-identification (Re-ID) [1], [2], [3], [4], [5], the task of identifying individuals across a collection of pedestrian images taken by non-overlapping cameras, plays a critical role in applications such as criminal investigation, infectious disease monitoring, and shopping behavior analysis. It is a significant area of focus in the field of computer vision.

The person re-identification task presents considerable challenges, such as variations in image backgrounds, pose changes, viewpoint discrepancies, and occlusions resulting from dynamic changes in the pedestrian's environment.

Deep learning techniques are deployed by researchers to tackle these challenges, with solutions categorized into three

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

types: global feature-based [6], local feature-based [7], and attention-based methods [8]. Global feature-based techniques for person re-identification employ convolutional neural networks (CNNs) to represent pedestrian images as feature vectors, which are then processed through distance learning for identification. Chen et al. [9] enhanced the triplet loss function by introducing quadruplet loss, which effectively addresses its constraints and boosts accuracy in person re-identification. He et al. [10] introduced the Vision Transformer for extracting pedestrian features, combined with triplet loss, thereby enhancing the robustness of person re-identification. Local feature-based person re-identification entails segmenting pedestrian images into multiple parts and extracting features for each part, effectively addressing occlusion problems. Narayan et al. [11] introduced a novel approach that expands the horizon of re-identification

techniques by emphasizing the integration of human appearance, face biometrics, and location constraints across camera networks. Sun et al. [12] proposed horizontally partitioning pedestrian features into six blocks and applying identity loss for each block, showcasing the effectiveness of body part segmentation in enhancing recognition accuracy. Luo et al. [13] tackled misalignment of pedestrian body parts with the AlignedReID model, dynamically aligning local features using the shortest path method, thereby further improving person re-identification accuracy.

Recently, attention mechanisms have shown promise in multiple computer vision areas, notably in person re-identification. Hu et al. [14] introduced SENet, successfully implementing a lightweight channel attention mechanism to enhance network performance. Woo et al. [15] proposed the Convolutional Block Attention Module (CBAM), enriching attention maps by incorporating max-pooling features for channel attention and spatial attention. Triplet Attention [16] improves feature quality by interacting with three dimensions of the tensor and applying dimension attention to each dimension. Person re-identification with fusion attention mechanisms showcases the capability to suppress irrelevant features and emphasize relevant pedestrian features. Chen et al. [17] introduced the Mixed Higher-order Attention Network, enhancing attention discriminability and richness, consequently improving person re-identification accuracy. Zhang et al. [18] designed a relation-aware global attention module, inferring relationships between each feature position and others globally, thereby determining the importance of each feature position in aiding the network to extract crucial pedestrian information.

Research on person re-identification has primarily focused on cloud computing scenarios. In cloud computing environments, video data storage and processing are centralized in remote data centers with robust storage and computational capabilities for executing complex models. However, in the future, person re-identification will extend to edge computing environments. Edge computing distributes data processing tasks to devices at the network edge, like IoT devices, smartphones, or local servers. This decentralization reduces latency, enhances processing speed, and fulfills real-time or near-real-time application needs. In contrast to prior approaches, person re-identification for edge computing necessitates lightweight models.

Current research on lightweight person re-identification models primarily focuses on reducing model parameters and computational complexity. Li et al. [19] proposed the HA-CNN model, capable of extracting multiple complementary attention features to maximize the potential complementary effects of person re-identification while maintaining a lightweight design. Quan et al. [20] introduced the Auto-ReID model, applying Neural Architecture Search (NAS) technology to search for an efficient feature extraction network for person re-identification. Wang et al. [21] proposed a coarse-to-fine selection method, using short pedestrian features for initial screening and long features for detailed

screening, employing knowledge distillation methods [22], [23], [24] to compress knowledge from large networks into smaller ones, thereby significantly reducing the complexity of person re-identification in both feature extraction and distance measurement stages. Wu et al. [25] presented a novel framework in “Distilled Person Re-Identification: Towards a More Scalable System”, employing a Multi-teacher Adaptive Similarity Distillation (MASD) approach to effectively refine the process of transferring knowledge from complex models to simpler and more scalable systems. This work underscores the potential of distillation techniques not only in reducing the model size but also in enhancing the efficiency and scalability of re-identification systems, thereby addressing some of the critical challenges in deploying these models for real-time applications. Zhao et al. [26] introduced the Saliency-Guided Iterative Asymmetric Mutual Hashing (SIAMH) model, utilizing ReNeSt-50 as the teacher network and ResNet50 as the student network. Through self-distillation, this model enhanced the quality of hash features, thereby improving person re-identification accuracy. However, these methods currently have limitations, such as the accuracy of HA-CNN does not meet the current state-of-the-art levels, the parameter count of the Auto-ReID model still lagging behind true lightweight models, and the use of knowledge distillation-based models requiring pre-training of a heavyweight teacher network, significantly increasing training costs and time. SIAMH only improves the speed of person re-identification in the distance measurement stage; the model still employs ResNet50 without reducing parameter count or increasing feature extraction speed.

To enhance speed while preserving accuracy, this paper proposes a lightweight person re-identification network with attention knowledge distillation, named ADLN (Attention Knowledge Distillation Lightweight Network). In both the training and testing stages, ADLN involves only one network, significantly reducing model size and substantially improving inference speed while maintaining accuracy. Its specific contributions are as follows:

- (1) Introducing a Dimension Interactive Attention mechanism (DIA) embedded in deep networks to enhance the accuracy of person re-identification.
- (2) Employing self-distillation during training to distill attention features as prior knowledge into shallow networks. During testing, reducing the parameters of the network skeleton containing attention mechanisms to significantly decrease inference time for Re-ID.
- (3) Jointly optimizing network parameters using cross-entropy loss, weighted regularization triplet loss, and center loss to alleviate intra-class differences.

II. ATTENTION KNOWLEDGE DISTILLATION LIGHTWEIGHT NETWORK

Figure 1 illustrates the training stage of ADLN, which consists of four steps. In the first step, the Dimension Interactive Attention mechanism is integrated into a network

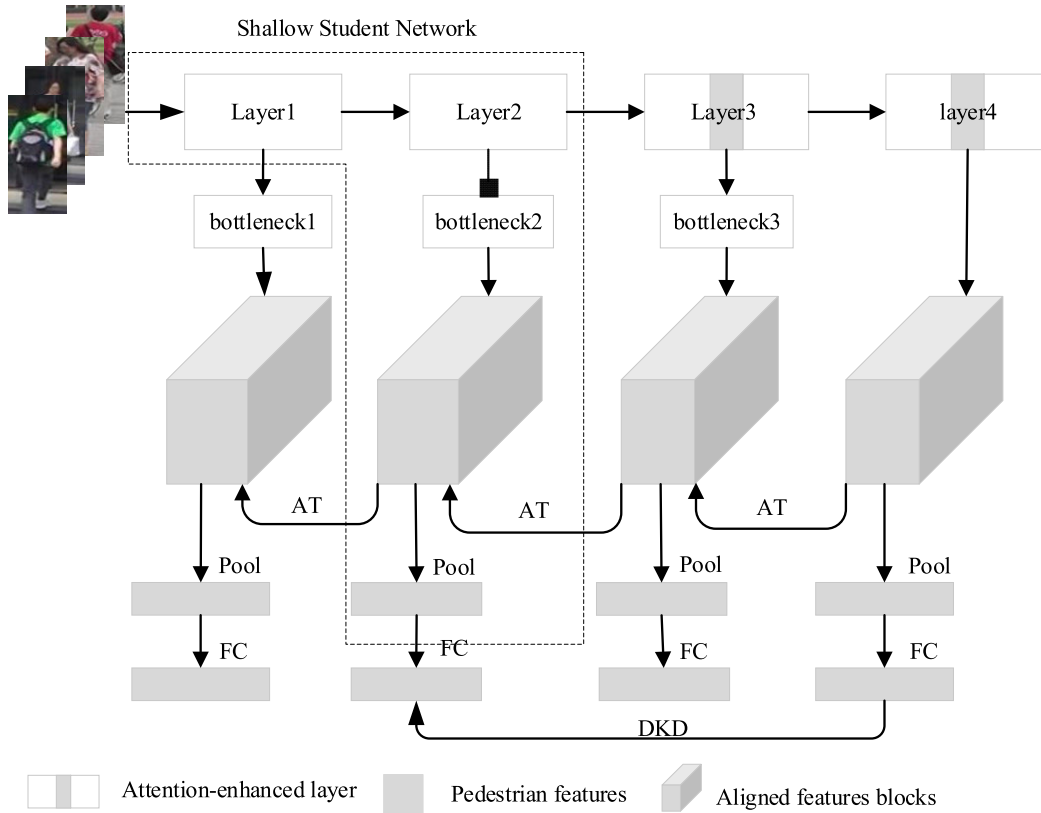


FIGURE 1. Lightweight person re-identification feature extraction network.

with ResNet50 as the backbone to enhance Person Re-identification accuracy. The second step adds a bottleneck to the last layer of the first three layers to address different output feature dimensions across network layers, facilitating subsequent self-distillation. The third step involves training the network by treating the entire network as the teacher network and the shallow network (shown as the first two layers in the figure, but practically, it can be the first or first three layers) as the student network. Self-distillation compresses attention knowledge from the teacher network to the shallow network. Distillation methods include Attention Transfer Knowledge Distillation (AT) and Decoupled Knowledge Distillation (DKD). AT uses a layer-by-layer distillation approach, transferring knowledge from deep networks to shallow networks. DKD uses the entire output features of the teacher network to distill knowledge into the student network. The fourth step involves reducing the deep network with attention to only use the shallow student network for testing, as shown in the dashed box in Figure 1.

A. DIMENSION INTERACTION ATTENTION MECHANISM

The purpose of introducing the dimension interaction attention mechanism is to enhance the recognition rate of the teacher network, thereby improving the performance of the student network it trains. The dimension interaction attention

proposed in this paper is extracted from the first two branches of the Triplet Attention, specifically the interaction branch between the channel dimension and the spatial dimension. The third branch of the Triplet Attention is dedicated to spatial attention. However, after the interaction between the C and H dimensions, as well as the C and W dimensions, in the first two branches, spatial attention has already been effectively achieved. For person re-identification, the third branch can be considered redundant, and better results are obtained by removing this redundant branch.

Figure 2 illustrates a schematic diagram of the Dimension Interaction Attention mechanism. In the figure, “Permutation” represents the tensor rotation operation, “Pool” indicates pooling operations (both average and max pooling), “Conv” denotes convolutional operations for dimension interaction with a 7×7 kernel, and “Sigmoid” is the activation function. We choose sigmoid as the activation function because it inherits several advantages from the Triplet Attention mechanism and offers several advantages in the context of our dimensional interaction mechanism. For an input feature tensor $x \in \mathbb{R}^{C \times H \times W}$, where C represents the numbers of channels, and H and W represents spatial dimensions, the first branch undergoes five operations, as illustrated in the branch below Figure 2. First, the feature x_1 undergoes a rotation operation, resulting in rotated features \hat{x}_1 with a shape of $H \times C \times W$. Next, a pooling operation is applied to

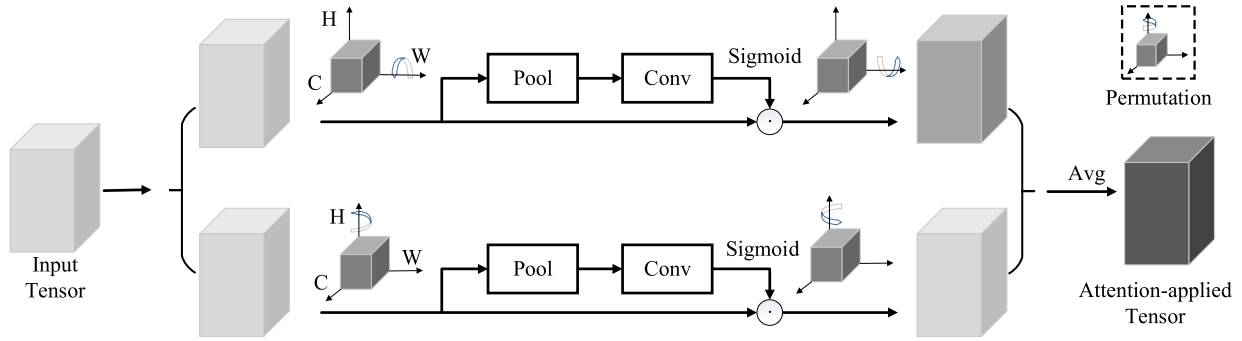


FIGURE 2. Dimensional interactive attention mechanism.

the rotated features along the H dimension, producing pooled features \hat{x}_1^* with a shape of $2 \times C \times W$. Then, a kernel size 7×7 standard convolutional operation ψ_1 is used to learn the relationship between the C and W dimensions, interacting between the C and W dimensions to obtain interactive features with a shape of $1 \times C \times W$. These features pass through a sigmoid activation function σ to generate activation maps, which, when multiplied with the rotated features \hat{x}_1 , yield attention features with a shape of $H \times C \times W$. Finally, a rotation operation is applied to match the input feature’s shape, obtaining attention features of the same shape.

Similarly, the second branch mirrors the operations of the first branch, allowing interaction between the C and H dimensions.

After the two branches are processed, the outputs from two branches are aggregated using simple averaging. The process to obtain the refined attention-applied tensor $y \in \mathbb{R}^{C \times H \times W}$ from dimension interaction attention for an input tensor $x \in \mathbb{R}^{C \times H \times W}$ can be represented by (1)

$$y = \frac{1}{2} \left(\overline{\hat{x}_1 \sigma(\psi_1(\hat{x}_1^*))} + \overline{\hat{x}_2 \sigma(\psi_2(\hat{x}_2^*))} \right) \quad (1)$$

Where σ represents the sigmoid activation function; ψ_1 and ψ_2 represent the standard two-dimensional convolutional layers defined by kernel size 7 in the two branches of DIA. Simplifying (1), y becomes:

$$y = \frac{1}{2} \left(\overline{\hat{x}_1 \omega_1} + \overline{\hat{x}_2 \omega_2} \right) = \frac{1}{2} (\overline{y_1} + \overline{y_2}) \quad (2)$$

Where ω_1 and ω_2 are the two-dimension interaction attention weights computed in DIA. The $\overline{y_1}$ and $\overline{y_2}$ in (2) represents the 90° clockwise rotation to retain the original input shape of $C \times H \times W$.

The dimension interaction attention exhibits characteristics such as cross-dimensional interaction and the establishment of dependencies between dimensions. Moreover, it simultaneously achieves channel attention and spatial attention through dimension attention, this approach significantly improves the discriminability and robustness of pedestrian features by leveraging dimension attention.

B. KNOWLEDGE DISTILLATION

This method employs self-distillation to transfer attention knowledge from the deep layers of the network to the shallow layers, thereby enhancing the recognition capability of the student network. The advantage of self-distillation is that it allows training the student network concurrently with the teacher network during online distillation. This eliminates the need for offline distillation, where one would have to first train the teacher network and then use the trained teacher network to train the student network, thereby reducing training costs.

Attention transfer knowledge distillation utilizes attention as the carrier for knowledge transfer, transferring the knowledge from the teacher network to the student network. Specifically, by taking the output of a certain layer of the network, represented as a three-dimensional tensor $A \in \mathbb{R}^{(C \times H \times W)}$, where C is the number of channels, and $H \times W$ is the size of the tensor. Attention transfer involves mapping A into a 2D spatial attention map, where the size of the 2D spatial attention map is $H \times W$. Subsequently, knowledge transfer is performed through the spatial attention map, and the processing of the spatial attention map is described by (3).

$$T = \sum_{i=1}^C |A_i|^p \quad (3)$$

In the equation, $A_i = A(i, :, :)$ represents the i -th feature map among C feature maps, T is the spatial attention map, p is the power, set to 2 in this paper, and $|\bullet|$ denotes the absolute value. During training, the l_2 loss is calculated between the spatial attention map of a certain layer in the network and the spatial attention map of its subsequent layer, as described in (4).

$$L_{AT} = \sum_{j \in I, j > 1} \left\| \frac{Q^j}{\|Q^j\|_2} - \frac{Q^{j-1}}{\|Q^{j-1}\|_2} \right\|_2 \quad (4)$$

Here, $Q^j = \text{vec}(T^j)$ represents the vectorized form of the spatial attention map for the j -th layer, and L_{AT} is the final loss function for attention migration knowledge distillation, where $I \in \{1, 2, 3, 4\}$.

Decoupled Knowledge Distillation involves splitting the fundamental knowledge distillation into two parts for separate distillation. One part focuses on distilling the similarity

between the probabilities of the target class and the non-target classes, while the other part concentrates on distilling the similarity within the probabilities of the non-target classes. Given a training sample belonging to a total of M classes with the target class being t , during training, its logits obtained through the network and Softmax function are denoted as $P = [p_1, \dots, p_{(t-1)}, p_t, \dots, p_M] \in \mathbb{R}^{(1 \times M)}$. These logits are divided into two parts: the first part, denoted as $b = [p_t, p_{\setminus t}] \in \mathbb{R}^{(1 \times 2)}$, represents the probability vector of the target class, and the second part, denoted as $\hat{P} = [p_1, \dots, p_{t-1}, p_{t+1}, \dots, p_M] \in \mathbb{R}^{1 \times (M-1)}$, represents the probability vector of the non-target classes. The traditional knowledge distillation loss is modified as shown in (5).

$$\begin{aligned} L_{KD} &= \text{KL}(b^T || b^S) + (1 - p_t^T) \text{KL}(\hat{P}^T || \hat{P}^S) \\ b^T &= [p_t^T, p_{\setminus t}^T] \in \mathbb{R}^{1 \times 2} \\ b^S &= [p_t^S, p_{\setminus t}^S] \in \mathbb{R}^{1 \times 2} \\ \hat{P}^T &= [p_1^T, \dots, p_{t-1}^T, p_t^T, L, p_M^T] \in \mathbb{R}^{1 \times (M-1)} \\ \hat{P}^S &= [p_1^S, \dots, p_{t-1}^S, p_t^S, L, p_M^S] \in \mathbb{R}^{1 \times (M-1)} \end{aligned} \quad (5)$$

Where KL represents the Kullback-Leibler divergence, p_i^T denotes the probability of the sample being classified as the i -th class in the teacher network, p_i^S denotes the probability of the sample being classified as the i -th class in the student network, $p_{\setminus t}^T$ represents the probability of the sample not being the target class in the teacher network, and $p_{\setminus t}^S$ represents the probability of the sample not being the target class in the student network.

Decoupled Knowledge Distillation aims to reduce the coupling between the two parts and increases the emphasis on distilling non-target class information. Its distillation loss is defined as shown in (6).

$$L_{DKD} = \alpha \text{KL}(b^T || b^S) + \beta \text{KL}(\hat{P}^T || \hat{P}^S) \quad (6)$$

The two distillation methods were inspired by references [22] and [23].

C. LOSS FUNCTION

To better learn the parameters of the model, this paper proposes the joint training of the network using a combination of cross-entropy loss, weighted regularized triplet loss, and center loss. These three loss functions are commonly used in the field of person re-identification. The cross-entropy loss is expressed as the following equation:

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K y_{i,k} \log(p_{i,k}) \quad (7)$$

Where $y_{i,k}$ represents whether the identity of the i -th image is k , N is the total number of pedestrian classes in the dataset, and $p_{i,k}$ represents the probability that the identity of the i -th image is k . To overcome the problem of model overfitting, this paper employs cross-entropy loss with label smoothing to enhance the model's generalization ability.

Weighted regularized triplet loss is also one of the commonly used loss functions in pedestrian re-identification. It inherits the advantages of triplet loss in optimizing the distance between positive and negative samples and avoids introducing margin parameters. Its expression is as follows:

$$\begin{aligned} L_{wrt}(i) &= \log(1 + \exp(\sum_j w_{ij}^p d_{ij}^p - \sum_k w_{ik}^n d_{ik}^n)) \\ w_{ij}^p &= \frac{\exp(d_{ij}^p)}{\sum_{d_{ij}^p \in P_i} \exp(d_{ij}^p)}, w_{ik}^n = \frac{\exp(d_{ik}^n)}{\sum_{d_{ik}^n \in N_i} \exp(d_{ik}^n)} \end{aligned} \quad (8)$$

In the equation, i represents each anchor image in the batch, P_i denotes the set of positive samples, N_i represents the set of negative samples, and d_{ij}^p and d_{ik}^n respectively represent the distance between the anchor image and the positive sample image and the negative sample image.

To make intra-class features more compact, this paper introduces center loss. Center loss achieves the goal of reducing intra-class distance by learning a central feature point for each class during training, continually pulling together samples from the same class. The calculation formula for center loss is as follows:

$$L_{center} = \frac{1}{2} \sum_{i=1}^B \|f_i - c_{y_i}\|_2^2 \quad (9)$$

In the formula, f_i represents the feature extracted from the deep network for the i -th sample, y_i represents the label of the i -th sample, and c_{y_i} represents the high-dimensional feature center corresponding to the class of y_i . The B represents the batch size.

As shown in Figure 1, the output of each layer of the network needs to be constrained using the three mentioned loss functions. Therefore, the final loss function is as follows:

$$L = \sum_{i=1}^4 (L_{cls}^i + L_{wrt}^i + L_{center}^i) + L_{AT} + L_{DKD} \quad (10)$$

Where L_{cls}^i , L_{wrt}^i and L_{center}^i respectively represent the cross-entropy loss, weighted regularized triplet loss, and center loss for the i -th layer.

III. EXPERIMENT

A. DATASET AND METRICS

To validate the effectiveness of our proposed method, ADLN, experiments were conducted on the widely used Market1501 [27] and DukeMTMC-ReID [28] datasets.

The Market1501 dataset was obtained using the DMP pedestrian detection method, collected from six cameras, and comprises 32,668 images from 1,501 pedestrians. The dataset is divided into a training set with 12,936 images from 751 individuals and a test set with 19,732 images from the remaining 750 individuals. The test set includes 3,368 query images and 19,364 gallery images.

The DukeMTMC-ReID dataset is a subset of the DukeMTMC dataset designed specifically for person re-identification. It consists of data from eight cameras capturing

TABLE 1. Comparison results on Market1501 and DukeMTMC ReID datasets.

Method	Market1501		Params	Flops	DukeMTMC-ReID		
	Rank-1	mAP			Rank-1	mAP	
PCB	93.8	81.	$\geq 23.5\text{M}$	$\geq 4.06\text{G}$	83.3	69.2	
BagTricks	94.5	85.9	$\geq 23.5\text{M}$	$\geq 4.06\text{G}$	86.4	76.4	
AlignedReID++	92.8	89.4	$\geq 23.5\text{M}$	$\geq 4.06\text{G}$	85.2	81.2	
RGA-SC	96.1	88.4	$\geq 23.5\text{M}$	$\geq 4.06\text{G}$	-	-	
AGW	95.1	87.8	$\geq 23.5\text{M}$	$\geq 4.06\text{G}$	89.0	79.6	
MHN-6(PCB)	95.1	85.0	$\geq 23.5\text{M}$	$\geq 4.06\text{G}$	89.1	77.2	
SCSN ^[29]	95.7	88.5	$\geq 23.5\text{M}$	$\geq 4.06\text{G}$	91.0	79.0	
CBDB-Net ^[30]	94.4	85.0	$\geq 23.5\text{M}$	$\geq 4.06\text{G}$	87.7	74.3	
LAG-Net	95.6	89.5	$\geq 23.5\text{M}$	$\geq 4.06\text{G}$	90.4	81.6	
Faster-ReID	93.7	84.0	11.6M	1.18G	87.6	74.8	
HA-CNN	91.2	75.7	2.7M	1.30G	80.5	63.8	
Auto-ReID	94.5	85.1	11.3M	-	88.5	75.1	
ADLN	Res1	73.0	49.5	0.48M	0.58G	62.3	46.7
	Res2	92.4	82.5	1.92M	1.28G	83.0	70.8
	Res3	95.2	89.0	9.40M	2.26G	89.5	80.2
	Res4	95.8	89.2	23.51M	4.08G	91.3	82.5

1,404 pedestrians detected by two or more cameras. The training set comprises 16,522 images from 702 individuals, while the test set contains 19,889 images from the same 702 individuals. The test set includes 2,228 query images and 17,661 gallery images.

To ensure fairness in the experiments, the evaluation metrics used are the Rank-1 accuracy and mean Average Precision (mAP), which are commonly employed in person re-identification. Additionally, the network size is represented by the number of parameters (Params), and the deep neural network's inference speed is measured in terms of floating-point operations per second (Flops).

B. EXPERIMENTAL SETUP

This paper utilizes a ResNet50 network pretrained on the ImageNet dataset as the fundamental backbone for the experimental network. The stride of the last bottleneck block is set to 1. During the training process, three data augmentation methods are considered: random cropping, horizontal flipping, and erasing. The margin for triplet loss and label-smoothed regularization rate are set to 0.3 and 0.1, respectively. The input image size is set to 256×128 , and a learning rate with a warm-up strategy is employed. The learning rate linearly increases from 4×10^{-6} to 4×10^{-4} in the first 10 epochs, followed by exponential decay with a factor of 0.1 at the 40th, 80th, and 120th epochs. The total number of training iterations is 160, and the Adam optimizer is used to optimize model parameters with an $L2$ regularization weight decay factor of 5×10^{-4} .

During testing, the average features between original and horizontally flipped test images are used. The features after the Batch Normalization (BN) layer serve as the pedestrian retrieval features, and cosine distance is employed to measure the distance between features. This experiment is implemented using the PyTorch 1.8 deep learning framework and accelerated on an NVIDIA 3090 GPU.

C. COMPARATIVE EXPERIMENT

This experiment compared the proposed method with selected mainstream person re-identification (Re-ID) approaches to validate its effectiveness. The comparative results are presented in Table 1, where methods based on local features include PCB, AlignedReID++, MHN-6(PCB); attention-based models include RGA-SC, AGW, and lightweight person re-identification networks include HA-CNN, Auto-ReID. The last four rows in the table represent the results of the proposed method, ranging from a one-layer network backbone to a four-layer network backbone. Res4 denotes the four-layer network configuration, wherein the DIA mechanism is implemented within the last two layers of the ResNet50 backbone. Res1 represents the one-layer network configuration, comprising solely the first layer of ResNet50. In all experiments in this paper, a single-frame query mode is used. Except for lightweight networks, all other networks use ResNet50 as the backbone.

Based on the Market1501 dataset, the comparison results are shown in the left two columns of Table 1. The four-layer skeleton model proposed by ADLN achieves Rank-1 and mAP of 95.8% and 89.2%, respectively. Compared with the

attention models RGA-SC and AGW, Rank-1 decreases by 0.3% and increases by 0.7%, and mAP increases by 0.8% and 1.4%, respectively. Compared with the local feature method MHN-6(PCB), Rank-1 and mAP increase by 0.7% and 4.2%. Compared with other models, there are improvements in both metrics to varying degrees. The three-layer skeleton model of ADLN achieves a reduction of 0.6% in Rank-1 and 0.2% in mAP, with a nearly 50% reduction in network parameters and computations. Compared with the lightweight model Auto-ReID, it outperforms in both parameter quantity and accuracy. The two-layer skeleton model of ADLN achieves a reduction of 3.4% in Rank-1 and 6.7% in mAP, with only 1/10 of the original network's parameters and approximately 1/3 of the computation. Compared with the lightweight model HA-CNN, it outperforms in parameters, computations, and accuracy. The one-layer skeleton model of ADLN has a considerable loss in accuracy, but the network model and computation are significantly reduced, making it applicable under certain conditions.

Based on the DukeMTMC-ReID dataset, the comparison results are shown in the right two columns of Table 1. The four-layer skeleton model proposed by ADLN achieves Rank-1 and mAP of 91.3% and 82.5%, respectively. Compared with the attention model AGW, Rank-1 and mAP increase by 2.3% and 2.9%, respectively. Compared with the local feature method MHN-6(PCB), Rank-1 and mAP increase by 2.2% and 5.3%. Compared with other models, there are improvements in both metrics to varying degrees. The three-layer skeleton model of ADLN achieves a reduction of 1.8% in Rank-1 and 2.3% in mAP, with a nearly 50% reduction in network parameters and computations. Compared with the lightweight models Auto-ReID and Faster-ReID, it outperforms in both parameter quantity and accuracy. The two-layer skeleton model of ADLN achieves a reduction of 8.3% in Rank-1 and 11.7% in mAP, with only 1/10 of the original network's parameters and approximately 1/3 of the computation. Compared with the lightweight model HA-CNN, it outperforms in parameters, computations, and accuracy. The one-layer skeleton model of ADLN has a considerable loss in accuracy, but the network model and computation are significantly reduced.

According to the above comparison results on these two datasets, the attention model proposed in this paper effectively improves the accuracy of person re-identification, and the self-distillation method significantly reduces the network parameters and saves computational costs while maintaining limited accuracy loss. In practical applications, determining the appropriate number of layers for the ADLN model requires a careful balance between network parameters and accuracy.

D. ABLATION EXPERIMENT

To further verify the effectiveness of the ADLN algorithm, ablation experiments were conducted on the Market1501 dataset, and the results are presented in Table 2. The baseline method employed in Table 2 is the Attention Pyramid

TABLE 2. Ablation experiments on dataset market1501.

Method	Rank-1	mAP	Params	Flops
Baseline	95.0	86.1	23.5M	4.06G
Baseline +DIA	95.8	89.2	23.5M	4.08G
Res2	90.5	76.4	1.92M	1.28G
Res2+AT	91.1	77.5	1.92M	1.28G
Res2+AT+DKD	91.5	79.1	1.92M	1.28G
Res2+DIA+AT+DKD	92.4	82.5	1.92M	1.28G

networks (APNet). The student networks in these ablation experiments consist of the first two subnetworks of the network backbone.

After incorporating the dimension interaction attention module into the baseline, there are improvements of 0.8% in Rank-1 and 3.1% in mAP, with no significant increase in network parameters and computational complexity, demonstrating the effectiveness of the attention module. After knowledge distillation with attention transfer, the Rank-1 and mAP improve by 0.6% and 1.1%, respectively. Furthermore, with decoupled knowledge distillation, the Rank-1 and mAP continue to improve by 0.4% and 1.6%. These results indicate the effectiveness of both knowledge distillation methods. If the attention mechanism is added to the teacher network's deeper layers, the student network's Rank-1 and mAP increase by 0.9% and 3.4%, respectively, demonstrating that both knowledge distillation methods effectively compress the deep attention knowledge of the teacher network into the shallow student network.

IV. CONCLUSION

This paper introduces the ADLN (Attention-based Distillation for Lightweight Network) method to tackle challenges in existing deep learning models for person re-identification, such as high parameter count, slow inference speed, and difficulties in deployment on edge devices. The ADLN method integrates a dimension-interaction attention mechanism to improve person re-identification accuracy. During training, a self-distillation approach is employed, progressively distilling attention knowledge from deep layers to shallow layers. During testing, the deep layers with attention are discarded, reducing network parameters and enhancing the speed of person feature extraction.

Experimental results on the Market1501 and DukeMTMC-ReID datasets demonstrate that the proposed method significantly reduces the model's parameter count while maintaining accuracy, making it suitable for edge computing applications.

Despite its notable contributions, this work is not without limitations. One of the main constraints lies in balancing the lightweight nature of the model with its accuracy. While the ADLN model achieves significant reductions in parameter

count and computational demands, the most lightweight configurations (e.g., Res1 and Res2 models) exhibit a marked decrease in accuracy compared to more complex configurations. This trade-off highlights the challenge of balancing efficiency and effectiveness, particularly in scenarios where the highest possible accuracy is paramount.

Additionally, while the ADLN model marks a significant step towards optimizing person re-identification for edge computing, there remains room for improvement in further accelerating the metric learning stage. Future work could explore more advanced techniques for speeding up this phase without compromising the model's discriminative capability.

REFERENCES

- [1] 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.
- [2] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 1092–1108, Apr. 2020.
- [3] C.-X. Ren, B. Liang, P. Ge, Y. Zhai, and Z. Lei, "Domain adaptive person re-identification via camera style generation and label propagation," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1290–1302, 2020.
- [4] L. Youjiao, Z. Li, and Z. Jing, "Overview of pedestrian re-recognition technology," *J. Autom.*, vol. 44, no. 9, pp. 1554–1568, 2018.
- [5] L. Hao, J. Wei, and F. Xing, "Research progress of pedestrian recognition based on deep learning," *J. Autom.*, vol. 45, no. 11, pp. 8295–8302, 2019.
- [6] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 149–1487.
- [7] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global–local alignment descriptor for scalable person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 986–999, Apr. 2019.
- [8] X. Gong, Z. Yao, X. Li, Y. Fan, B. Luo, J. Fan, and B. Lao, "LAG-net: Multi-granularity network for person re-identification via local attention system," *IEEE Trans. Multimedia*, vol. 24, pp. 217–229, 2022.
- [9] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Honolulu, HI, USA: IEEE, Jul. 2017, pp. 1320–1329.
- [10] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14993–15002.
- [11] N. Narayan, N. Sankaran, D. Arpit, K. Dantu, S. Setlur, and V. Govindaraju, "Person re-identification for improved multi-person multi-camera tracking by continuous entity association," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 566–572.
- [12] Y. Sun, L. Zheng, and Y. Yang, "Beyond part models: Person retrieval with refined part pooling and a strong convolutional baseline," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2017, pp. 501–518.
- [13] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang, "AlignedReID++: Dynamically matching local information for person re-identification," *Pattern Recognit.*, vol. 94, pp. 53–61, Oct. 2019.
- [14] 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.
- [15] S. Woo, J. Park, and J. Y. Lee, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2018, pp. 3–19.
- [16] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3138–3147.
- [17] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 371–381.
- [18] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3183–3192.
- [19] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 2285–2294.
- [20] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, "Auto-ReID: Searching for a part-aware ConvNet for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 3749–3758.
- [21] G. Wang, S. Gong, J. Cheng, and Z. Hou, "Faster person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 279–292.
- [22] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network", 2015, *arXiv:1503.02531*.
- [23] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 11943–11952.
- [24] Z. Sergey and K. Nikos, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–13.
- [25] A. Wu, W.-S. Zheng, X. Guo, and J.-H. Lai, "Distilled person re-identification: Towards a more scalable system," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1187–1196.
- [26] C. Zhao, Y. Tu, Z. Lai, F. Shen, H. T. Shen, and D. Miao, "Salience-guided iterative asymmetric mutual hashing for fast person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 7776–7789, 2021.
- [27] 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)*, Washington, DC, USA, Dec. 2015, pp. 1116–1124.
- [28] E. Ristani, F. Solera, and R. Zou, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, Cham, Switzerland: Springer, 2016, pp. 17–35.
- [29] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang, "Salience-guided cascaded suppression network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3297–3307.
- [30] H. Tan, X. Liu, Y. Bian, H. Wang, and B. Yin, "Incomplete descriptor mining with elastic loss for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 160–171, Aug. 2022.



WANG JIN (Member, IEEE) was born in Nantong, Jiangsu, China, in 1981. He received the Ph.D. degree. He holds the position as an Associate Professor. He is a Senior Member of China Computer Federation (CCF). His primary research interests include encompass computer vision and pattern recognition.



DONG YANBIN was born in Sanmenxia, Henan, China, in 2001. He is currently pursuing the master's degree. His primary research interests include computer vision and pattern recognition.



CHEN HAIMING was born in Xuzhou, Jiangsu, China, in 1995. He is currently pursuing the master's degree. He is a member of China Computer Federation (CCF). His primary research interest includes computer vision.

...