Multi-Grained feature aggregation based on Transformer for unsupervised person re-identification

Zhongmin Liu*,Changkai Zhang

School of Electrical Engineering and Information Engineering, Lanzhou University of Technology,
Lanzhou,Gansu,730050,China
**corresponding author:liuzhmx@163.com*

Abstract: Person re-identification aims to retrieve specific person targets across different surveillance cameras. Due to problems such as posture changes, object occlusion, and background interference, the person re-identification effect is poor. A multi-grained feature aggregation unsupervised person re-identification based on Transformer is proposed to make full use of the extracted person features. First, a Dual-Channel Attention module is designed to enable the network to adaptively adjust the receptive field size based on multiple scales of input information, facilitating the capture of connections between different parts of the person's body. This enhances the network's ability to extract person feature information, enabling it to obtain more critical image information and output more representative person expression features. Next, an Explicit Visual Center module is proposed to capture global information and aggregate essential local information, strengthening the network's feature representation and thereby improving the model's generalization capability. Finally, validation are conducted on popular datasets such as Market1501,DukeMTMC-reID, and MSMT17. The results demonstrate that the improved model achieves higher performance metrics, yielding greater recognition accuracy and better representation of person features. Code is available at https://gitee.com/zhchkk/mgfa

Keywords: Feature Aggregation; Multi-Grained Features; Unsupervised Learning; Person Re-Identification; Attention Mechanism

1. INTRODUCTION

The task of Person Re-identification (ReID) aims to retrieve specific images or videos of individuals across different cameras. In other words, given an image or video of a person, the goal is to locate that person's image or video within a large-scale database captured by multiple surveillance cameras. This technology finds wide applications in criminal investigations and intelligent monitoring (Lin et al., 2020). (Wu et al., 2016) propose a deep end-to-end neu-ral network to simultaneously learn high-level features and a corresponding similarity metric for person re-identification. The network takes a pair of raw RGB images as input, and outputs a similarity value indicating whether the two input images depict the same person.While supervised ReID methods have made significant progress with the development of Convolutional Neural Networks (CNNs), they often require costly manual annotations and extensive labeling efforts, limiting their practical utility. Therefore, to address the real-world applications of person re-identification, researchers have introduced unsupervised person re-identification, aiming to learn models directly from unlabeled data. Due to its good potential, it has received extensive attention from the academic community in recent years, which makes it more practical and better scalable in actual scenarios. Over the past few years, significant progress has been made, mainly due to the leverage of pseudo labeling (Lin et al., 2019) and contrastive learning (Chen et al., 2020; He et al., 2020; Wu et al., 2018) techniques. However, existing unsupervised methods often focus on the design of various contrastive losses (Chen et al., 2021; Dai et al., 2021; Wang et al., 2021; Wang et al., 2022) and the refinement of noisy pseudo labels (Cho et al., 2022; Wu et al., 2021; Zhang et al., 2021). Most of them pay little attention to the improvement of their feature extraction networks, which are crucial for identification as well.

On the contrary, the architecture of feature extraction backbones has been extensively investigated in supervised person Re-ID. For example, besides bag of tricks (BoT) (Luo et al., 2019), partition-based (Cheng et al., 2016; Sun et al., 2018) or multi-granularity (Wang et al., 2018; Zheng et al., 2019) networks are developed to capture fine-grained cues, and attention schemes (Chen et al., 2019b; Li et al., 2018; Si et al., 2018) are integrated to concentrate on discriminative parts. Recently, self-attention or transformer mechanisms (He et al., 2021; Lai et al., 2021; Li et al., 2021; Luo et al., 2021) have also been successfully applied to supervised Re-ID. Some of them (Lai et al., 2021; Li et al., 2021;) integrate transformers with convolutional neural networks (CNNs) and the others (He et al., 2021; Luo et al., 2021) construct pure transformer architectures to explore long-range contexts. It has been validated that both fine-grained cues and long-range contexts greatly boost the performance of supervised Re-ID.

Inspired by the above supervised person re-identification, especially by the CNN-based Multiple Granularity Network (MGN) (Wang et al., 2018) and the pure transformer networks (He et al., 2021; Luo et al., 2021; Sharma et al., 2021; Zhu et al., 2023), we intend to investigate the way of extracting multi-grained features from a pure transformer network to address the more challenging unsupervised Re-ID problem. Vision Transformer (ViT) is used as the backbone network, and is slightly modified on this basis to adapt to the person re-identification task. By constructing a dual-branch architecture based on the backbone network to learn local features and global features, multi-granularity features are effectively learned from the Transformer network. Finally, offline and online comparative learning losses are defined through global and local features. However, the performance of purely unsupervised methods largely depends on the quality of clustering and the adequacy of granular information, resulting in insufficient feature extraction, which in turn affects the accuracy of person recognition.

To solve the above problems, this paper proposes a Transformer-based Multi-Grained Feature Aggregation unsupervised person re-identification network model (MGFANet). MGFANet can not only realize the aggregation of image features and improve network efficiency but also can further mine the connection between different parts of the person's body to obtain more abundant person feature information. Although some recent works also attempt to learn partlevel features from pure transformer networksfor supervised Re-ID or leverage CNN-based part-level features for unsupervised Re-ID (Yang et al., 2019; Yang et al., 2022), our method distinguishes itself from them in the following aspects:

1) This paper introduces a Transformer-based unsupervised person re-identification network model called MGFANet. This network is capable of aggregating image features, exploring potential relationships among person features, and enhancing the model's ability to extract information from person images.

2) This paper introduces the Dual Channel Attention module and the Explicit Visual Center module to enhance the feature extraction capabilities of the network model. The Dual Channel Attention module is designed to improve the model's ability to extract features from person images, while the Explicit Visual Center module is used to aggregate critical local information from person images.

3) The network model MGFANet proposed in this paper demonstrates superior performance compared to the majority of existing unsupervised person re-identification methods on three popular datasets: Market1501, DukeMTMC-reID, and MSMT17.

2. RELATED WORK

2.1 Person re-identification based on attention mechanism

Attention mechanisms are widely applied in various deep learning domains, including information processing and speech recognition. These mechanisms enable models to dynamically focus on specific parts of input data relevant to the task at hand. (Sun et al., 2019) introduced a Visibility-aware Part-level Model (VPM) that utilizes attention mechanisms. It employs a localization method to pinpoint visible regions within a given image and subsequently learns local features based on these visible regions. (Chen et al., 2019a) proposed High-Order Attention (HOA), which involves modeling and leveraging complex and high-order statistical information within attention mechanisms. This approach captures subtle differences among persons, thereby enhancing the model's discriminative power and feature richness. (Huang et al., 2020) presented a 3D attention model that concurrently considers channel and spatial information. This enables better complementarity in attention extraction, ultimately enhancing feature recognition capabilities. (Shi et al., 2023) proposed a lightweight network with a progressive attention mechanism, which gradually segments the features into local blocks of different granularity, allowing for the learning of discriminative features at each granularity level. This progressive approach enhances the network’s ability to perceive foreground information from coarse to fine levels and improves feature matching capability. To explicitly alleviate the impact of the person changing clothes on re-identification, (Zhou et al., 2022) presents a cloth-irrelevant harmonious attention network (CIHANet) that learns cloth-irrelevant knowledge. Firstly, with the help of human parsing, the color information of human clothing is removed to generate black clothes images. Secondly, the raw person images are used to learn features with more color-based appearance knowledge, while the black clothes images are used to learn features with more cloth-irrelevant knowledge. Then, to fuse the knowledge of two distinct streams, we propose the harmonious attention module, including mutual learning attention and salience guided attention mechanisms. The mutual learning attention mechanism adaptively selects identity-relevant features across feature channels to make two streams interact with each other. The salience guided attention mechanism highlights the cloth-irrelevant areas by transferring the spatial knowledge from the black clothes stream to the raw images stream.

2.2 Unsupervised person re-identification

To reduce the cost of manual labeling and improve scalability in real-world scenarios, unsupervised person re-identification aims to train a discriminative model on unlabeled datasets. However, due to the lack of training labels, discovering the latent relationships between data instances is crucial in person re-identification. (Li et al., 2019) proposed an Unsupervised Tracklet Association Learning (UTAL) model to automatically eliminate person ID labels. The UTAL model achieves effective person re-identification capabilities through single-camera tracklet recognition learning and cross-camera tracklet association learning. (Chen et al., 2018) introduced a method that jointly optimizes two margin-based association losses to constrain the associations of each frame and discover more reliable cross-camera trajectories automatically. (Wu et al., 2019) presented an Unsupervised Graph Association (UGA) model for mining view-invariant representations across individuals through cross-view association learning at the low level, reducing interference from noisy associations. (Dong et al., 2023) proposed a dual pseudo label refinement framework for unsupervised domain adaptive person re-identification. It has two pseudo label refinement modules, one learns cross consistency of corresponding features between two collaborative networks and the other explores mutual consistency between global and local feature spaces. By working complementarily and jointly, the two modules optimize the quality of training dataset.

2.3 Transformer-based person re-identification

The Transformer is a sequence modeling model based on self-attention mechanisms and has been widely applied in Natural Language Processing (NLP) and Computer Vision (CV). In recent years, the Transformer has been employed in unsupervised person re-identification for feature extraction from person images. (Lai et al., 2021) proposed an Adaptive Part Partitioning (APD) model to better extract local features. APD mainly consists of two key modules: a Transformer-based local merging (TPM) module and a Part Mask Generation (PMG) module. TPM first adaptively assigns patch tokens of the same semantic object to the identical part. Then PMG combines these identical parts together and generates several non-overlapping masks for robust part division. (Zhang et al., 2021) designed a Hierarchical Aggregation Transformers (HAT) model that integrates low-level details as global priors for high-level semantic information. (Ma et al., 2021) proposed a Pose-guided Inter- and Intra-part Relational Transformer (PIRT) for occluded person re-identification, which utilizes a Transformer to build part-aware long-term relationships.

3. PROPOSED METHOD

This paper presents an improvement upon TMGF (Transformer-Based Multi-Grained Features) (Li et al., 2021) and introduces the MGFA model for unsupervised person re-identification. In the Transformer feature encoding phase, the DCA module is introduced to adaptively adjust the receptive field size based on multiple scales of input information, enhancing useful features and improving the network model's representation capability. The EVC module is incorporated to aggregate deep and shallow features, obtaining a comprehensive feature representation. The overall structure of the MGFA model is depicted in Figure 1, comprising a ViT, MGN and unsupervised loss functions.

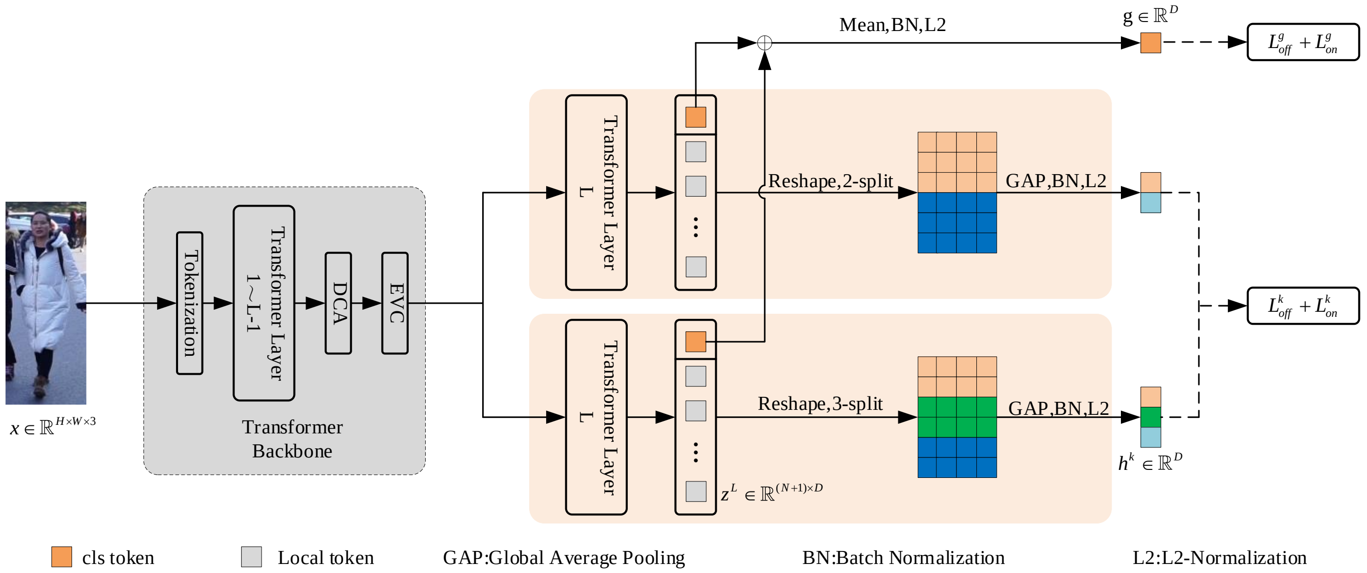

Fig. 1. Overall structure block diagram of the model.

The ViT is used as the backbone network model. Input images $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ are first processed through a convolutional stem with Instance-Batch Normalization (IBN) to generate feature maps $\mathbf{x}' \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$, where H represents height, W represents width, and C represents the number of channels. Subsequently, the feature maps are divided into $N = \dfrac{HW}{P^2}$ non-overlapping patches, each of size $\dfrac{P}{2} \times \dfrac{P}{2}$. Each patch is projected into a D-dimensional feature $\mathbf{f} \in \mathbb{R}^D$ as an embedding label. A learnable class token *cls* is added to the sequence of patch labels. Finally, position embeddings and additional camera embeddings are appended to the patch labels, forming the input for the Transformer network (He et al.2021). The entire tokenization process is represented as follows:

$$\begin{aligned} &\mathbf{f}_i = \psi_i(\text{ICS}(\mathbf{x})), \quad i = 1, \ldots, N \\ &\mathbf{z}^0 = [cls; \mathbf{f}_1; \ldots; \mathbf{f}_N] + \mathbf{p} + \lambda_c \mathbf{c} \end{aligned} \tag{1}$$

Here, ICS represents the convolutional stem with Instance-Batch Normalization, and ψ_i represents the partition and projection operation. $\mathbf{p} \in \mathbb{R}^{(N+1) \times D}$ is the position embeddings, $\mathbf{c} \in \mathbb{R}^{(N+1) \times D}$ is the camera embeddings, and λ_c is a weighted hyperparameter.

The embedded tokens $\mathbf{z}^0$ are input into a Transformer network composed of *L-1* Transformer layers. Each layer consists of Multi-head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) modules.

$$\begin{aligned} \hat{\mathbf{z}}^{l-1} &= \text{MSA}(\text{LN}(\mathbf{z}^{l-1})) + \mathbf{z}^{l-1} \\ \mathbf{z}^{l} &= \text{MLP}(\text{LN}(\hat{\mathbf{z}}^{l-1})) + \hat{\mathbf{z}}^{l-1} \\ \mathbf{z}_E &= \text{EVC}(\text{DCA}(\mathbf{z}^{l})) \\ \hat{\mathbf{z}}^{L} &= \text{MSA}(\text{LN}(\mathbf{z}_E)) + \mathbf{z}^{l-1} \\ \mathbf{z}^{L} &= \text{MLP}(\text{LN}(\hat{\mathbf{z}}^{L})) + \hat{\mathbf{z}}^{L} \end{aligned} \tag{2}$$

Here, LN represents layer normalization, $l \in \{1, \ldots, L-1\}$. Layer Normalization refers to normalizing the input of all neurons in batches, that is, making the data in the layer obey a normal distribution with a mean of 0 and a variance of 1, which helps to speed up task training. This method is based on Normalizing by sample, rather than initial normalizing by scale, can improve the system's robustness to scaling swing changes. Multi-Head Self-Attention mechanism aims to enhance the model's expressive ability and generalization ability. It obtains a richer representation by using multiple independent attention heads, calculating attention weights separately, and splicing or weighting their results. Multilayer Perceptron is called a feedforward neural network. It is a machine learning model based on neural networks that performs high-level abstraction and classification of input data through multi-layered non-linear transformations.

Therefore, we denote the final output of the transformer network as:

$$\mathbf{z}^L = [cls^L; \mathbf{f}_1^L; \ldots; \mathbf{f}_N^L] \tag{3}$$

Subsequently, the output of the Transformer network is input into the multi-grained module to extract multi-grained features. Each branch has a copy of the L-th Transformer layer, which outputs a global token cls^L and N local tokens $[\mathbf{f}_1^L; \ldots; \mathbf{f}_N^L]$. the local tokens substantially correspond to the original input patches. Therefore, reshaping the local tokens in each branch and dividing them into horizontal stripes generates local features. The output local tokens from the i-th branch denoted as $[\mathbf{f}_{i,1}^L; \ldots; \mathbf{f}_{i,N}^L]$, which are reconstructed into feature maps $\mathbf{f}_i'$ in size of $\frac{H}{P} \times \frac{W}{P} \times D$, then $\mathbf{f}_i'$ are divided into K_i horizontal parts and subjected to average pooling to obtain a D-dimensional feature. this procedure is described as:

$$\begin{aligned} &\mathbf{f}_i' = \text{Reshape}([\mathbf{f}_{i,1}^L; \ldots; \mathbf{f}_{i,N}^L]), \quad i = 1, 2; \\ &\mathbf{h}_{i,k} = \text{AvgPool}(\text{Split}(\mathbf{f}_i', k)), \quad k = 1, \ldots, K_i \end{aligned} \tag{4}$$

Here, $\mathbf{h}_{i,k}$ represents the k-th local feature for the i-th branch. The collection of local features is denoted as $\{\mathbf{h}^k\}_{k=1}^K$, where $K = K_1 + K_2$.

The global features are obtained by averaging the two branches' global tokens:

$$\mathbf{g} = \frac{1}{2}(cls_1^L + cls_2^L) \tag{5}$$

Finally, the global features and local features are normalized using Batch Normalization (BN) and L_2 normalization layers to obtain unit benchmarks. We detail architectures in Table 1.

Table 1. Detail architectures.

MGAF (384, 128)			
In	Out	Out.size	Layers
x	x'	(192, 64)	Tokenization
x'	z^{l-1}	(192, 64)	Transformer 1~L-1
z^{l-1}	z^D	(192, 64)	DCA
z^D	z^E	(256, 256)	EVC
z^E	z^L	(256, 256)	Transformer L
z^L	f_i'	(256, 256)	MGN

3.1 Dual-Channel Attention Module

In the unsupervised person re-identification task, the attention mechanism can effectively suppress irrelevant background information and better focus on person features. Most existing attention methods focus on developing more complex attention modules to achieve better performance, which inevitably increases the complexity of the model. In order to overcome the contradiction between performance and complexity, while adaptively adjusting the perception The size of the field, a dual-channel attention mechanism is proposed. Mainly split, fuse and select through three operations, as shown in Figure 2.

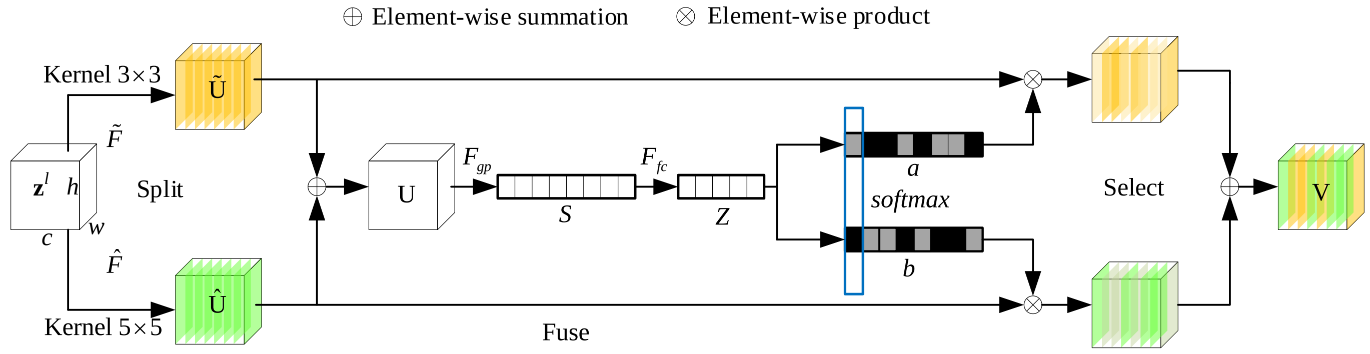

Fig. 2. Structure of DCA module.

Split: For any given feature map $\mathbf{z}^l \in \mathbb{R}^{H' \times W' \times C'}$, First, perform two conversions: $\tilde{F}: \mathbf{z}^l \rightarrow \tilde{\mathbf{U}} \in \mathbb{R}^{H \times W \times C}$ and $\hat{F}: \mathbf{z}^l \rightarrow \hat{\mathbf{U}} \in \mathbb{R}^{H \times W \times C}$, with convolutional kernel sizes of 3 and 5, $\tilde{F}$ and $\hat{F}$ are composed of efficient grouped/depthwise convolutions, Batch Normalization and ReLU function in sequence..

Fuse: the results from both branches are fused by element-wise summation:

$$\mathbf{U} = \tilde{\mathbf{U}} + \hat{\mathbf{U}} \tag{6}$$

To obtain the global information $\mathbf{S}$ ($\mathbf{S} \in \mathbb{R}^C$) of length C by using global average pooling, and to compute S_c by reducing $\mathbf{U}$ to spatial dimensions H×W:

$$S_c = F_{gp}(\mathbf{U}_c) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{U}_c(i,j) \tag{7}$$

Here, F_{gp} is the global average pooling operation.

Then, S_c is compressed into feature $\mathbf{Z}$ ($\mathbf{Z} \in \mathbb{R}^{d\times 1}$) through the fully connected layer, which improves the efficiency while reducing the dimensionality:

$$\mathbf{Z} = F_{fc}(\mathbf{S}) = \sigma(B(\mathbf{W}_{\mathrm{S}})) \tag{8}$$

Here, σ is the ReLU function, B denotes the Batch Normalization, $\mathbf{W} \in \mathbb{R}^{d\times C}$.

Select: to adaptively select information of different spatial dimensions, soft attention across channels guided by the compressed feature $\mathbf{Z}$ is utilized. Therefore, the softmax operation is applied on the two branches:

$$a_c = \frac{e^{\mathbf{A}_c\mathbf{Z}}}{e^{\mathbf{A}_c\mathbf{Z}} + e^{\mathbf{B}_c\mathbf{Z}}}, b_c = \frac{e^{\mathbf{B}_c\mathbf{Z}}}{e^{\mathbf{A}_c\mathbf{Z}} + e^{\mathbf{B}_c\mathbf{Z}}} \tag{9}$$

Here, $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{C\times d}$ and a, b denote the soft attention vectors for $\tilde{\mathbf{U}}$ and $\hat{\mathbf{U}}$, respectively. $\mathbf{A}_c \in \mathbb{R}^{1\times d}$ is the c-th row of $\mathbf{A}$, a_c is the c-th element of a, $\mathbf{B}_c \in \mathbb{R}^{1\times d}$ is the c-th row of $\mathbf{B}$, and b_c is the c-th element of b. Under both branches, matrix B is redundant, because $a_c + b_c = 1$. The final feature map $\mathbf{V}$ is obtained through the attention weights on various kernels:

$$\mathbf{V}_c = a_c \cdot \tilde{\mathbf{U}} + b_c \cdot \hat{\mathbf{U}},\ a_c + b_c = 1 \tag{10}$$

Here, $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_c]$, $\mathbf{V}_{\mathbf{c}} \in \mathbb{R}^{H\times W}$.

3.2 Explicit Visual Center Module

The Transformer-based person re-identification method shows its effectiveness and superiority in feature extraction. However, most of the existing methods pay too much attention to the inter-layer feature interaction and ignore the intra-layer feature representation. Although some methods try to learn a compact intra-layer feature representation in the model of ViT, they ignore the very important task of person re-identification. Local corner area. To solve this problem, a display visual center module is proposed, which mainly consists of a lightweight MLP and a Learnable Visual Center (LVC) connected in parallel, as shown in Figure 3.

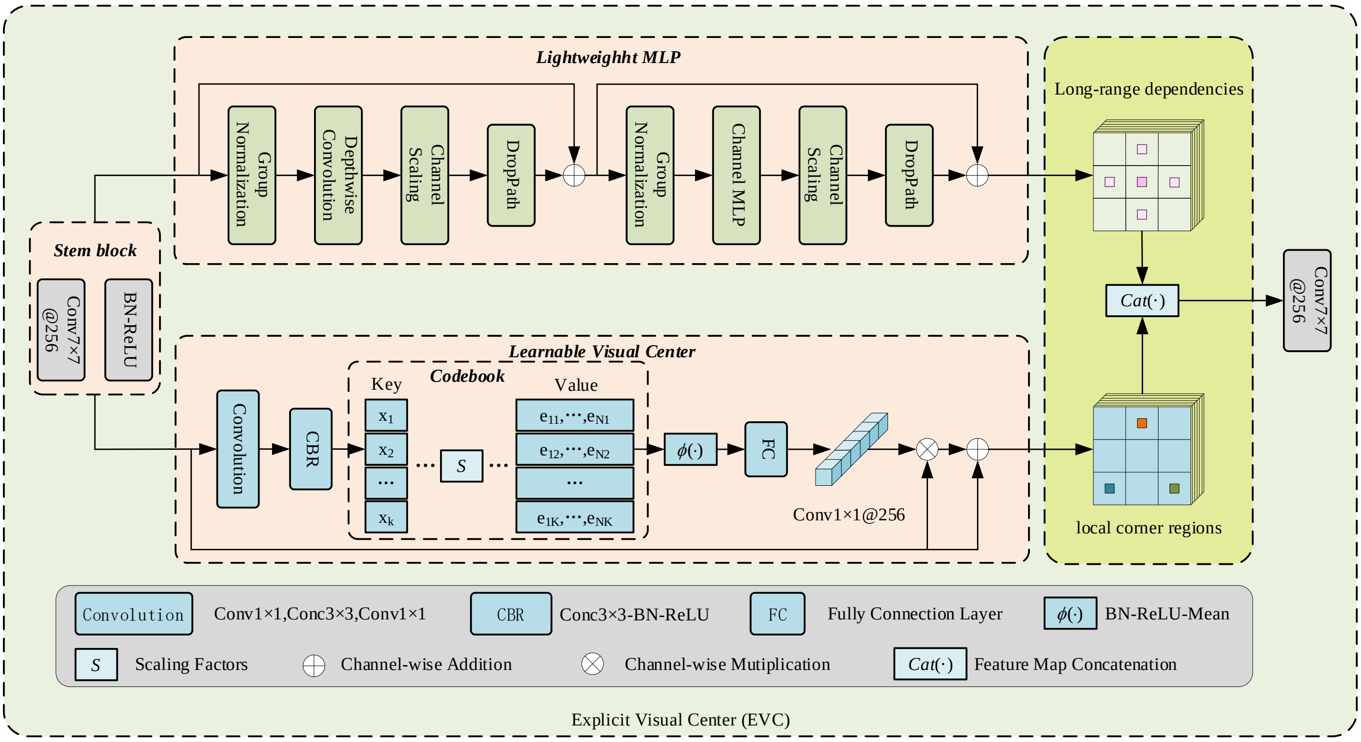

Fig. 3. Explicit visual center module structure diagram.

The lightweight MLP module is used to capture global long-term dependencies (global information), and the LVC module preserves local corner information (local information) by aggregating local features within layers. The stem module between the two modules is used for feature smoothing. The feature maps of the two modules are spliced together along the channel dimension as the output of EVC. The above processes can be formulated as:

$$\mathbf{X} = \mathrm{cat}(\mathrm{MLP}(\mathbf{X}_{\mathrm{in}}); \mathrm{LVC}(\mathbf{X}_{\mathrm{in}})) \tag{11}$$

Here, $\mathbf{X}$ represents the output of EVC, cat(·) represents parallel concatenation of feature maps along the channel dimension, MLP($\mathbf{X}_{\mathrm{in}}$) and LVC($\mathbf{X}_{\mathrm{in}}$) represent the output features of the lightweight MLP and learnable visual center used, respectively. $\mathbf{X}_{\mathrm{in}}$ is the output of the Stem block, which is obtained by:

$$\mathbf{X}_{\mathrm{in}} = \sigma(\mathrm{BN}(\mathrm{Conv}7\times 7(\mathbf{V}))) \tag{12}$$

Here, Conv7×7(·) represents the 7×7 convolution with a stride of 1 and the channel size is set to 256, BN(·) represents the batch normalization, and $\sigma(\cdot)$ represents the ReLU activation function.

The lightweight MLP mainly consists of two residual modules: a depthwise convolution-based module and a channel MLP-based module, where the input of the channel MLP-based module is the output of a depthwise convolution-based module. The two modules finally pass channel scaling operations and DropPath operations to improve feature generalization and robustness capabilities. for the depthwise convolution-based module, the features output from the Stem module $\mathbf{X}_{\text{in}}$ are first fed to the group normalization layer, and then through the depth convolution layer. Compared with the traditional spatial convolution, the depth convolution can improve the feature representation ability, and at the same time Reduce computing costs. Then, channel scaling and DropPath are performed, and finally, a residual connection of $\mathbf{X}_{\text{in}}$ is implemented. The above processes can be formulated as:

$$\tilde{\mathbf{X}}_{\text{in}} = \text{DConv}(\text{GN}(\mathbf{X}_{\text{in}})) + \mathbf{X}_{\text{in}} \tag{13}$$

Here, $\tilde{\mathbf{X}}_{\text{in}}$ is the output of the depth convolution module, $\text{GN}(\cdot)$ is the group normalization, and $\text{DConv}(\cdot)$ is the depthwise convolution with a convolution kernel of 1×1.

For the channel MLP-based module, the output features $\tilde{\mathbf{X}}_{\text{in}}$ based on the depthwise convolution-based module are first fed to the group normalization layer and then passed through the channel MLP layer. Compared with the spatial MLP, the Channel MLP can not only effectively reduce computational complexity, but also meet the requirements of general visual tasks. Then, channel scaling and DropPath are performed, and finally, the residual connection of $\tilde{\mathbf{X}}_{\text{in}}$ is implemented. The above processes can be formulated as:

$$\text{MLP}(\mathbf{X}_{\text{in}}) = \text{CMLP}(\text{GN}(\tilde{\mathbf{X}}_{\text{in}})) + \tilde{\mathbf{X}}_{\text{in}} \tag{14}$$

Here, $\text{CMLP}(\cdot)$ is the channel MLP.

LVC is an encoder with an intrinsic dictionary and has two components: 1) inherent codebook: $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K\}$, where $N = H \times W$ is the total spatial number of input features, where H and W represent the height and width of the feature map space size, respectively.; 2) Learnable visual center scale factor: $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$. The output of the stem module is first encoded through a combination of a set of convolutional layers (consisting of 1×1 convolution, 3×3 convolution, and 1×1 convolution). Then, the encoded features are processed with a CBR block, which consists of a 3×3 convolution with a BN layer and a ReLU activation function. Through the above steps, the processed encoding features $\hat{\mathbf{X}}_{\text{in}}$ are input into the Codebook, and a set of scaling factors S is used to continuously map to $\hat{\mathbf{x}}_{\text{i}}$ and $\mathbf{b}_{\text{k}}$ the corresponding position information. The information of the whole image with respect to the k-th codeword can be calculated by:

$$e_k = \sum_{i=1}^{N} \frac{e^{-S_k \|\hat{x}_i - b_k\|^2}}{\sum_{j=1}^{K} e^{-S_k \|\hat{x}_i - b_k\|^2}} (\hat{x}_i - b_k) \tag{15}$$

Here, $\hat{\mathbf{x}}_{\text{i}}$ is the i-th pixel point, $\mathbf{b}_{\text{k}}$ is the k-th learnable visual codeword, and $\mathbf{s}_{\text{k}}$ is the k-th scaling factor. $\hat{\mathbf{x}}_{\text{i}} - \mathbf{b}_{\text{k}}$ is information about the position of each pixel relative to the codeword. K is the total number of visual centers. we use ϕ to fuse all $\mathbf{e}_k$, which ϕ contains BN layer and ReLU layer and average pooling layer. the full information of the whole image with respect to the K codewords is calculated as follows.:

$$\mathbf{e} = \sum_{k=1}^{K} \phi(\mathbf{e}_k) \tag{16}$$

After obtaining the output of the codebook, $\mathbf{e}$ is further fed to a fully connected layer and a 1×1 convolutional layer to predict features that highlight key classes. After that, channel multiplication between the output $\mathbf{X}_{\text{in}}$ of the Stem block and the scaling factor coefficient δ is used, the above processes are expressed as:

$$\mathbf{Z} = \mathbf{X}_{\text{in}} \otimes (\delta(\text{Conv1}\times 1(\mathbf{e}))) \tag{17}$$

Here, Conv1×1 is a 1×1 convolution, and $\delta(\cdot)$ is a sigmoid function. $\otimes$ is channel-wise multiplication. Finally, we perform a channel-wise addition between the Stem block output $\mathbf{X}_{\text{in}}$ and the local area feature $\mathbf{Z}$, the above processes are expressed as:

$$\text{LVC}(\mathbf{X}_{\text{in}}) = \mathbf{X}_{\text{in}} \oplus \mathbf{Z} \tag{18}$$

Here, $\oplus$ is the channel-wise addition.

3.3 Loss Function

This paper uses offline and online comparative learning losses. For image $\mathbf{X}_{\text{i}}$, a global feature $\mathbf{g}_i$ and a set of local features $\{h^k\}_{k=1}^{K}$ were extracted, offline association directly retrieves a positive proxy set P_1 based on the offline clustering and segmentation results, while obtaining a negative proxy set Q_1 from the remaining hard negative proxies. P_1 and Q_1 store the indexes of related positive and negative agents respectively. Then the offline contrastive loss is defined as:

$$L_{off}^{g} = -\sum_{i=1}^{B} \left(\frac{1}{|P_1|} \sum_{u \in P_1} \log \frac{S(u, \mathbf{g}_i)}{\sum_{p \in P_1} S(p, \mathbf{g}_i) + \sum_{q \in Q_1} S(q, \mathbf{g}_i)} \right) \tag{19}$$

Here, $S(u, \mathbf{g}_i) = \exp(K[u]^T \mathbf{g}_i / \tau)$, K represents the agent-level memory bank, τ is the temperature factor, $\mathbf{g}_i$ represents the global feature, $|\cdot|$ represents the cardinality of the set, and B represents the batch size.

However, offline correlations are noisy due to imperfect clustering results. To eliminate the noise, an online correlation strategy is proposed. It adopts an instance-agent balanced similarity and camera-aware nearest neighbor scheme to dynamically select a positive proxy set P_2 and a negative proxy set Q_2 for each anchor image X_i. Then the online contrastive loss is defined as:

$$L_{on}^{g} = -\sum_{i=1}^{B}\left(\frac{1}{|P_2|}\sum_{u\in P_2}\log\frac{S(u,\mathbf{g}_i)}{\sum\limits_{p\in P_2}S(p,\mathbf{g}_i)+\sum\limits_{q\in Q_2}S(q,\mathbf{g}_i)}\right) \tag{20}$$

Since the loss function introduced above is defined based on global features, in order to utilize local features, an additional memory library is constructed and two types of losses are defined for each local information while keeping the clustering steps unchanged. Therefore, the entire loss for training is as follows:

$$L = L_{off}^{g} + L_{on}^{g} + \lambda_p \frac{1}{K}\sum_{k=1}^{K}(L_{off}^{g} + L_{on}^{g}) \tag{21}$$

Here, λ_p is the weighting factor that balances global loss and local loss.

4. RESULTS AND ANALYSIS

4.1 Datasets and Evaluation Metrics

To verify the effectiveness of the proposed method, validation were conducted on the datasets Market1501 (Zheng et al., 2015), DukeMTMC-reID (Ristani et al., 2016) and MSMT17 (Wei et al., 2018). The two datasets, Market1501 and DukeMTMC-reID, only capture outdoor scenes on university campuses, while MSMT17 contains both indoor and outdoor scenes, making it more challenging. Table 2 lists the number of cameras in the three datasets and the number of identities and images contained in the training set, query set and gallery set. The training set is used for unsupervised learning. During the testing phase, each image in the query set is matched with similar images in the gallery set.

Table 2. The number of cameras, IDs and images on three datasets.

Dataset	Training Set			Query Set		Gallery Set	
	#Camera	#ID	#Image	#ID	#Image	#ID	#Image
Market1501	6	751	12936	750	3368	751	15913
DukeMTMC-reID	8	702	16522	702	2228	1110	17661
MSMT17	15	1041	32621	3060	11659	3060	82161

The widely used two evaluation indicators of mean average precision (mAP) and cumulative matching characteristic curve (Cumulative Matching Characteristic, CMC) are used to evaluate the performance of the results. CMC evaluates the results through Rank-1, Rank-5, and Rank-10.

4.2 Implementation Details

The operating system is Linux 540136generic, the graphics card model is NVIDIA GeForce RTX 3090 Ti, and the video memory is 24GB. The pytorch framework is used to complete the model training. During training, the update rate μ=0.2, the temperature factor τ = 0.07, the batch_size is set to 32, and the weight of the local loss is λ_p = 0.1. The model was trained by the SGD optimizer 50 times with a momentum of 0.9, a learning rate of 0.00035, and a weight decay of 0.0005.

4.3 Ablation Study

To verify the impact of different attention mechanisms on this results, attention mechanisms such as SE (Hu et al., 2018), CBAM (Woo et al., 2018), and SPA (Woo et al., 2018) were introduced. The results are shown in Table 3. The channel attention mechanism is better than the spatial attention mechanism in performance. Compared with SPA, the CBAM attention mechanism is better than SPA in every indicator. Compared with SE, the DCA used in this article has an evaluation index that is about 0.3% higher on the dataset Market1501 and DukeMTMC-reID, and an evaluation index of about 1.5% higher on the dataset MSMT17.

Table 3. Results of ablation study on mainstream datasets using different attention mechanisms.

Method	Market1501				DukeMTMC-reID				MSMT17			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
SE	89.3	95.3	97.9	98.6	76.6	86.7	92.6	94.1	57.8	82.5	88.9	90.5
CBAM	89.0	95.2	98.0	98.4	76.4	86.6	92.2	94.0	56.2	79.5	87.7	89.8
SPA	88.5	94.8	97.8	98.5	75.7	85.6	92.5	93.8	55.7	79.2	87.5	89.6
DCA	**89.7**	**95.8**	**98.1**	98.6	**76.8**	**86.9**	**93.1**	**94.4**	**59.1**	**83.6**	**90.5**	**92.5**

To evaluate the impact of lightweight MLP and LVC in the EVC module on person recognition accuracy, ablation study were performed on the Market1501, DukeMTMC-reID and MSMT17 datasets respectively. The mAP, Rank-1, Rank-5 and Rank-10 obtained by each group are shown in Table 4. It can be seen from the results that compared with the benchmark model, lightweight MLP and LVC both improve the evaluation index accordingly. The improvement effect of LVC is more obvious than that of lightweight MLP. When lightweight MLP and LVC are connected in parallel (EVC module), the evaluation index reaches the optimal level compared with the two components of lightweight MLP and LVC.

To verify the effectiveness of each component in this article's model, ablation study were conducted using a single query mode on the Market1501, DukeMTMC-reID and MSMT17 datasets. The mAP, Rank-1, Rank-5 and Rank-10 obtained by each group are shown in Table 5. The DCA, EVC and DCA+EVC modules were introduced into the baseline model for ablation study, which improved the recognition accuracy of person images. For the dataset Market1501, the introduction of the DCA module based on the baseline model increased mAP by 0.2%, Rank-1 and Rank-5 increased by 0.35% and 0.1% respectively, and Rank-10 decreased by 0.1%; for the dataset DukeMTMC- reID, the DCA module was introduced based on the baseline model to keep mAP unchanged, and Rank-1, Rank-5 and Rank-10 increased by 0.1%, 0.2% and 0.3% respectively; for the dataset MSMT17, based on the baseline model The introduction of the DCA module increased mAP by 0.9%, and Rank-1, Rank-5 and Rank-10 increased by 0.3%, 0.3% and 0.4% respectively. For the dataset Market1501, the introduction of the EVC module based on the baseline model increased mAP by 1.4%, and Rank-1, Rank-5 and Rank-10 increased by 0.9%, 0.6% and 0.4% respectively; for the dataset DukeMTMC-reID , the introduction of the EVC module based on the baseline model increased mAP by 1.3%, and Rank-1, Rank-5 and Rank-10 increased by 0.9%, 1.3% and 0.5% respectively; for the dataset MSMT17, based on the baseline model The introduction of the EVC module increased mAP by 1.1%, and Rank-1, Rank-5 and Rank-10 increased by 1.1%, 0.7% and 0.9% respectively. For the dataset Market1501, the introduction of the DCA+EVC module based on the baseline model increased mAP by 1.6%, and Rank-1, Rank-5 and Rank-10 increased by 0.9%, 1.0% and 0.5% respectively; for the dataset DukeMTMC -reID, introducing the DCA+EVC module based on the baseline model increased mAP by 1.5%, Rank-1, Rank-5 and Rank-10 increased by 1.4%, 1.0% and 0.6% respectively; for the dataset MSMT17, in The introduction of the DCA+EVC module based on the baseline model increased mAP by 1.6%, and Rank-1, Rank-5 and Rank-10 increased by 1.3%, 0.9% and 1.4% respectively.

Table 4. Ablation study results of MLP, LVC and EVC on mainstream datasets.

Method	Market1501				DukeMTMC-reID				MSMT17			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
MLP	90.1	95.7	98.3	98.8	77.3	87.1	93.3	94.3	58.5	83.6	90.4	92.3
LVC	90.5	96.2	98.6	99.0	77.8	87.4	94.1	94.5	58.9	84.1	90.8	92.8
EVC	**90.9**	**96.4**	**98.7**	**99.1**	**78.1**	**87.6**	**94.2**	**94.6**	**59.3**	**84.4**	**90.9**	**93.0**

Table 5. Results of ablation study on datasets.

Method	Market1501				DukeMTMC-reID				MSMT17			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Baseline	89.5	95.5	98.0	98.7	76.8	86.7	92.9	94.1	58.2	83.3	90.2	92.1
DCA	89.7	95.8	98.1	98.6	76.8	86.9	93.1	94.4	59.1	83.6	90.5	92.5
EVC	90.9	96.4	98.7	99.1	78.1	87.6	**94.2**	94.6	59.3	84.4	90.9	93.0
DCA+EVC	**91.1**	96.4	**99.0**	**99.2**	**78.3**	**88.1**	93.9	**94.7**	**59.8**	**84.6**	**91.1**	**93.5**

4.4 Comparative Study

To further verify the effectiveness of this algorithm, we compared it with a variety of advanced person re-identification algorithms such as BUC, HCT (Zeng et al., 2020), CAP (Wang et al., 2021), RLCC (Zhang et al., 2021), MGH (Wu et al., 2021), MCRN (Wu et al., 2022), MGCE-HCL (Sun et al., 2021), PPLR (Cho et al., 2022), etc. on three mainstream datasets: Market1501, DukeMTMC-reID and MSMT17. The results are shown in Table 6. It can be seen from the data in the table that the method proposed in this article has a certain improvement in performance. The mAP, Rank-1, Rank-5 and Rank-10 on the Market1501 data set are 91.1%, 96.4% and 99.0 respectively. % and 99.2%. Compared with the PPLR algorithm with the highest recognition accuracy, mAP, Rank-1, Rank-5 and Rank-10 have increased by 6.7%, 2.1%, 1.2% and 0.5% respectively; on the DukeMTMC-reID dataset mAP, Rank-1, Rank-5 and Rank-10 are 78.3%, 88.1%, 93.9% and 94.7% respectively. Compared with the MGH algorithm with the highest recognition accuracy, mAP, Rank-1, Rank-5 and Rank-10 are respectively Increased by 8.1%, 4.4%, 1.8% and 1.0%; mAP, Rank-1, Rank-5 and Rank-10 on the MSMT17 dataset were 59.8%, 84.6%, 91.1% and 93.5% respectively, compared to recognition For the PPLR algorithm with the highest accuracy, mAP, Rank-1, Rank-5 and Rank-10 increased by 17.6%, 11.3%, 7.6% and 7.0% respectively. The evaluation indicators have been improved accordingly on the three mainstream datasets, indicating that the method in this paper is effective in improving the performance of the person re-identification task.

In order to verify the generalization of the proposed method in the neural network model, CNN is used as the backbone network, and experiments are conducted on three mainstream data sets: Market1501, DukeMTMC-reID and MSMT17. CNN+MGN is used as the benchmark network. By adding different modules Verification is carried out and the results are shown in Table 7. the results show that the method proposed in this article has certain improvement in performance.

Table 6. Performance comparison of different models on datasets.

Method	Market1501				DukeMTMC-reID				MSMT17			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
BUC	38.3	66.2	79.6	84.5	27.5	47.4	62.6	68.4	-	-	-	-
HCT	56.4	80.0	91.6	95.2	50.7	69.6	83.4	87.4	-	-	-	-
CAP	79.2	91.4	96.3	97.7	67.3	81.1	89.3	91.8	36.9	67.4	78.0	81.4
RLCC	77.7	90.8	96.3	97.5	69.2	83.2	91.6	93.8	27.9	56.5	68.4	73.1
MGH	81.7	93.2	96.8	98.1	70.2	83.7	92.1	93.7	40.6	70.2	81.2	84.5
MCRN	80.8	92.5	-	-	69.9	83.5	-	-	31.2	63.6	-	-
MGCE-HCL	79.6	92.1	-	-	67.5	82.5	-	-	-	-	-	-
PPLR	84.4	94.3	97.8	98.7	-	-	-	-	42.2	73.3	83.5	86.5
TMGF	89.5	95.5	98.0	98.7	76.8	86.7	92.9	94.1	58.2	83.3	90.2	92.1
Ours	**91.1**	**96.4**	**99.0**	**99.2**	**78.3**	**88.1**	**93.9**	**94.7**	**59.8**	**84.6**	**91.1**	**93.5**

“-” indicates no results reported

Table 7. Results of CNN-based model.

Method	Market1501				DukeMTMC-reID				MSMT17			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Baseline	62.5	72.5	76.9	77.6	53.9	60.8	65.9	66.8	48.3	53.2	60.6	62.5
DCA	62.7	73.2	77.1	78.8	56.4	66.5	69.1	70.6	50.1	56.7	65.6	68.3
EVC	64.3	75.8	78.2	80.1	58.4	68.6	71.2	73.8	54.9	59.2	66.8	70.4
DCA+EVC	**65.8**	**77.1**	**79.6**	**81.2**	**60.5**	**71.3**	**72.8**	**75.1**	**55.7**	**60.4**	**68.6**	**72.5**

5. CONCLUSION

In order to solve the problem of insufficient person feature extraction in the unsupervised person re-identification process, this paper improves the TMGF network model and proposes a Transformer-based multi-Grained feature aggregation unsupervised person re-identification MGFA model. First, a dual-channel attention module is designed to realize the interaction and correlation between different channels to extract more critical information of person images; then an explicit visual center module is proposed to capture global information and aggregate key local information to enhance the characteristics of the network representation, thereby improving the generalization ability of the model. In the ablation study, compared with the baseline model, mAP, Rank-1, Rank-5 and Rank-10 on the dataset Market1501 increased by 1.6%, 0.9%, 1.0% and 0.5% respectively; on the dataset DukeMTMC-reID, mAP, Rank-1, Rank-5 and Rank-10 increased by 1.5%, 1.4%, 1.0% and 0.6% respectively; on the dataset MSMT17, mAP, Rank-1, Rank-5 and Rank-10 increased by 1.6%, 1.3%, 0.9% and 1.4% respectively. Therefore, this method can effectively extract person features and the recognition effect is more obvious.

REFERENCES

Chen, B., Deng, W., & Hu, J. (2019a). Mixed high-order attention network for person re-identification. *In Proceedings of the IEEE/CVF international conference on computer vision*, pp. 371-381.

Chen, H., Lagadec, B., & Bremond, F. (2021). Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 14960-14969.

Cheng, D., Gong, Y., Zhou, S., Wang, J., & Zheng, N. (2016). Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the iEEE conference on computer vision and pattern recognition* pp. 1335-1344.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* pp. 1597-1607. PMLR.

Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., ... & Wang, Z. (2019b). Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* pp. 8351-8361.

Chen, Y., Zhu, X., & Gong, S. (2018). Deep association learning for unsupervised video person re-identification. *arXiv preprint arXiv*:1808.07301.

Cho, Y., Kim, W. J., Hong, S., & Yoon, S. E. (2022). Part-based pseudo label refinement for unsupervised person re-identification. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7308-7318.

Dai, Z., Wang, G., Yuan, W., Zhu, S., & Tan, P. (2022). Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision* pp. 1142-1160.

Dong, W., Qu, P., & Li, B. (2023). Dual Pseudo Label Refinement for Unsupervised Domain Adaptive Person Re-identification. *IEEE Access*.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition pp. 9729-9738.

He, S., Luo, H., Wang, P., Wang, F., Li, H., & Jiang, W. (2021). Transreid: Transformer-based object re-identification. *In Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15013-15022.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132-7141.

Huang, Y., Lian, S., Zhang, S., Hu, H., Chen, D., & Su, T. (2020). Three-dimension transmissible attention network for person re-identification. *IEEE transactions on circuits and systems for video technology*, 30(12), 4540-4553.

Lai, S., Chai, Z., & Wei, X. (2021). Transformer meets part model: Adaptive part division for person re-identification. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4150-4157.

Li, M., Zhu, X., & Gong, S. (2019). Unsupervised tracklet person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 42(7), 1770-1782.

Li, J., Wang, M., & Gong, X. (2023). Transformer Based Multi-Grained Features for Unsupervised Person Re-Identification. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 42-50.

Li, W., Zhu, X., & Gong, S. (2018). Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 2285-2294.

Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., & Wu, F. (2021). Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 2898-2907.

Lin, Y., Dong, X., Zheng, L., Yan, Y., & Yang, Y. (2019). A bottom-up clustering approach to unsupervised person re-identification. *In Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, No. 01, pp. 8738-8745.

Lin, Y., Xie, L., Wu, Y., Yan, C., & Tian, Q. (2020). Unsupervised person re-identification via softened similarity learning. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3390-3399.

Luo, H., Gu, Y., Liao, X., Lai, S., & Jiang, W. (2019). 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* pp. 0-0.

Luo, H., Wang, P., Xu, Y., Ding, F., Zhou, Y., Wang, F., ... & Jin, R. (2021). Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv*:2111.12084.

Ma, Z., Zhao, Y., & Li, J. (2021). Pose-guided inter-and intra-part relational transformer for occluded person re-identification. *In Proceedings of the 29th ACM international conference on multimedia*, pp. 1487-1496.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. *In European conference on computer vision*, pp. 17-35.

Shi, C., Niu, D., Gong, H., Zhang, M., Cao, Z., & Jin, Y. (2023). Person Re-identification Lightweight Network Based on Progressive Attention Mechanism. In 2023 *6th International Symposium on Autonomous Systems* (ISAS) pp. 1-6. IEEE.

Sharma, C., Kapil, S. R., & Chapman, D. Person re-identification with a locally aware transformer. arXiv 2021. *arXiv preprint arXiv*:2106.03720.

Si, J., Zhang, H., Li, C. G., Kuen, J., Kong, X., Kot, A. C., & Wang, G. (2018). Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 5363-5372.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision* (ECCV) pp. 480-496.

Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., & Sun, J. (2019). 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*, pp. 393-402.

Sun, H., Li, M., & Li, C. G. (2021). Hybrid contrastive learning with cluster ensemble for unsupervised person re-identification. *In Asian Conference on Pattern Recognition*, pp. 532-546.

Wang, G., Yuan, Y., Chen, X., Li, J., & Zhou, X. (2018). Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia* pp. 274-282.

Wang, M., Lai, B., Huang, J., Gong, X., & Hua, X. S. (2021). Camera-aware proxies for unsupervised person re-identification. *In Proceedings of the AAAI conference on artificial intelligence*, Vol. 35, No. 4, pp. 2764-2772.

Wang, M., Li, J., Lai, B., Gong, X., & Hua, X. S. (2022). Offline-online associated camera-aware proxies for unsupervised person re-identification. *IEEE Transactions on Image Processing*, 31, 6548-6561.

Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer gan to bridge domain gap for person re-identification. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 79-88.

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. *In Proceedings of the European conference on computer vision (ECCV)*, pp. 3-19.

Wu, J., Yang, Y., Liu, H., Liao, S., Lei, Z., & Li, S. Z. (2019). Unsupervised graph association for person re-identification. *In Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8321-8330.

Wu, L., Shen, C., & Hengel, A. V. D. (2016). Personnet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv*:1601.07255.

Wu, Y., Wu, X., Li, X., & Tian, J. (2021). MGH: Metadata guided hypergraph modeling for unsupervised person re-identification. *In Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1571-1580.

Wu, Y., Huang, T., Yao, H., Zhang, C., Shao, Y., Han, C., ... & Sang, N. (2022). Multi-centroid representation network for domain adaptive person re-id. *In Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, No. 3, pp. 2750-2758.

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 3733-3742.

Yang, Q., Yu, H. X., Wu, A., & Zheng, W. S. (2019). Patch-based discriminative feature learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 3633-3642.

Yang, Z., Jin, X., Zheng, K., & Zhao, F. (2022). Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 14298-14307.

Zeng, K., Ning, M., Wang, Y., & Guo, Y. (2020). Hierarchical clustering with hard-batch triplet loss for person re-identification. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13657-13665.

Zhang, X., Ge, Y., Qiao, Y., & Li, H. (2021). Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3436-3445.

Zhang, G., Zhang, P., Qi, J., & Lu, H. (2021). Hat: Hierarchical aggregation transformers for person re-identification. *In Proceedings of the 29th ACM International Conference on Multimedia*, pp. 516-525.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. *In Proceedings of the IEEE international conference on computer vision*, pp. 1116-1124.

Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., ... & Ji, R. (2019). Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 8514-8522.

Zhou, Z., Liu, H., Shi, W., Tang, H., & Shi, X. (2022). A Cloth-Irrelevant Harmonious Attention Network for Cloth-Changing Person Re-identification. In 2022 *26th International Conference on Pattern Recognition (ICPR) pp. 989-995. IEEE.*

Zhu, K., Guo, H., Zhang, S., Wang, Y., Liu, J., Wang, J., & Tang, M. (2023). Aaformer: Auto-aligned transformer for person re-identification. *IEEE Transactions on Neural Networks and Learning Systems.*