applied sciences

MDPI

Article

An Adaptive Partitioning and Multi-Granularity Network for Video-Based Person Re-Identification

Bailiang Huang, Yan Piao * and Yanfeng Tang

School of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun 130012, China
* Correspondence: piaoyan@cust.edu.cn

Abstract: Person re-identification (Re-ID) is a key technology used in the field of intelligent surveillance. The existing Re-ID methods are mainly realized by using convolutional neural networks (CNNs), but the feature information is easily lost in the operation process due to the down-sampling structure design in CNNs. Moreover, CNNs can only process one local neighbourhood at a time, which makes the global perception of the network poor. To overcome these shortcomings, in this study, we apply a pure transformer to a video-based Re-ID task by proposing an adaptive partitioning and multi-granularity (APMG) network framework. To enable the pure transformer structure better at adapting to the Re-ID task, we propose a new correlation-adaptive partitioning (CAP) of feature embedding modules that can adaptively partition person images according to structural correlations and thus retain the structure and semantics of local feature information in the images. To improve the Re-ID performance of the network, we also propose a multi-granularity (MG) module to better capture people feature information at different levels of granularity. We performed validation trials on three video-based benchmark datasets. The results show that the network structure based on the pure transformer can adapt to Re-ID tasks well, and our APMG network outperforms other state-of-the-art methods.

Keywords: machine vision; deep learning; video-based Re-ID; transformer

Citation: Huang, B.; Piao, Y.; Tang, Y. An Adaptive Partitioning and Multi-Granularity Network for Video-Based Person Re-Identification. *Appl. Sci.* **2022**, *12*, 12503. https://doi.org/10.3390/app122312503

Academic Editor: Antonio Sarasa Cabezuelo

Received: 10 November 2022
Accepted: 2 December 2022
Published: 6 December 2022

1. Introduction

The rapid increase in urbanisation in recent years has resulted in increasing numbers of people migrating to cities. The dense populations in cities are thus a major challenge for urban security management, which is addressed by the continuous deployment in cities of security camera monitoring equipment. This equipment enables cities to be searched for targets or the movements of suspects to be tracked by manually reviewing surveillance images. However, the continuous increase in the extent of camera monitoring means that massive amounts of monitoring data are generated every day, which are difficult to process manually. Re-ID technology, which was developed in the field of intelligent monitoring, is designed to perform such processing tasks, consisting of image-based Re-ID tasks and video-based Re-ID tasks. Compared with image-based tasks, video-based tasks have a higher research value as they produce greater continuity between frames before and after an image, they can provide dynamic information on a person, and they are more similar to a real scene. Therefore, the network framework proposed in this paper is designed for video-based Re-ID tasks.

Most of the existing approaches to video-based Re-ID tasks use a CNN as a backbone network. However, with the continuous application and development of transformer network structures in image classification [1,2], object detection [3,4], and other computer vision fields, experiments have shown that a transformer is as effective as a CNN in extracting feature information from images [2,5]. A transformer network structure was originally proposed for natural language processing (NLP) tasks [6], where each word

is converted into a word vector and used as input for parallel computation. In the field of computer vision, the network with a transformer structure partitions the image and converts it into multiple feature embeddings that serve as input. A transformer differs from a CNN by having no down-sampling structure and thus effectively preserves the original feature information in an image. Moreover, a CNN increases the receptive field of a network by superposition convolution, whereas the structural design of a transformer has the characteristics of the global receptive field, and its self-attention mechanism has a better representation ability for managing the extracted feature information through high-order correlation modelling. Considering these significant advantages, we apply a pure transformer as the backbone for our APMG network framework.

In networks using a transformer as a backbone structure [2,7–9], the image is usually partitioned by a fixed-size sliding window to generate the input feature embeddings. There are two problems with the use of such transformer networks for image feature extraction. First, the rigid partitioning approach may destroy the structure of local features of human images in the instance-level image Re-ID task. The key to improving re-identification accuracy is to let the network learn the various feature information of people, as changes in the local structure of people inevitably affect the ability of a network to extract feature information. Furthermore, in each frame of a video sequence, a change in the shooting angle or the displacement of people may cause the fixed segmentation method to change the content captured in the sliding window, resulting in inconsistent semantic information on feature embedding [10]. Second, the self-attention structure in a transformer has strong global awareness, so a network usually takes the class token learning of the global feature information as the final output [5]. This may ignore some fine-grained feature information in an image, thereby limiting the re-identification performance of a network. Therefore, we think that the transformer needs to be improved to better adapt to the image Re-ID task.

To solve the first problem, this paper proposes a simple and effective CAP module to partition the feature embedding modules, which performs the partitioning, as shown in Figure 1. In the CAP module, the closely correlated parts in the window are partitioned together, which preserves the structural integrity of the local feature information in the person image and maintains the consistency of the semantic information before and after the local feature segmentation. We also use two sets of parameters to learn the correlation of calculations in the window, such that the partitioning of feature embeddings in different windows is adaptive. To solve the second problem, this paper proposes an MG module, in which the feature embeddings are cut horizontally into groups and the class tokens are embedded in different groups to learn the feature information at different levels of granularity. We also design an aggregation method that is incorporated into the MG module, which aggregates the same levels of granularity in the video sequence that are present in the final feature representation. Thus, overall, this paper makes the following four main contributions:

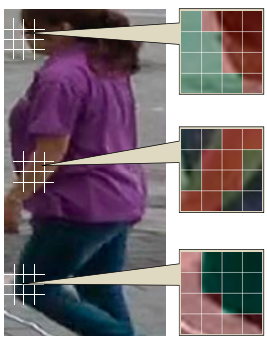

Figure 1. Correlation-adaptive partitioning of feature embeddings.

1. We propose a pure transformer-based APMG network for video Re-ID tasks;
2. We propose a CAP module to adapt the partitioning of feature embeddings according to the correlation, which maintains the consistency of the overall structure and the semantic information of feature embeddings;
3. We propose an MG module that can capture feature information at different levels of granularity, thereby improving the re-identification performance of the network.
4. We perform experiments on three widely used video Re-ID datasets, which demonstrated that our APMG network achieves a better performance than other state-of-the-art models.

The rest of this paper is arranged as follows: In Section 2, we review the related work on Re-ID and transformer. In Section 3, we propose an APMG network based on the pure transformer structure, and introduce two brand-new modules, CAP and MG, in detail. In Section 4, we introduce three datasets for experimental testing and make comprehensive analyses of the experimental results. In Section 5, we summarize the work content of this paper and look forward to future work.

2. Related Work

The key to network design for Re-ID tasks, involving an image dataset, is to enable the extraction of the distinguishing feature information of people [11–13]. However, in a video sequence dataset, people have multiple and continuous image data, which constitutes additional information about people in time and space. Therefore, in recent years, new methods have been developed for video-based Re-ID tasks to alleviate the negative impact on Re-ID performance of realistic problems, such as person occlusion and person pose change, by focusing on extracting the spatial or temporal features of people from video sequences. For example, the graph convolution network (GCN) structure is used to establish connections between an image in each frame [14–17]. The spatial-temporal GCN network (STGCN) [18] divides person-feature information into multiple feature points horizontally and then uses the GCN structure to construct the feature points of multiple person images in the video sequence into spatial and temporal branches. Subsequently, the STGCN extracts the spatiotemporal information that complements the appearance information of a person, which greatly improves the performance, relative to that of a GCN. Another mainstream method, to enhance the ability of a network to extract spatiotemporal person information, is the attention mechanism [19–22]. In the salient-to-broad module designed in the scale-insensitive convolutional neural network (SINet) [23], for example, differential amplification is used to enhance the difference between each image in each frame and thereby improve the ability of the network to extract the feature information from image sequences. Three-dimensional convolution is also used to capture and fuse temporal person-feature information [24–27]. However, a network using a 3D convolution structure has high computational demands, and it is easy to overfit and difficult to train. Other approaches, such as extracting person-attribute information [28,29] and extracting fine-grained person information [30,31], have also been developed for enhancing a network's ability to collect different feature information and thus achieve an enhanced performance.

One of the mainstream methods in the field of NLP uses a transformer structure [6], which captures the global dependency of sequences. Vision Transformer [5] was the first network to apply transformer backbone structures to image tasks. The transformer structure retains the original feature information better than a CNN, which has led to transformers being applied in the field of computer vision [32,33]. To keep the semantic information before and after the image split Zhiyang et al. [10] proposed variable-size windows to split patches. At present, the main improvement method involves combining a CNN and a transformer by adding convolution structures to a self-attention mechanism to improve the feature re-identification ability of a network [34,35]. Shuting et al. [7] applied the transformer structure to the Re-ID task for the first time and introduced a variable to learn and compensate for the image differences caused by different cameras. A few studies have applied transformers in video-based Re-ID tasks. Tianyu et al. [36] introduced

a global attention learning branch to learn the relationship between patches in an image in each frame, and Xianghao et al. [37] extracted feature information at different scales and different at levels of granularity by partitioning feature embedding from different directions. However, these studies have used a fixed-sized sliding window to segment person images for generating input feature embeddings, which ignores the consistency of semantic information before and after an overall structure and the local feature segmentation of people. To the best of our knowledge, our APMG network is the first to partition feature embeddings according to correlations in Re-ID tasks.

3. Model and Methods

3.1. Model Frame Structure

Our backbone network for video-based Re-ID tasks follows the structural design of a pure transformer used in the Vision Transformer [5] and has the overall framework as shown in Figure 2. Based on the characteristics of a strong overall structure of images, we design a new method to partition image-feature embedding according to their correlations. Then, to improve the re-identification accuracy and robustness of a network, we design a simple and practical multi-grained method.

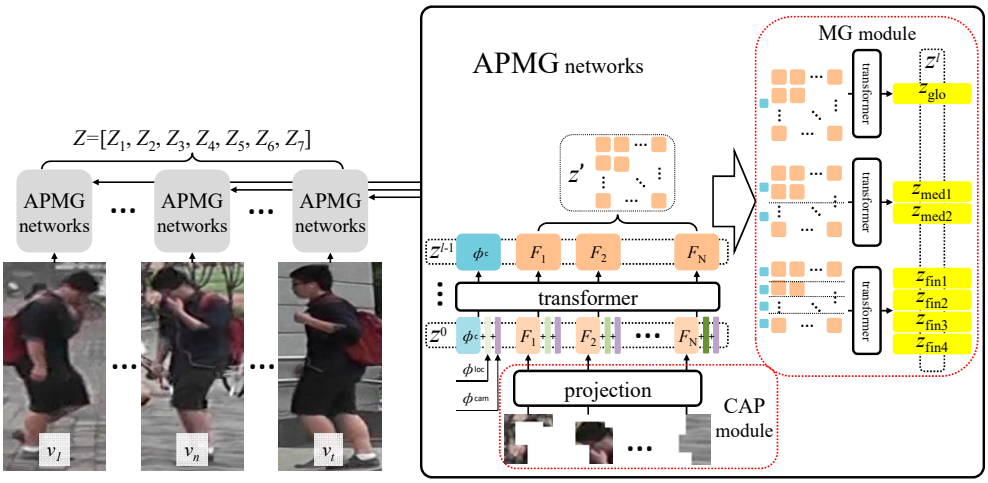

Figure 2. The overall framework of the APMG network.

The model is shown in Figure 2. Given a person video sequence v, which contains t frames of images $v = \{v_1, v_2, \ldots, v_t\}$, for any one image frame, the CAP module is first partitioned into N feature embeddings $\{F_1, F_2, \ldots, F_N\}$, where the size of each feature embedding is $1 \times c$. A learnable random variable $\Phi^c \in R^{1\times c}$ is introduced as the final feature expression of the whole image. The Φ^c and N feature embeddings are successively expanded and spliced horizontally to form the sequence $F = [\Phi^c, F_1, F_2, \ldots, F_N]$, $F \in R^{(1+N)\times c}$. Where Φ^c is placed at the top of the sequence. To compensate for the problem of the position information in the image being lost after feature embeddings are flattened, a learnable random variable $\Phi^{loc} \in R^{(N+1)\times c}$ is introduced into the image to supplement the position information of the feature embeddings. Moreover, we follow the design in TransReID [7] to introduce a variable $\Phi^{cam\prime} \in R^{1\times c}$ to save the camera information. To match the sequence size of Φ^c and N feature embeddings, $\Phi^{cam\prime}$ is copied N+1 times and changed to $\Phi^{cam} \in R^{(N+1)\times c}$. For the person image, the first layer input of the network z^0 can be expressed as follows:

$$z^0 = F + \Phi^{loc} + \Phi^{cam} \tag{1}$$

We then use the $l-1$ transformer structure to calculate the input z^0. Each transformer layer is composed of a multi-head self-attention (MSA) with a residual structure and a

multi-layer perception (MLP) with a residual structure. For layer i of the network, the output z^i can be expressed as follows:

$$z^{i\prime} = z^{i-1} + \mathrm{MSA}(z^{i-1}) \tag{2}$$

$$z^{i} = z^{i\prime} + \mathrm{MLP}(z^{i\prime}) \tag{3}$$

After $l-1$ transformer operations, the class token learns the feature information of the whole image and is finally divided by the MG module into eight class tokens with different levels of granularity.

3.2. Correlation-Adaptive Partitioning of Feature Embeddings Module

Considering the continuity of video sequences and to improve the structural integrity of patches, we propose a CAP module. The specific implementation process is as follows: For the input t-frame video sequence $v = \{v_1, v_2, \ldots, v_t\}$, we first use the $a \times a$ sliding window to preprocess each video image to generate patches without superimposed pixels $v_{\mathrm{a}} = \{v_{\mathrm{a}1}, v_{\mathrm{a}2}, \ldots, v_{\mathrm{a}t}\}$, $v_{\mathrm{a}n} \in \mathrm{R}^{h\times w\times c}$. We then use the $b \times b$ sliding window to partition the patches of each frame, p feature embeddings are partitioned from each sliding window and arranged horizontally in turn. As the structure of the two-layer sliding window enlarges the receptive field to a certain extent, we do not set the superposition region in the process of partitioning patches. For any one frame in a video sequence v, N feature embeddings are generated by the CAP module:

$$N = \frac{h}{b} \times \frac{w}{b} \times p = H \times W \times p \tag{4}$$

There are $s = b \times b$ patches in the $b \times b$ sliding window. To determine the integrity of the feature embedding within the sliding window for a patch $x_m \in \mathrm{R}^{c\times 1}$, we take the correlation among patches as the basis for inference. We consider that the strong continuity of the feature information between frames in video sequences can be used to help partition the feature embeddings in the sliding window if all the patches in the sliding window at the same position in the video sequence are taken into account. Therefore, for any one patch x_m in the sliding window of $v_{\mathrm{a}n}$, the number of patches for which the correlation is calculated is $b \times b \times t$. However, the influence of changes in light intensity and viewing angle, occlusion, and other factors means that there may be a large amount of information redundancy in the video sequence, which will affect the partitioning of feature embedding in the sliding window. Furthermore, the large number of patches distributed along the video sequence increases the computational burden of a network. We solve these two problems by adopting the method of mean operation in the direction of the video sequence. The patch sequence $v_{\mathrm{a}} = \{v_{\mathrm{a}1}, v_{\mathrm{a}2}, \ldots, v_{\mathrm{a}t}\}$ is pooled into a reference patche matrix $V_{\mathrm{a}} \in \mathrm{R}^{h\times w\times c}$. Thus, when calculating the correlation, patches in x_m only successively operate with patches in the sliding window at the same position in V_{a}. This method preserves the feature information of edge and texture structure and reduces the information redundancy in the video sequence. Moreover, the number of patches for computing correlations in the sliding window is reduced from $b \times b \times t$ to $b \times b$, which greatly reduces the computational cost of the network.

For any patches x_m from the sliding window $b \times b$ in $v_{\mathrm{a}n}$, the correlation is calculated as follows:

$$X_m = \sum_{n=1}^{s} \left[\mathrm{ReLU}\left(W_{\mathrm{x}} x_m\right)\right]^T \cdot \left[\mathrm{ReLU}(W_{\mathrm{y}} y_n)\right] \tag{5}$$

where y_n is a patch in the sliding window at the same position in V_{a}; $W_{\mathrm{x}} \in \mathrm{R}^{(c/e)\times c}$ and $W_{\mathrm{y}} \in \mathrm{R}^{(c/e)\times c}$ are two adaptive correlation learning matrices, where e is the positive integer coefficient of the compressed computation. ReLU is the linear rectification activation function. For any one $b \times b$ sliding window in van, we successively calculate the score of the correlation between patches x_m and s reference patches as the output result in

an accumulative manner. Thus, the correlation calculation result sequence of s patches is $X = \{X_1, X_2, \ldots, X_s\}$, $X_m \in \mathrm{R}^{1\times s}$. The patch x_m and correlation calculation results X_m correspond ($x_1 \rightleftharpoons X_1$). Considering the integrity of the feature embeddings within a sliding window, we regard the patches with a similar correlation score as having a stronger integrity among patches, such that they can be partitioned into one feature embedding. To achieve this, we order the correlation scores in sequence X in size, from largest to smallest, to generate the sequence $X'=\{X'_1, X'_2, \ldots, X'_s\}$, $X'_n \in \mathrm{R}^{1\times s}$. At the same time, we arrange s corresponding patches in the $b\times b$ sliding window according to the order in X', and the matrix $h \in \mathrm{R}^{1\times s\times c}$ is obtained by splicing along the transverse direction. Finally, the $1 \times (s/p)$ sliding window with a step size of s/p is used to divide h into p feature embeddings. The size of each feature embedding is $F \in \mathrm{R}^{1\times c}$.

The implementation process is shown in Figure 3, which illustrates the first sliding window result in v_{an}. For easy understanding and presentation, $b = 2$, $p = 2$ in the figure.

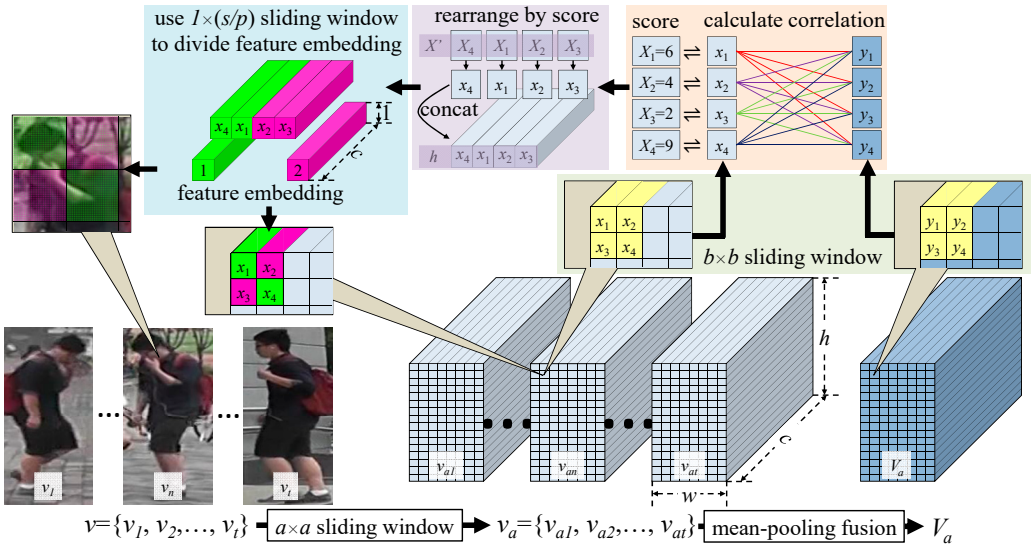

Figure 3. Schematic diagram of correlation-adaptive partitioning of feature embeddings.

3.3. Multi-Granularity Module

Due to the environmental conditions of video image acquisition, many challenges for feature extraction arise in practice, such as occlusion and dislocation. Thus, the effective feature extraction area may be a very small part of a whole image. To improve re-identification accuracy, we explore the feature information in image and video sequences at different levels of granularity and propose an MG module. Given that the overall structure of people can be divided horizontally, we use a horizontal strategy to restructure the token sequence. The rearranged tokens z_n' are divided into global-grained tokens, medium-grained tokens, and fine-grained tokens, according to the size of a window. These tokens share class token Φ^c.

For the n-th frame image in a video sequence v, the input token sequence of the last layer (layer l) is set to ${z_n}^{l-1} = [{\Phi^c}_n{}^{l-1}, {F_{n1}}^{l-1}, {F_{n2}}^{l-1}, \ldots, {F_{nN}}^{l-1}]$. We extract N feature embeddings $z_n = \{{F_{n1}}^{l-1}, {F_{n2}}^{l-1}, \ldots, {F_{nN}}^{l-1}\}$ and rearrange them according to the original corresponding positions in the image to obtain $z_n' \in \mathrm{R}^{H\times(W\times p)}$:

$$z_n = \{{F_{n1}}^{l-1}, {F_{n2}}^{l-1}, \cdots, {F_{nN}}^{l-1}\} \overset{\text{Rearrange}}{\rightarrow} z_n' = \begin{bmatrix} {F_{n1}}^{l-1} & {F_{n2}}^{l-1} & \cdots & {F_{n(W\times p)}}^{l-1} \\ {F_{n(W\times p+1)}}^{l-1} & {F_{n(W\times p+2)}}^{l-1} & \cdots & {F_{n(2\times W\times p)}}^{l-1} \\ \vdots & \vdots & \ddots & \vdots \\ {F_{n((H-1)\times W\times p+1)}}^{l-1} & {F_{n((H-1)\times W\times p+2)}}^{l-1} & \cdots & {F_{nN}}^{l-1} \end{bmatrix} \quad (6)$$

For global-grained tokens, we use a window of size $H \times (W \times p)$ to partition $z_n{}'$ into a whole. ${\Phi^c}_n{}^{l-1}$ and $z_n{}'$ are flattened together along the horizontal direction and then spliced to form the global-grained token sequence $z_{glo,n}{}^{l-1} = [{\Phi^c}_n{}^{l-1}, F_{n1}{}^{l-1}, F_{n2}{}^{l-1}, \ldots, F_{nN}{}^{l-1}]$, $z_{glo,n}{}^{l-1} \in \mathrm{R}^{(1+N)\times c}$. After l_{glo} layers of backbone transformer operations, the final output global-grained token sequence is obtained: $z_{glo,n}{}^{l} = [{\Phi^c}_n{}^{l}, F_{n1}{}^{l}, F_{n2}{}^{l}, \ldots, F_{nN}{}^{l}]$, $z_{glo,n}{}^{l} \in \mathrm{R}^{(1+N)\times c}$. Following previous studies, we use the class token as feature representations of the token sequence. For the video sequence v, the set of global-grained tokens from t-frame images can be obtained as follows:

$$z_{glo} = \left\{ {\Phi^c}_{glo,1}{}^{l}, {\Phi^c}_{glo,2}{}^{l}, \cdots, {\Phi^c}_{glo,t}{}^{l} \right\} \tag{7}$$

For medium-grained tokens, we use a window of size $(H/2) \times (W \times p)$ to divide $z_n{}'$ into two parts. ${\Phi^c}_n{}^{l-1}$ is combined with each of the two parts and, after expansion, two groups of medium-grained token sequences, $z_{\mathrm{med1},n}{}^{l-1}$ and $z_{\mathrm{med2},n}{}^{l-1}$, are obtained. The first medium-grained token sequence is $z_{\mathrm{med1},n}{}^{l-1} = [{\Phi^c}_n{}^{l-1}, F_{n,1}{}^{l-1}, F_{n,2}{}^{l-1}, \ldots, F_{n,(N/2)}{}^{l-1}]$. Through the l_{med} layer operation of the backbone transformer, class tokens learn the relationship and feature information between medium-grained tokens. Finally, two groups of medium-grained token sequences, $z_{\mathrm{med1},n}{}^{l}$ and $z_{\mathrm{med2},n}{}^{l}$, are outputs. We use a class token as the feature representation of the token sequence. For the video sequence v, the set of granular tokens z_{med1} and z_{med2} in the two groups that obtain t frames of images can be expressed as follows:

$$z_{\mathrm{med}\ x} = \left\{ {\Phi^c}_{\mathrm{med}\ x,1}{}^{l}, {\Phi^c}_{\mathrm{med}\ x,2}{}^{l}, \cdots, {\Phi^c}_{\mathrm{med}\ x,t}{}^{l} \right\}, x = 1, 2 \tag{8}$$

For fine-grained tokens, we use a window of size $(H/4) \times (W \times p)$ to divide $z_n{}'$ into four parts. As with the operations for medium-grained tokens, four sets of fine-grained tokens (z_{fin1}, z_{fin2}, z_{fin3}, and z_{fin4}) are finally obtained after the l_{fin} layer operation in the backbone transformer, which can be expressed as:

$$z_{\mathrm{fin}\ x} = \left\{ {\Phi^c}_{\mathrm{fin}\ x,1}{}^{l}, {\Phi^c}_{\mathrm{fin}\ x,2}{}^{l}, \cdots, {\Phi^c}_{\mathrm{fin}\ x,t}{}^{l} \right\}, x = 1, 2, 3, 4 \tag{9}$$

For the n-th frame image in video sequence v, the extraction process of three different kinds of granularity tokens is shown in Figure 4.

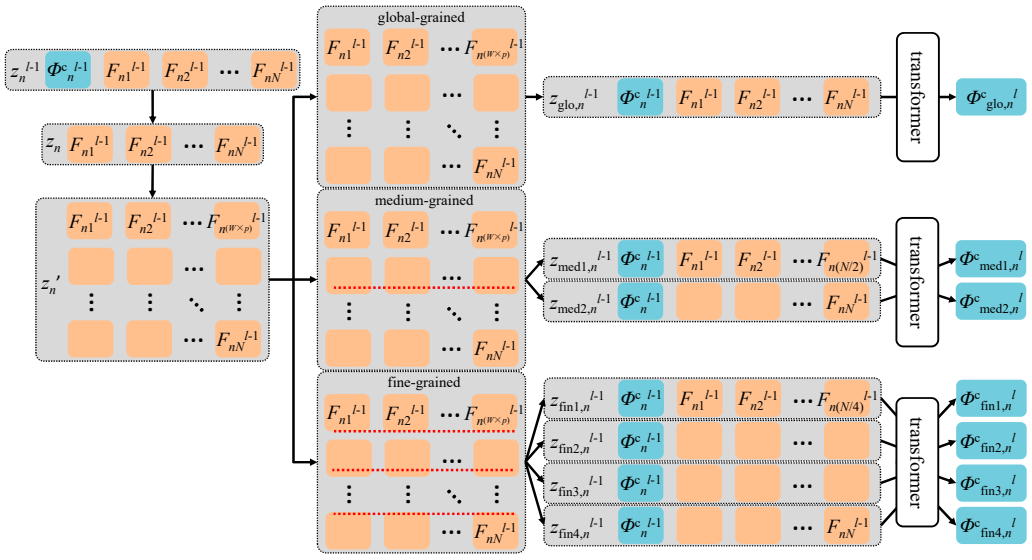

Figure 4. The extraction process of three different kinds of granularity tokens.

When compared with tokens with a global level of granularity, tokens with a finer level of granularity can extract and learn more detailed features locally in an image, meaning that they have better local feature awareness and can obtain multi-granular feature information. Thus, after learning a whole video sequence, such tokens can also distinguish people. In a video sequence, an image is affected by changes in illumination, shooting angle, occlusion, and other factors. Thus, each image in a frame provides different information on an overall feature. Nonetheless, when people observe a video sequence of person recognition, they can determine the integrity and importance of a feature in each image according to the image and information in each successive frame. This inspires us to design a way of aggregating granular tokens of the same category along a video sequence in terms of 'importance'. We measure importance by the cosine similarity between the same category of granular tokens and aggregate them as a weighted sum. Finally, we combine the three granularity tokens horizontally to form a set of multi-granular tokens Z, which represents the overall feature of video sequence v. The calculations involved are as follows:

$$\rho'_{glo,n} = \sum_{m=1,m\neq n}^{t} \cos(\Phi^{c}{}_{glo,n}^{\,l}, \Phi^{c}{}_{glo,m}^{\,l}) \quad \rho_{glo,n} = \frac{\exp(\rho'_{glo,n})}{\sum_{n=1}^{t} \exp(\rho'_{glo,n})} \qquad n, m \in t \tag{10}$$

$$\rho'_{\text{med } x,n} = \sum_{m=1,m\neq n}^{t} \cos(\Phi^{c}{}_{\text{med } x,n}^{\,l}, \Phi^{c}{}_{\text{med } x,m}^{\,l}) \quad \rho_{\text{med } x,n} = \frac{\exp(\rho'_{\text{med } x,n})}{\sum_{n=1}^{t} \exp(\rho'_{\text{med } x,n})} \qquad x = 1,2 \qquad n, m \in t \tag{11}$$

$$\rho'_{\text{fin } x,n} = \sum_{m=1,m\neq n}^{t} \cos(\Phi^{c}{}_{\text{fin } x,n}^{\,l}, \Phi^{c}{}_{\text{fin } x,m}^{\,l}) \quad \rho_{\text{fin } x,n} = \frac{\exp(\rho'_{\text{fin } x,n})}{\sum_{n=1}^{t} \exp(\rho'_{\text{fin } x,n})} \qquad x = 1,2,3,4 \qquad n, m \in t \tag{12}$$

$$Z_1 = \sum_{n=1}^{t} (\Phi^{c}{}_{glo,n}^{\,l} \times \rho_{glo,n}) \tag{13}$$

$$Z_{1+x} = \sum_{n=1}^{t} (\Phi^{c}{}_{\text{med } x,n}^{\,l} \times \rho_{\text{med } x,n}) \qquad x = 1,2 \tag{14}$$

$$Z_{3+x} = \sum_{n=1}^{t} (\Phi^{c}{}_{\text{fin } x,n}^{\,l} \times \rho_{\text{fin } x,n}) \qquad x = 1,2,3,4 \tag{15}$$

$$Z = [Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7] \tag{16}$$

In the training process, we use cross-entropy loss L_{cls} without label smoothing and ternary loss L_{tri} with soft edges to train the global and six local-grained tokens. The total loss L is calculated as follows:

$$L = L_{\text{cls}}(Z_1) + L_{\text{tri}}(Z_1) + \frac{1}{8} \times \sum_{x=2}^{7} (L_{\text{cls}}(Z_x) + L_{\text{tri}}(Z_x)) \tag{17}$$

During inference, we use the set of multi-granular tokens Z as the overall feature representation of the video sequence v.

4. Experiments

4.1. Datasets and Evaluation Protocols

To verify the effectiveness of the model, we conducted experiments on three widely used benchmark datasets: the Motion Analysis and Re-identification Set (MARS) [38], the Imagery Library for Intelligent Detection Systems-VID (iLIDS-VID) [39], and the Person Re-ID Dataset 2011 (PRID-2011) [40].

The MARS dataset is a video extension of the Market-1501 dataset and contains more than one million frames of people images, making it the largest publicly available video-based dataset. The data were collected by six cameras on the campus of Tsinghua University. The dataset contains 20,478 video sequences of 1261 identities: 625 identities for training and 636 identities for testing. The video sequence of each identity was captured by at least two cameras. The large amount of image information was processed by the deformable parts model detector and generalised maximum multi-clique problem tracker, and there are errors and poor execution in the figure cutting.

The iLIDS-VID dataset was collected in an aviation terminal hall, with the video sequence collected by two cameras with non-overlapping shooting fields. The dataset contains 300 identities and each of the two cameras captured a video sequence of 23–192 frames for each identity. Each video sequence has an average of 100 image frames. Although iLIDS-VID is a small dataset, with only 300 independent identities, the video sequence contains a large number of people wearing similar clothes and with overlapping occlusion, background clutter, and obvious lighting changes; these features make it a highly realistic and challenging dataset for Re-ID tasks.

The PRID-2011 dataset was collected and collated by the Austrian Institute of Technology University, Austria. The video sequences were collected at crosswalks and sidewalks by two surveillance cameras with non-overlapping shooting fields. Each video sequence contains 5–675 frames of images, with an average of 100 frames. The dataset contains 385 identity video images taken by camera A and 749 identity video images taken by camera B; only 200 identities were captured by both cameras. As the acquisition environment of PRID-2011 was a relatively spacious outdoor street scene, the background of its person images is simple and clean and there is less person occlusion than in the other two datasets.

To evaluate the re-identification performance of APMG network, we use the cumulative matching characteristic (CMC) curve and mean average precision (mAP) evaluation indicators. When using the MARS dataset, we followed the default training and testing groupings of the dataset and used the CMC curve and mAP to evaluate the model performance. When using iLIDS-VID and PRID-2011 datasets, we randomly divided the dataset into two subsets for training and testing groups. The experiment was repeated 10 times, and the CMC curve (representing the average result of 10 experiments and for which we took Rank-1 and Rank-5) was used to evaluate the model performance.

4.2. Implementation Details

Our network was written in Python 3.7. The module and training environment in the network architecture was PyTorch 1.8.0. The main hardware was an NVIDIA GeForce RTX 3090 with 24 GB of memory. The backbone network used ViT-B16 [5], which was pre-trained on ImageNet to improve the performance and accuracy of the network. In the preprocessing of the dataset, we divided the video into eight segments, randomly took one frame from each segment, and used these eight frames to form a new video sequence to represent the video. The image size in the video sequence was uniformly adjusted to 256 × 128 pixels, and random erasure, random cropping, and horizontal flipping were used for data enhancement. The batch size was set to 64 and contained eight video sequences. The overall structure of APMG network is shown in Figure 2. In the CAP module, the sliding window for preprocessing each frame of the video image was 4 × 4 (a = 4) and the number of generated patches $v_{an} \in R^{h \times w \times c}$ without overlapping pixels was 64 × 32 × 768. The sliding window for splitting patches was 4 × 4 (b = 4) and the quantity of feature embeddings partitioned in the window was p = 2. After the CAP module had completed its task, the number of feature embeddings generated for each frame image was N = 256. The number of network layers was l = 12. For network training, the learning rate was initially set to 0.09, and the cosine algorithm was used to adjust the learning rate. The optimiser used stochastic gradient descent.

4.3. Comparison with State-of-the-Art Methods

To verify the performance, we selected 14 recently developed models to apply to the MARS, iLIDS-VID, and PRID-2011 datasets for the experimental comparison. The selected models are the attribute-driven feature disentangling (ADFD) [28] and appearance and motion enhancement (AMEM) [29], which uses person-attribute information; the multi-granular hypergraphs (MGH) [16], STGCN [18], and correlation and topology learning (CTL) [17], which use a GCN; the appearance-preserving 3D convolution (AP3D) [25], spatiotemporal representation factorisation (STRF) [26], and self-separated network to align parts for 3D convolution (SSN3D) [27] models, which use a three-dimensional convolution module; dense interaction learning (DenseIL) [21], spatiotemporal transformer (STT) [36], and SINet [23] models, which use an attention mechanism to enhance the extraction of time- or person-feature information; and the multi-granularity reference-aided attentive feature aggregation (MG-RAFA) [30], global-guided reciprocal learning (GRL) [31], and pyramid in transformer (PiT) [37] models, which consider feature information at different levels of granularity. The re-identification accuracy of each model is shown in Table 1. Our APMG network achieved the best performance in each of the three benchmark datasets. In the MARS dataset, its mAP scores are more than 1.4% higher than those of the next best model, and on the iLIDS-VID and PRID-2011 datasets, its Rank-5 scores are close to 100%.

Table 1. The re-identification accuracy of each model.

Methods		MARS			iLIDS-VID		PRID-2011	
		mAP	**Rank-1**	**Rank-5**	**Rank-1**	**Rank-5**	**Rank-1**	**Rank-5**
Attribute	ADFD	78.2	87.0	95.4	86.3	97.4	93.9	99.5
	AMEM	79.3	86.7	94.0	87.2	97.7	93.3	98.7
Graph	MGH	85.8	90.0	96.7	85.6	97.1	94.8	99.3
	STGCN	83.7	89.9	96.4	-	-	-	-
	CTL	86.7	91.4	96.8	89.2	97.0	-	-
3D convolution	AP3D	85.6	90.7	-	88.7	-	-	-
	STRF	86.1	90.3	-	89.3	-	-	-
	SSN3D	86.2	90.1	96.6	88.9	97.3	-	-
Attention	DenseIL	87.0	90.8	97.1	92.0	98.0	-	-
	STT	86.3	88.7	-	-	-	-	-
	SINet	86.2	91.0	-	92.5	-	96.5	-
Grained	MG-RAFA	85.9	88.8	97.0	88.6	98.0	95.9	99.7
	GRL	84.8	91.0	96.7	90.4	98.3	96.2	99.7
	PiT	86.8	90.2	97.2	92.1	98.9	-	-
APMG networks		88.4	92.1	97.8	93.6	98.7	97.1	99.8

In Table 1, STT, PiT, and our APMG network are all based on a transformer structure. When compared with other CNN-based networks, the experimental results show that these three networks had the best performance, which proves the feasibility and effectiveness of a transformer structure for Re-ID tasks. When compared with the PiT network, our method is similar, as it attaches importance to the granularity of feature information; however, our method adds the aggregation of feature information according to 'importance', which enhances the feature representation at each level of granularity for a person video sequence. Moreover, we use the CAP module to generate feature embeddings instead of a fixed sliding window in the Pit network, which effectively improves the overall structure and semantic information consistency of feature embeddings and makes the network capture more detailed with comprehensive information between feature embeddings. The overall performance of the APMG network is therefore superior to that of PiT. Networks based on a GCN structure can effectively capture the dependencies of people in space and time, but these models adopt a CNN-based structure for feature extraction. Moreover, as the

points in a GCN structure are represented by coarsely grained features, a GCN network does not make full use of the detailed distinguishable feature information in an image. In contrast, our APMG network makes use of various features with different levels of granularity and also pays attention to the importance of features in time. Furthermore, the APMG network's transformer structure and CAP module allows the network to learn the global features of an image without losing the original information. For these reasons, the architecture and performance of the APMG network are superior to that of GCN-based networks. When compared with attention-based methods, the transformer in the APMG network has the structural characteristics of the global receptive field, which can extract enough feature information from each frame. It is not necessary to expand the attention area in the current image through the previous frames as SINet does. When compared with attribute-based methods, the APMG network's mAP score in the MARS dataset increased by 9.1%. The available attribute information in person images mainly depends on the extraction of the network, it may have a large error due to the change in visual angle. However, the APMG network mainly extracts the multi-granularity information of people, and through the multi-granularity information fusion between frames, it can effectively reduce the influence of the visual angle on the network performance. In summary, the APMG network has achieved the most advanced performance.

To verify the effectiveness of the APMG network, we randomly select three video sequences from the MARS dataset and examine the activation visualisation results for these sequences, as shown in Figure 5. The first group of images in the figure shows that the APMG network effectively extracts distinguishable feature information of people, such as a ponytail, a schoolbag, an arm, and other feature information. Figure 5 also shows that the APMG network can extract person-feature information from an image with a cluttered background (such as in the third image in the first group) or when the person is blocked (such as in the second image in the second group).

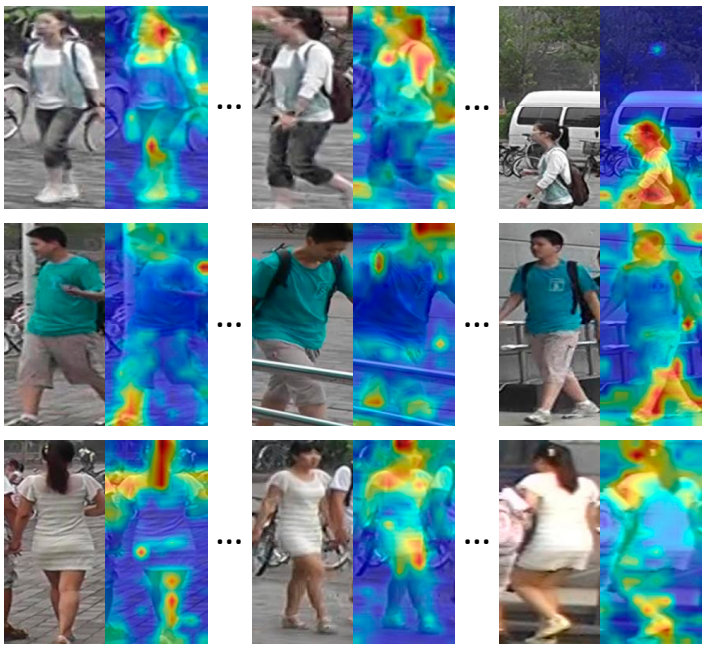

Figure 5. Activation visualisation results from three video sequences.

4.4. Ablation Study

We performed ablation experiments on the CAP and MG modules in the APMG network, again using the MARS, iLIDS-VID, and PRID-2011 datasets.

For the ablation experiment on the CAP module, we set up five experimental settings. In setting 1, we cancelled the two-layer sliding window structure and instead used a 16×16 sliding window. In settings 2 and 3, we used two different sizes of two-layer sliding

window structures. In setting 4, we removed the correlation-adaptive partitioning of the feature embedding. In setting 5, we used all patches on the sequence corresponding to the position of the sliding window to calculate the correlation. To ensure the accuracy of the experiments, we kept the number of feature embeddings consistent across the five settings. The results are shown in Table 2.

Table 2. Ablation experiment results for CAP module.

Settings	Parameter Quantity	MARS	iLIDS-VID	PRID-2011
		mAP	Rank-1	Rank-1
4 × 4 and 4 × 4 CAP	126 M	88.4	93.6	97.1
1. 16 × 16	125 M	84.1	90.7	95.4
2. 2 × 2 and 8 × 8	148 M	88.2	94.0	97.0
3. 8 × 8 and 2 × 2	131 M	86.7	92.4	96.3
4. 4 × 4 and 4 × 2 no partition	123 M	86.2	89.1	93.2
5. use all patches to calculate the correlation	154 M	88.5	91.9	97.5

The re-identification accuracy results for settings 1, 2, and 3 show that our two-layer sliding window structure is effective. Under setting 2, the re-identification accuracy in iLIDS-VID was improved by 0.4% because the first layer adopted a 2 × 2 sliding window, which made the generated patches more detailed. However, the 8 × 8 sliding window greatly increased the number of network parameters needed for partitioning the feature embedding. Considering the overall performance and the number of network parameters in the three datasets, the scheme in setting 2 was discarded. Under setting 4, the re-identification accuracy decreased by as much as 4.5% in iLIDS-VID, which demonstrates the benefits of our correlation-adaptive partitioning of feature embeddings. Under setting 5, the re-identification accuracy was slightly improved for the MARS and PRID-2011 datasets because taking all patches into account when calculating the correlation increases the detail of the reference used for defining feature embeddings. However, this increased the number of network parameters and thus increased the computation effort, which introduced a large amount of information redundancy into the time series and thus decreased the robustness of the network. In iLIDS-VID, which is characterised by more complex background information, the re-identification accuracy decreased by 1.7%. In terms of the overall performance of the network on the three datasets and the number of network parameters, our CAP module was the best performer. We randomly selected two images in the MARS dataset for the visualisation of activations, as shown in Figure 6. The figure shows that the CAP module effectively partitioned the feature embeddings and improved the network's extraction of the distinguishable feature information.

Figure 6. Activation visualisation results. From left to right: original image, setting 4, and CAP.

For the ablation experiment on the MG module, we set up four groups of experiments with varying partition granularities in window size. Setting 1 used a single size 16 × 16 window (global-granularity token). Setting 2 was setting 1 plus two size 8 × 16 windows (medium-grained tokens). Setting 3 was setting 2 plus four size 4 × 16 windows (fine-grained tokens) and eight size 2 × 16 windows. Setting 4 lacked the weighted sum

aggregation operation designed in the MG module and used a simple mean method instead. The experimental results are shown in Table 3.

Table 3. The ablation experiment results of MG module.

Settings	Parameter Quantity	MARS	iLIDS-VID	PRID-2011
		mAP	Rank-1	Rank-1
16 × 16, 8 × 16, 4 × 16 MG	126 M	88.4	93.6	97.1
1. 16 × 16	114 M	86.6	90.0	94.9
2. 16 × 16, 8 × 16	119 M	87.5	92.7	95.8
3. 16 × 16, 8 × 16, 4 × 16, 2 × 16	135 M	88.2	92.9	97.2
4. MG without weighted sum	121 M	87.0	90.8	95.3

The results in Table 3 show that as the granularity was continuously refined, the number of network parameters increased rapidly, and the re-identification performance gradually increased. A bottleneck was reached in setting 3, where the re-identification performance only slightly improved in the PRID-2011 dataset. When the granularity was too fine, the extracted feature information could lose its overall structure and the granular feature information could become trivial, which increased the redundant information in the image. Considering the re-identification performance improvement and the number of network parameters, we abandoned the size 2 window. When compared with Setting 4, our aggregation operation achieved an improvement of more than 1.4% on all three datasets. This demonstrated that the MG module enhances the final feature representation of video sequences.

5. Conclusions and Prospect

In this study, we propose a network framework named APMG for video Re-ID tasks. The APMG network comprises two modules: a CAP module, which partitions feature embeddings, and an MG module, which extracts feature information at different levels of granularity. To verify the re-identification performance of the APMG network, we conducted experiments on the MARS, iLIDS-VID, and PRID-2011 datasets. The results show that the performance of the APMG network is superior to other state-of-the-art methods, which also proves the feasibility of applying transformer structures in the Re-ID task. The application of transformer structures in Re-ID is at the initial stage. We believe that a more effective transformer-based network structure can be designed and improved by absorbing a lot of mature experience based on CNN methods, which can be further explored. In future work, we will introduce sliding windows with variable sizes into the ACP module to generate feature embeddings, and to make the network adapt to image tasks more effectively.

Author Contributions: Conceptualization, B.H.; methodology, B.H.; software, B.H.; validation, B.H., Y.P. and Y.T.; formal analysis, B.H.; investigation, B.H.; resources, B.H. and Y.P.; data curation, B.H.; writing—original draft preparation, B.H.; writing—review and editing, B.H. and Y.P.; visualization, B.H.; supervision, B.H. and Y.T.; project administration, Y.P.; funding acquisition, Y.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Jilin Provincial Department of Science and Technology grant number YDZJ202102CXJD062, and Jilin Province Machine Vision Intelligent Manufacturing and Inspection Technology Innovation Centre grant number 20180623039TC.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://blog.csdn.net/qq_34132310/article/details/83869605 (accessed on 9 November 2022).

Acknowledgments: This work was supported by Changchun University of Science and Technology, Jilin Provincial Department of Science and Technology, and Jilin Province Machine Vision Intelligent Manufacturing and Inspection Technology Innovation Centre.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 10012–10022.
2. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jegou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MA, USA, 17–23 July 2021; pp. 10347–10357.
3. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
4. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 213–229.
5. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *2017*, 1–11.
7. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. Transreid: Transformer-based object re-identification. In Proceedings of the IEEE/CVF international conference on computer vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15013–15022.
8. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.; Tay, F.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 558–567.
9. Wang, W.; Xie, E.; Li, X.; Fan, D.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
10. Chen, Z.; Zhu, Y.; Zhao, C.; Hu, G.; Zeng, W.; Wang, J.; Tang, M. Dpt: Deformable patch-based transformer for visual recognition. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–25 October 2021; pp. 2899–2907.
11. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-scale feature learning for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3702–3712.
12. Chen, X.; Liu, X.; Liu, W.; Zhang, X.P.; Zhang, Y.; Mei, T. Explainable Person Re-Identification with Attribute-guided Metric Distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 11813–11822.
13. Nguyen, B.X.; Nguyen, B.D.; Do, T.; Tjiputra, E.; Tran, Q.D.; Nguyen, A. Graph-based person signature for person re-identifications. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3492–3501.
14. Shen, Y.; Li, H.; Yi, S.; Chen, D.; Wang, X. Person re-identification with deep similarity-guided graph neural network. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 486–504.
15. Yan, Y.; Zhang, Q.; Ni, B.; Zhang, W.; Xu, M.; Yang, X. Learning context graph for person search. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Nashville, TN, USA, 20–25 June 2019; pp. 2158–2167.
16. Yan, Y.; Qin, J.; Chen, J.; Liu, L.; Zhu, F.; Tai, Y.; Shao, L. Learning multi-granular hypergraphs for video-based person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Nashville, TN, USA, 20–25 June 2020; pp. 2899–2908.
17. Liu, J.; Zha, Z.J.; Wu, W.; Zheng, K.; Sun, Q. Spatial-temporal correlation and topology learning for person re-identification in videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4370–4379.
18. Yang, J.; Zheng, W.S.; Yang, Q.; Chen, Y.C.; Tian, Q. Spatial-temporal graph convolutional network for video-based person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2020; pp. 3289–3299.
19. Subramaniam, A.; Nambiar, A.; Mittal, A. Co-segmentation inspired attention networks for video-based person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 562–572.
20. Wang, Y.; Zhang, P.; Gao, S.; Geng, X.; Lu, H.; Wang, D. Pyramid spatial-temporal aggregation for video-based person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 12026–12035.

21. He, T.; Jin, X.; Shen, X.; Huang, J.; Chen, Z.; Hua, X.S. Dense interaction learning for video-based person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 1490–1501.
22. Eom, C.; Lee, G.; Lee, J.; Ham, B. Video-based person re-identification with spatial and temporal memory networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12036–12045.
23. Bai, S.; Ma, B.; Chang, H.; Huang, R.; Chen, X. Salient-to-Broad Transition for Video Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2022; pp. 7339–7348.
24. Li, J.; Zhang, S.; Huang, T. Multi-scale 3d convolution network for video based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8618–8625.
25. Gu, X.; Chang, H.; Ma, B.; Zhang, H.; Chen, X. Appearance-preserving 3d convolution for video-based person re-identification. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 228–243.
26. Aich, A.; Zheng, M.; Karanam, S.; Chen, T.; Roy-Chowdhury, A.K.; Wu, Z. Spatio-temporal representation factorization for video-based person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 152–162.
27. Jiang, X.; Qiao, Y.; Yan, J.; Li, Q.; Zheng, W.; Chen, D. SSN3D: Self-separated network to align parts for 3D convolution in video person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2021; Volume 35, pp. 1691–1699.
28. Zhao, Y.; Shen, X.; Jin, Z.; Lu, H.; Hua, X.S. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2019; pp. 4913–4922.
29. Li, S.; Yu, H.; Hu, H. Appearance and motion enhancement for video-based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11394–11401.
30. Zhang, Z.; Lan, C.; Zeng, W.; Chen, Z. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2020; pp. 10407–10416.
31. Liu, X.; Zhang, P.; Yu, C.; Lu, H.; Yang, X. Watching you: Global-guided reciprocal learning for video-based person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13334–13343.
32. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on visual transformer. *arXiv* **2020**, arXiv:2012.12556.
33. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM Comput. Surv.* **2022**, *54*, 1–41. [CrossRef]
34. Chu, X.; Zhang, B.; Tian, Z.; Wei, X.; Xia, H. Do we really need explicit position encodings for vision transformers. *arXiv* **2021**, arXiv:2102.10882.
35. Srinivas, A.; Lin, T.Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck transformers for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16519–16529.
36. Zhang, T.; Wei, L.; Xie, L.; Zhuang, Z.; Zhang, Y.; Li, B.; Tian, Q. Spatiotemporal transformer for video-based person re-identification. *arXiv* **2021**, arXiv:2103.16469.
37. Zang, X.; Li, G.; Gao, W. Multi-direction and Multi-scale Pyramid in Transformer for Video-based Pedestrian Retrieval. *IEEE Trans. Ind. Inform.* **2022**, *18*, 8776–8785. [CrossRef]
38. Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; Tian, Q. Mars: A video benchmark for large-scale person re-identification. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 868–884.
39. Wang, T.; Gong, S.; Zhu, X.; Wang, S. Person re-identification by video ranking. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 688–703.
40. Hirzer, M.; Beleznai, C.; Roth, P.M.; Bischof, H. Person re-identification by descriptive and discriminative classification. In Proceedings of the Scandinavian Conference on Image Analysis, Ystad, Sweden, 9 May 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 91–102.