applied sciences

MDPI

Article

PVTReID: A Quick Person Reidentification-Based Pyramid Vision Transformer

Ke Han, Qianlong Wang *, Mingming Zhu and Xiyan Zhang

School of Software, North China University of Water Resources and Electric Power, Zhengzhou 450046, China; hanke@ncwu.edu.cn (K.H.); 13525723704@163.com (M.Z.); 19138568973@163.com (X.Z.)
* Correspondence: wql19971225@gmail.com

Abstract: Person re-identification (ReID) has attracted the attention of a large number of researchers due to its wide range of applications. However, due to the difficulty of extracting robust features and the complexity of the feature extraction process, ReID is difficult to truly apply in practice. In this paper, we utilize Pyramid Vision Transformer (PVT) as the backbone for feature extraction and propose a PVT-based ReID method in conjunction with other studies. First, we establish a basic model using powerful methods verified on CNN-based ReID. Second, to further improve the robustness of the features extracted from the PVT backbone, we design two new modules: (1) a local feature clustering (LFC) module is used to select the most discrete local features and cluster them individually by calculating the distance between local and global features, and (2) side information embeddings (SIE) are used to encode nonvisual information and send it to the network for use training in order to reduce its impact on the features. Our experiments show that the proposed PVTReID achieves an mAP of 63.2% on MSMT17 and 80.5% on DukeMTMC-reID. In addition, we evaluated the inference speed for images achieved by different methods, proving that image inference is faster with our proposed method. These results clearly illustrate that using PVT as a backbone network with LFC and SIE modules can improve inference speed while extracting robust features.

Keywords: ReID; Pyramid Vision Transformer; local feature clustering; side information embeddings

Citation: Han, K.; Wang, Q.; Zhu, M.; Zhang, X. PVTReID: A Quick Person Reidentification-Based Pyramid Vision Transformer. *Appl. Sci.* **2023**, *13*, 9751. https://doi.org/10.3390/app13179751

Academic Editor: Andrea Prati

Received: 31 July 2023
Revised: 23 August 2023
Accepted: 26 August 2023
Published: 29 August 2023

1. Introduction

In recent years, more and more attention is being paid to the public safety; in this context, it is very important to improve person retrieval ability. Person re-identification (ReID) is a technology for searching for specific people from images captured by different cameras. It can complement face recognition technology and meet the needs of current intelligent monitoring. In addition to being used for tracking the whereabouts of criminal suspects on monitoring networks, ReID can be used for tracking in smart devices. Due to the influence of visual factors such as lighting, posture, occlusion, and resolution [1] and non-visual factors such as camera angle [2] and perspective, ReID faces many challenges in practical applications.

Extracting robust and discriminative person features is an important research part of ReID, which has long been dominated by CNN-based ReID [3–6]. In practical application, the effects of objective causes such as background interference, person blockage, and posture misalignment make it difficult for global feature to meet the requirements of person retrieval. The common practice of CNN-based ReID is to combine local feature extraction to obtain fine-grained features to make up for the shortcomings of global features. With the development of vision attention, the application of attention to ReID [7–9] has become popular. With the vast success of ViT [10] in the field of computer vision, a number of scholars have started to explore the application of ViT to the ReID task [11] by extracting global features through a Transformer.

In reviewing CNN-based methods and ViT-based ReID, we discovered three significant matters which are not well solved in ReID. First, the inference speed is slow. Most

CNN-based ReID methods combine local features to enhance its fine-grained ability to identify people; however, these additional structures greatly increase model complexity and consumption of computational resources [1]. Due to the large amount of computing resources required by Transformer, ViT-based ReID has slow feature extraction speed. Second, as ViT is designed for image classification tasks, the network needs to use a class token for classification to reduce preferences for specific image patches. Because it learns the information of image patches, it is overly reliant on global information, resulting in reduced usage of local features and decreased fine-grained identification ability [12]. Third, ViT uses position encoding to process spatial information. Position encoding corresponds to the input image size and resolution. Thus, position embedding requires retraining when dealing with images of different resolutions.

In order to solve the above problems, we present an ReID method using a Pyramid Vision Transformer (PVT)-based method. PVT [13] has achieved high accuracy on image classification tasks as well as on other downstream tasks. PVT uses a pyramidal feature structure [14] and zero padding [15] to solve the above-mentioned problems. First, unlike ViT, PVT uses a pyramid structure, which greatly reduces resource consumption and boosts the efficiency of feature extraction. Second, instead of using additional tokens as output vectors for feature extraction, PVT aggregates all local features trained by the Transformer to obtain a global feature representation. Through this operation, PVT increases the dependence of output features on local features while reducing model complexity and computational consumption and improving model training and inference efficiency. Third, PVT does not use absolute position encoding to represent the position information of image patches. Instead, it introduces zero-padding position encoding to learn position information and uses depth-wise convolution [16] to model position information.

In summary, PVT has great advantages; however, it needs to be modified in order to adapt to the unique challenges present in ReID, such as occlusion, pose changes, and camera changes. CNN-based methods alleviate the impact of these factors on person features in various ways, among which local feature methods [17] and side information [2] have been demonstrated to be effectual means of strengthening the robustness of features. At the same time, the features extracted by CNN-based methods are different from those extracted by the PVT-based method, and their semantic information differs. Furthermore, considering side information such as cameras can reduce the impact of nonvisual factors on the robustness of person features. If complex information constructed via CNN is directly used for PVT, the encoding ability and pyramidal structure of PVT cannot be fully utilized. To successfully solve these problems, it is necessary to design modules specifically for PVT.

For these reasons, we present a novel ReID method called PVTReID for obtaining robust person features. First, we build a strong basic network based on PVT with a few crucial adaptations. Second, in view of solving the problem of local person features that are indistinguishable from each other, we propose a local feature clustering module (LFC) by calculating the most discrete local features for further feature learning. LFC is used in the final stage of the framework to obtain person features together with the global features. Third, to further improve the robustness of the person features, we employ side information embeddings (SIE) to embed side information into the four encoding stages of the PVT.

The contributions of this paper are summarized as follows:

- We recommend a strong basic network that utilizes PVT for ReID with performance that is comparable to CNN-based methods. With the help of PVT, by using progressively smaller feature pyramids the PVT-based method reduces the computation of large feature maps, thereby alleviating the problem of slow inference for ReID person images.
- We introduce a local feature clustering (LFC) module to compute the most discrete of the local features and cluster them. With LFC, we further separate the feature representations of different people in the feature space to make the person features more robust.

- We perform side information embeddings (SIE) on the camera information and send this information to the feature extraction network for use in training, which reduces interference from nonvisual information in the resulting features and improves their robustness. In addition, we verify the effects of using camera information in training at different PVT encoding stages.
- The final PVTReID framework achieves 87.8%, 63.2%, and 80.5% mAP accuracy on Market-1501, MSMT17, and DukeMTMC-reID, respectively, and has faster inference speed compared to CNN-based methods.

2. Related Work

2.1. Person Reidentification

Many CNN-based ReID methods have been proposed in recent years, and have proven to have good performance. One popular pipeline is to build on top of a CNN backbone network (e.g., ResNet [18]) and optimize the network by designing a suitable loss function to extract person features.

Representation learning using global features is a very common ReID approach [19,20]. The representation learning method mainly regards ReID as an image classification task, regards each person ID as a category, and uses the global features extracted by the backbone to calculate the ID Loss [20]. Metric learning is a widely used method for image retrieval. Metric learning considers ReID as an image clustering problem, and aims to find the distance between two images in the feature space by learning. In the ReID feature space, metric learning shows that the distance between different images with the same person ID is less than the distance between different images with different person IDs. Triplet loss is a widely used metric learning loss, and many metric learning methods are based on the research and improvement of triplet loss [21,22]. The common idea of training the network by integrating metric learning and representation learning in person re-identification models has become popular. Luo et al. [23] devised with BNNeck, and Sun et al. [24] submitted Circle Loss, both of which have provided good presentations of the use of ID loss and triplet loss.

Representation learning and metric learning use the global features of people. In the case of misaligned posture, occlusion of images, and cases in which only local details are dissimilar, global features are prone to making mistakes. Local feature-based methods can solve these problems to an extent by mining fine-grained information. GLAD [25] extracts local features by dividing people into three parts: head, upper body, and lower body. PCB [5] segments the extracted person features through average pooling, then uses 1×1 convolution to obtain independent local features and predicts classification based on these local features. Local feature-based methods usually add a branch on the basis of global features; while this can extract rich person features, it increases the model inference time.

2.2. Vision Transformer

Ashish Vaswani et al. [26] first proposed the Transformer model to process sequence data in natural language processing (NLP). Inspired by this discovery, numerous researchers have explored its application in computer vision. Han [27] and Salman [28] investigated the application of Transformers in computer vision and showed its effectiveness in different tasks. ViT [10] divides an image into patches and flattens these patches into a sequence of one-dimensional vectors used as input to the Transformer. The uniqueness of ViT is that a learnable embedding is added to extract global features, reducing the preference for a certain image patch when classifying images; however, this reduces the model's ability to extract local features. PVT [13] uses a self-attention variant called Spatial-Reduced Attention (SRA). Based on this, PVT_V2 [29] obtains more continuous local image patches using overlapping patch embedding, eliminates the need for fixed position encoding through convolutional feed-forward networks, and compensates for the removal of position encoding through 3×3 convolution, making it flexible enough to handle inputs of various resolutions.

2.3. Side Information

In the process of ReID, it is very common to encounter changes in posture, and consequently in the resulting resolution of images due to different camera angles. In order to resolve these problems, previous works have used side information such as camera information and viewpoint information to enhance the robustness of the learned features. CBN [2] transforms person images from different cameras into the same subspace, effectively improving the distribution difference of images taken by different cameras. TransReID [11] sends side information encoding to the ViT for training, thereby reducing the influence of camera factors on the resulting features and improving their robustness.

3. Methods

As shown in Figure 1, ourproposed ReID framework for obtaining robust person features is designed on the basis of a PVT-based image classification network along with several useful improvements [23]. To further enhance the robustness of person feature training in the face of the many challenges involved, in Sections 3.2 and 3.3 we present the comprehensive design of a local feature clustering module (LFC) and side information embeddings (SIE) module. The resulting end-to-end network uses both modules at the same time, as displayed in Figure 2.

3.1. Basic PVT-Based Network

PVT [13,29] is a state-of-the-art image feature extraction method that is widely used for a variety of computer vision tasks. As an asymptotic feature pyramid method, image features are well learned in the performance of PVT while reducing the computation of large image feature maps. We use PVT as a backbone network to construct a powerful feature extraction network by adding representation learning and metric learning branches to enhance feature extraction capability while improving feature extraction speed.

We establish a PVT-based basic network for ReID using the general strong methods [23]. Our PVT-based ReID method consists of two main parts: feature extraction and feature utilization. PVT-V2 is chosen as the backbone network for feature extraction. As shown in Figure 1, it is mainly segmented into four parts with the same structure; the extraction process of each part can be expressed as follows:

$$Input : I_1 = x, I_i = O_{i-1} \tag{1}$$

$$F_i = PatchEmb_i(I_i) \tag{2}$$

$$O_i = Encoder(F_i) \tag{3}$$

For a given image $x \in \mathbb{R}^{H \times W \times C}$ used as input, where H, W, C respectively represent its height, width, and number of channels, respectively, we split the image into N patches with fixed size using PatchEmbed. Here, I_i represents the input vector of each part of the PVT. The image x is the first part of the input vector of the PVT backbone network, while the input vectors of the other parts are the output vectors O_{i-1} computed from the previous part. The input vector I_i is obtained by $PatchEmbed_i$ to obtain the vector F_i, and F_i is calculated by $Encoder_i$ to obtain the output vector O_i of each part. Repeating the above operations four times constitutes the PVT-based ReID backbone.

In Stage 1, a given input image $x \in \mathbb{R}^{H \times W \times 3}$ is first divided into $\frac{H \times W}{4^2}$ image patches of size $4 \times 4 \times 3$, which are input to PatchEmbed and then expanded into a feature sequence $F_1 = \frac{H \times W}{4^2} \times C_1$. The feature sequence is then input to the corresponding Encoder layer, and the final output is a feature map $O_1 = \frac{H}{4} \times \frac{W}{4} \times C_1$. The subsequent stage takes the feature output of the previous stage as input for the same operation and outputs the respectively feature maps F_2, F_3, and F_4, with a reduction scale of 8, 16, and 32 times with respect to the input image.

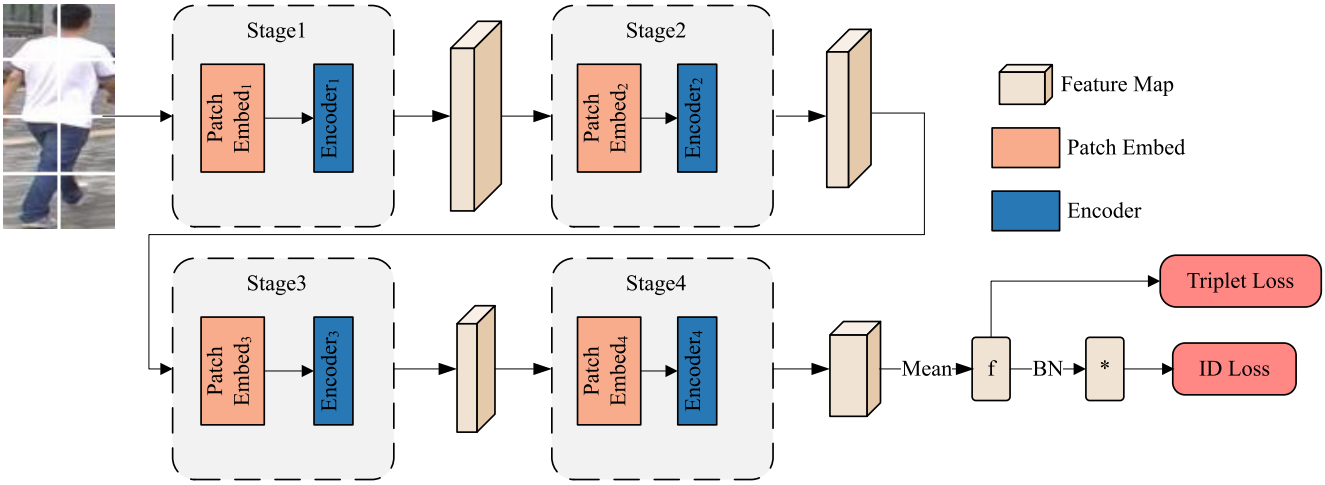


Figure 1. PVT-based basic framework. The whole model can be divided into two parts, namely, feature extraction and feature utilization. The feature extraction stage can be divided into four stages; each stage has a PatchEmbed and an Encoder composed of multiple Transformer layers. According to the pyramidal structure, the resolution of the output feature maps of the four stages decreases from the front (1/4 the original image size) to the back (1/32 the original image size). The final output feature maps are subjected to representation learning and metric learning, respectively. Inspired by [23], we introduce the BNNeck before the ID Loss. * is applied to the feature using Batch Normalization.

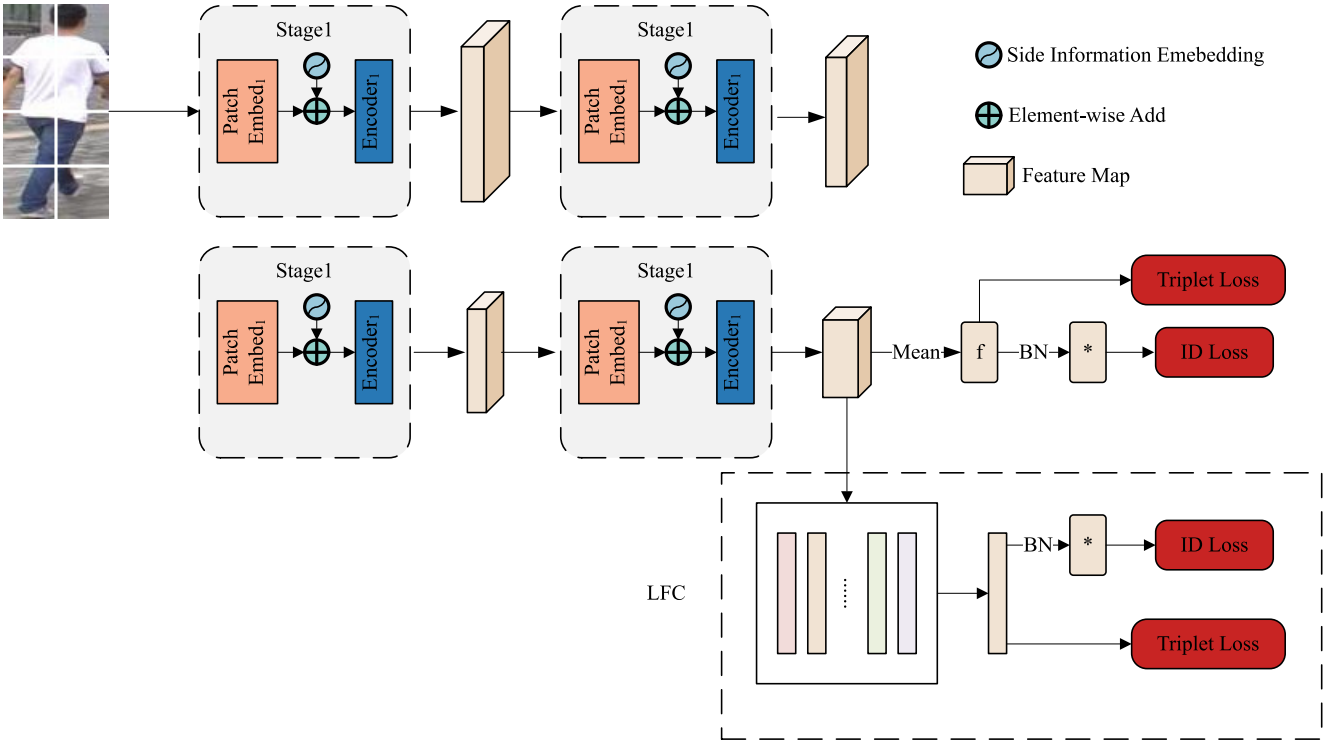


Figure 2. PVTReID Framework. Compared to the PVT-based basic framework, PVTReID adds an SIE module in the encoding phase and a local feature branch in the supervised learning phase. Nonvisual information such as camera information is encoded in the SIE. The SIE is added to the image patch embeddings element-by-element and fed into the Encoder. Supervised learning consists of two separate branches; one branch is a common branch that acquires global features, while the other uses the LFC module to compute the distance between local features and global features to obtain the most discrete local features. ReID loss is contributed by both global feature and local features.

Compared with ViT, which takes up a large amount of computational and storage resources, the PVT network adopts a pyramidal structure; as the network deepens, the number of channels in the feature maps gradually increases and the size of the feature maps gradually decreases. PVT obtains high-resolution feature maps in dense prediction tasks while reducing the computational resource consumption of feature maps with large sizes. In contrast to the ViT network, the PVT network does not add an extra learnable embedding token, instead aggregating the vector O_4 to obtain the global feature.

3.1.1. Patch Embed

Early Transformer models such as ViT divide images into non-overlapping image patches, which destroys the structural features of the local region. PVT uses sliding windows to generate overlapping pixel image patches, which is better able to preserve the integrity of the local features of images. PVT uses a pyramidal structure to divide the entire feature extraction process into four parts; each of these requires the encoding of a feature map. Assuming that the size of the image patch is $P \times P$, the step size is S, and S is less than P, the area where the two image patches overlap is $(P - S) \times P$. If an input feature map which has a resolution of $H \times W$, it is divided into N patches using a suitable P and S:

$$N = N_H \times N_W = \left\lfloor \frac{H + S - P}{S} \right\rfloor \times \left\lfloor \frac{W + S - P}{S} \right\rfloor \tag{4}$$

where N_H and N_W represent the numbers of splitting patches in height and width, respectively, $\lfloor . \rfloor$ is the floor function, and S is set to be smaller than P. When S is smaller, more patches can be obtained by dividing the input feature map, although more patches requires more computing power.

3.1.2. Feature Usage

We optimize the PVT-ReID basic network using the classification loss and triplet loss for the global feature. ReID can be considered as a classification task. Person IDs are regarded as categories of persons, and are used as labels to train the network. An amount of classification categories is equal to the total count of person IDs in the training set. We employ the cross-entropy loss along with label smoothing for the ID loss L_{ID} [30]:

$$L_{ID} = -\sum_{i}^{N} q_i \ln p_i \tag{5}$$

$$q_i = \begin{cases} 1 - \frac{N-1}{N}\varepsilon, if i = y \\ \frac{\varepsilon}{N}, otherwise \end{cases} \tag{6}$$

where the sum of the person IDs in the training set is N, p_i is the predicted probability of person ID, and y represents the true ID of the predicted person. Equation (6) represents the label smoothing operation on the ID, where ε is a hyperparameter. In this article, ε is set as 0.1. For a triplet set of person images $\{I_a, I_p, I_n\}$, we use the hard sample mining triplet loss L_T with a soft margin [21], as follows:

$$L_T = \log\left\{1 + exp\left[\max_{f_p}(\|f_a - f_p\|_2^2) - \min_{f_n}(\|f_a - f_n\|_2^2)\right]\right\} \tag{7}$$

where f_a, f_p, and f_n denote the feature representations of the person image I_a, the positive pair image I_p, and the negative pair image I_n, respectively. We choose the most distant positive pair and the closest negative pair in the mini-batch to calculate the triplet loss.

3.2. Local Feature Clustering Module

We add a local feature clustering (LFC) branch on the base of the global features. We derive the most discrete local features by calculating the distance between different

local features and global features, then optimize them using the cross-entropy loss and triplet loss to separate different local features and improve the features' fine-grained discriminative ability.

The framework based on PVT has achieved excellent results in ReID due to its powerful global feature extraction ability. However, with the problem of occlusion and posture misalignment, using global feature alone as the standard for distance measurement cannot meet the discrimination needs of difficult samples. Therefore, it is necessary to learn the local features of people in order to improve fine-grained discrimination ability. Stripe features and posture estimation have been widely used in CNN-based methods to extract the fine-grained features of people.

The feature of an image that PVT-based ReID extracts is $O_4 = [f_1, f_2, \ldots, f_{32}]$, and the global feature f is obtained through the mean operation. In order to obtain fine-grained local features, a straightforward approach is to use the discriminative ability of each local feature, that is, to cluster all local features which have the same person ID in the feature space. While the global feature f obtained after aggregating all local features is clustered as well, it cannot take into account all local features, and there will be situations in which local features cannot be distinguished, as shown in Figure 3a,c.

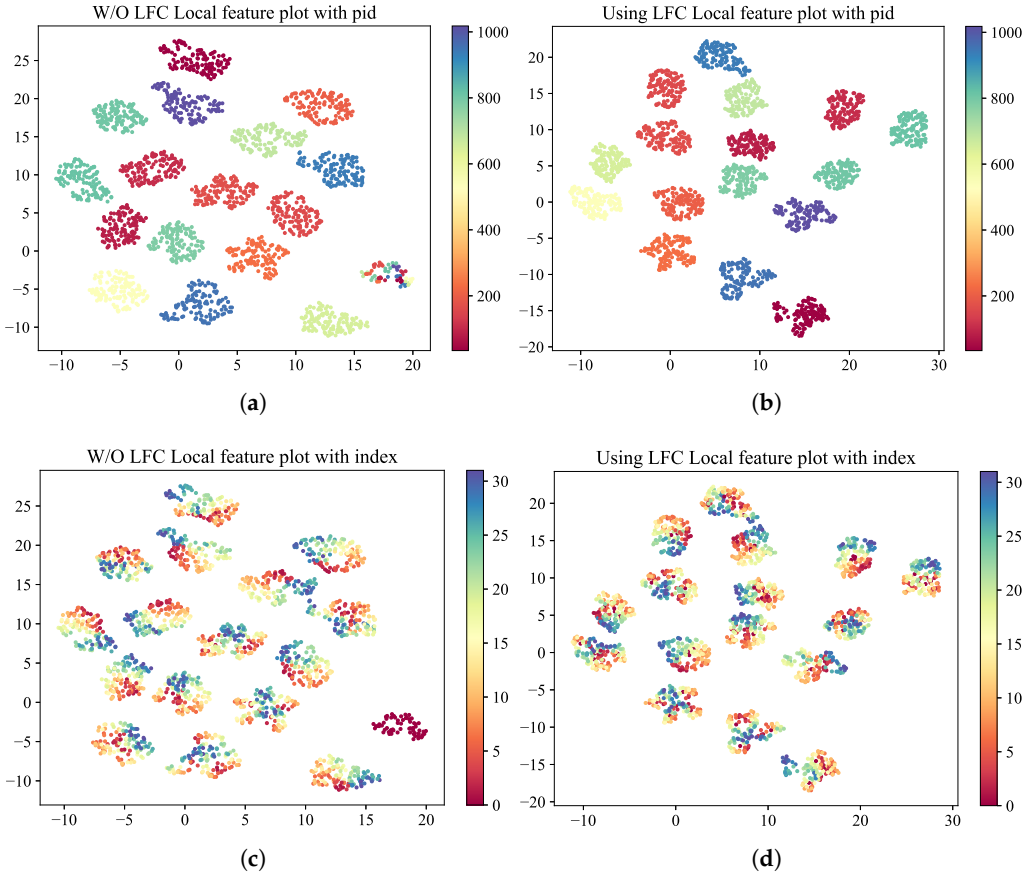


Figure 3. Local feature distribution. We used UMAP (n_neighbors = 10, min_dist = 1) [31] to reduce the dimension of local features: (**a**) without LFC and plot with ID; (**b**) with LFC and plot with ID; (**c**) without LFC and plot with local index; (**d**) with LFC and plot with index. By comparing the distribution of local features with and without the use of LFC, it can be concluded that local features clustered together with different IDs can be dispersed into clusters of the corresponding ID by clustering the most discrete local features.

In order to solve the problem that certain local features cannot be distinguished, we propose a local feature clustering (LFC) module. By computing the distance of each local feature f_i from the global feature f, we select the local feature with the farthest

distance. Then, through supervised learning we separate the aggregated local feature in Figure 3a,c. The process of selecting the farthest local feature f_m from the global feature f is shown below:

$$f_m = \underset{f_i \in \{f_1, f_2, \ldots, f_{32}\}}{\operatorname{argmax}} \|f_i - f\|_2^2 \tag{8}$$

Through Equation (8), we can obtain the most discrete local feature f_m. This means that if f_m is clustered with other local features of the same image, the global feature f obtained by aggregating the local features has better discriminative power.

As illustrated in Figure 2, another global branch which parallels to the LFC branch obtains f, which is the global feature from the CNN method, using the mean operation. Finally, the loss is computed using the ID loss and triplet loss for the global feature f and the most discrete local feature f_m, respectively, which are then added together using a specific factor. The total training loss is

$$L = L_{ID}(f) + L_T(f) + \eta[L_{ID}(f_m) + L_T(f_m)]. \tag{9}$$

During the inference process, the global feature f and the most discrete local feature f_m with respect to $[f; \eta f_m]$ are connected as the ultimate feature representations.

3.3. Side Information Embeddings

We encode the camera information using SIE, add it to the feature sequence obtained from PatchEmbed, and send it to Encoder to participate in training. The robustness of the features is improved by reducing interference caused by nonvisual information.

Although the feature representations are obtained through the PVT network, these features are affected by nonvisual information such as cameras, angles, etc. In other words, even a well-trained ReID framework might not be able to differentiate between people with the same ID who are captured by different cameras. To reduce the impact of nonvisual information on person features, we use side information embedding (SIE) to embed nonvisual information into the feature extraction network in order to extract more robust person features.

Specifically, assuming that an ReID dataset has N_C cameras in total, we initialize the learnable camera ID information embedding as $E_C \in R^{N_c \times D}$. When the camera ID of a person image is r, then the person ID encoding of this image is $E_C[r]$, and for all patches of the image its $E_C[r]$ is the same.

Embedding SIE into the network represents a problem. The simplest way is to directly add the patch embeddings and SIE. However, considering the need to balance the weight between vision information and SIE, the SIE coefficient needs to be set according to specific conditions. If the coefficient is too large, SIE will dominate and the role of vision information will be ignored. Similarly, if the coefficient is too small, the role of SIE will be ignored. The input sequence with camera ID information is sent to the Encoder as shown below:

$$I' = I + \lambda E_c(r). \tag{10}$$

where I is the input sequence of each encoding stage and λ is a hyperparameter used to balance SIE. This process takes place after the PatchEmbed Equation (2) and prior to the Encoder Equation (3). The final sequence of feature inputs is

$$F_i' = PatchEmb_i(I_i) + I'. \tag{11}$$

4. Experiments

4.1. Datasets

We conducted an evaluation of our proposed methods on three ReID datasets: Market-1501 [32], MSMT17 [33], and DukeMTMC-reID [34]. Each image in these datasets contains a camera ID; Table 1 shows the detailed dataset information.

Table 1. Statistics of ReID datasets used in our experiments.

Dataset	#Camera	#Image	#ID
Market-1501	6	32,668	1501
MSMT17	15	126,441	4101
DukeMTMC-reID	8	36,441	1404

4.2. Implementation

Except for special datasets, the image resolution of common ReID datasets is 256×128; thus, there is no need to process the image size. During image processing, we used random erasing [35], random cropping, random padding, and horizontal flipping [36] to process the training set images. There were 128 images per mini-batch and four images per person ID. In the training process of the model, we optimized the model using AdamW with a momentum of 0.9 and weight decay of 5×10^{-2}. The initial learning rate was set to 8×10^{-3} and decreased following the cosine schedule. More detailed information can be found in the Appendix A.

All of our experiments were performed on a system with one Nvidia RTX 3090 GPU using the PyTorch toolbox. The initial weights used for the PVT were pretrained on ImageNet-1K.

4.3. Evaluation Protocol

We evaluated the model in terms of its accuracy and inference speed. As is customary in the ReID research community, we evaluated the accuracy of all methods using the Cumulative Matching Characteristic (CMC) and the mean Accuracy Precision (mAP). In terms of inference speed, there is no unified standard for evaluation; in this paper, we evaluate the inference speed of the model using the feature extraction speed.

The CMC is calculated by summing the Acc_k of each query image and dividing it by the total number of query images, usually denoted as Rank-k, e.g., Rank-1 accuracy denotes the probability of correctly matching the top-ranked gallery image in the match list.

$$Acc_k = \begin{cases} 1 & top-k \\ 0 & othercases. \end{cases} \tag{12}$$

The mAP is used to evaluate overall model performance. For each query, we compute the average precision (AP). The mAP is obtained by calculating the average of the APs of all queries, which takes into account both the precision and recall of the query images. Therefore, mAP provides a more comprehensive evaluation.

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{13}$$

The time consumption of ReID required for inference is mainly divided into two parts, namely, the feature extraction time T_e and similarity computation time T_c. All ReID methods have the same T_c for the same dataset, and only T_e is taken into account. The inference speed of the ReID method is obtained using the total number of query images divided by T_e.

4.4. Results for the PVT-Based Basic Network

In this section, we compare the performance of different backbone networks on the ReID task, with the results shown in Table 2. Several different backbone networks were chosen as the feature extraction network for ReID to display the trade-off between computation consumption and performance: ViT-Base, PVT-V2-B2, and PVT-V2-B5, denoted as ViT-B, PVT-v2-b2, and PVT-v2-b5, respectively. In order to comprehensively compare the different backbone networks, we took into account the parameters, inference time, and performance.

Table 2. Backbone network comparison for ReID. The inference speed is expressed in terms of a comparison of each model with ResNet50, as only relative comparisons are required. All experiments were performed on an identical computer to allow for fair comparisons.

Backbone	Params (M)	Inference Speed	MSMT17	
			mAP	R1
ResNet50	23.5	1.0×	51.3	75.3
ResNet101	44.5	1.48×	53.8	77.0
ResNet152	60.2	1.96×	55.6	78.4
ResNeSt50	25.6	1.86×	61.2	82.0
ResNeSt200	68.6	3.12×	63.5	83.5
ViT-B	86.0	1.79×	61.0	81.8
PVT-V2-B2	25.4	0.82×	54.1	77.3
PVT-V2-B5	82.0	1.22×	60.2	81.2

From the above results, a huge gap in the ability of the models to extract person features can observed between ResNet and PVT. Compared with ResNet50, PVT2-B2 uses more parameters to achieve slightly better performance in terms of inference speed and accuracy. PVT2-B5 has similar performance to ResNest50 [37] with a significantly lower inference time (1.22× vs. 1.86×). PVT2-B5 uses fewer parameters and has a lower inference time (1.22× vs. 1.79×), achieving results comparable to ViT-B.

4.5. Ablation Study of LFC

The effectiveness of the proposed LFC module is validated in Table 3. LFC confers an improvement of +1.3% mAP on Market1501, +1.9% mAP on MSMT17 and +1.5% mAP on DukeMTMC-reID compared to the basic network. Comparing LFC and LFC without local features, it can be observed that when both are trained using LFC in the training stage, using only global features ("w/o local") in the inference stage results in slightly worse performance than the full version, while the inference times are similar.

Table 3. Local feature clustering ablation study; "w/o local" indicates that only global features were evaluated.

Backbone	Market1501		MSMT17		DukeMTMC-reID	
	mAP	R1	mAP	R1	mAP	R1
Basic	86.3	94.9	60.2	81.2	77.8	87.9
+LFC	87.6	95.1	62.1	82.1	79.3	88.6
+LFC w/o local	87.6	95.1	62.0	82.1	79.2	88.4

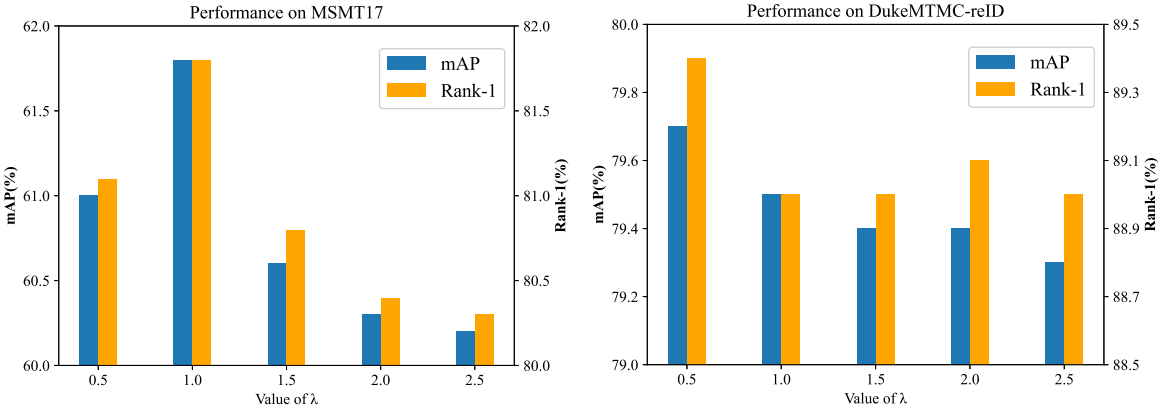
Figure 3a,c shows that most of the local features in the feature space are aggregated around the corresponding ID clusters, while a small portion are in a discrete state. After the introduction of the LFC module in the base network, Figure 3b,d shows the discrete local features clustered around the corresponding IDs, which indicates that the LFC module can alleviate the problem of discrete local features and improve the robustness of the resulting features.

4.6. Ablation Study of Camera Information

Studying the side information in the three datasets leads to the conclusion that only the camera information can be effectively applied as SIE information. In Table 4, we compared the performance of the camera SIE on the Market1501, MSMT17, and DukeMTMC-reID datasets while simultaneously investigating the effects of adding SIE at four different stages on the accuracy of PVT-based ReID. Figure 4, shows the results of our evaluation of the impact of camera information with weights λ on the model's performance on the MSMT17 and DukeMTMC-reID datasets.

Table 4. Ablation study of SIE. Because PVT has four embedding stages, we added SIE at several different stages. Note that λ in Equation (10) is 1.0. ✔ means that SIE modules are used by the coding stage.

Method	Embed Stage				Market1501		MSMT17		DukeMTMC-reID	
	1	2	3	4	mAP	R1	mAP	R1	mAP	R1
Basic					86.3	94.9	60.2	81.2	77.8	87.9
+SIE	✔				86.7	94.7	61.8	81.9	79.4	88.6
		✔			86.6	94.5	61.7	82	79.2	88.5
			✔		86.4	94.4	61.2	81.5	78.8	88.3
				✔	86.0	94.2	59.7	81.1	78.3	88.1



(**a**) MSMT17 (**b**) DukeMTMC-reID

Figure 4. Influence of λ on SIE.

4.6.1. Performance Analysis

From Table 4, it can be concluded that applying SIE in the first and second stages is much better than applying it in the third and fourth stages. In fact, using SIE in the fourth stage can even lead to a decrease in accuracy for PVT-based ReID. In summary, the earlier SIE is applied in PVT, the higher the model accuracy.

When SIE is encoded only in the first stage, the Basic + SIE improves by 0.7% in terms of rank-1 accuracy and 1.6% in terms of mAP on the MSMT17 dataset compared to the basic network. A very similar result is reached on the DukeMTMC-reID dataset, where Basic + SIE improves rank-1 accuracy by 0.7% and mAP by 1.6%. However, on the Market1501 it only obtains a 0.4% improvement in mAP. Because PVT is a variant of ViT, the problem of easy overfitting on small datasets persists. However, considering that the DukeMTMC-reID dataset is similar in size to Market-1501, the most likely reason that SIE does not improve as much on Market-1501 as on the other datasets is be that the number of cameras is too small in this case.

4.6.2. Ablation Study of λ

In Figure 4, when λ = 0, Basic achieves 60.2% mAP on MSMT17 and 77.8% mAP on DukeMTMC-reID. With increasing λ, Basic achieves 79.7% mAP on DukeMTMC-reID when λ = 0.5 and 61.8% mAP on MSMT17 when λ = 1.0. This performance indicates that SIE helps to reduce the impact of environmental factors on the robustness of person features and can help the model to learn invariant features. Continuing to increase the value of λ results in the model's performance decreasing. This is due SIE dominating person features and the role of vision information being weakened when λ is too large,.

4.7. Ablation Study of PVTReID

The effectiveness of the two proposed modules is evaluated in Table 5. Compared to the Basic network, the LFC module and SIE module increase performance by +1.3%/+1.9%/

+1.5 mAP and +0.4%/+1.6%/+1.9% mAP, respectively, on the Market1501/MSMT17/Duke MTMC-reID databases. With these two modules used together, PVTReID achieves 87.8% (+1.5%) mAP, 63.2% (+3.0%) mAP, and 80.5% (+2.7%) mAP on the Market1501, MSMT17, and DukeMTMC-reID databases, respectively. The effectiveness of our proposed ReID method and modules is further demonstrated by these experimental results.

Table 5. Results of PVTReID ablation study. ✘ means that the module is not used at method. ✔ means that the module is used at method.

Method	LFM	SIE	Market1501		MSMT17		DukeMTMC-reID	
			mAP	R1	mAP	R1	mAP	R1
Basic	✘	✘	86.3	94.9	60.2	81.2	77.8	87.9
	✔	✘	87.6	95.1	62.1	82.1	79.3	88.6
	✘	✔	86.7	94.7	61.8	81.9	79.7	89.3
PVTReID	✔	✔	87.8	95.0	63.2	82.3	80.5	90.0

4.8. Comparison to State-of-the-Art Methods

In Table 6, we compare our PVTReID to state-of-the-art methods on the three benchmarks datasets Market1501, MSMT17, and DukeMTMC-reID. On large datasets, the overall performance of PVTReID is significantly better than previous state-of-the-art methods. Specifically, on MSMT17, PVTReID achieves a 2.4% improvement in mAP. On DukeMTMC-reID, PVTReID obtains a 0.5% improvement in mAP. On smaller datasets such as Market-1501, PVTReID lags slightly behind a few other state-of-the-art methods.

Table 6. Comparison with state-of-the-art methods.

Backbone	Method	Size	Inference (Images/s)	Market1501		MSMT17		DukeMTMC-reID	
				mAP	R1	mAP	R1	mAP	R1
CNN	CBN [2]	256 × 128	338	77.3	91.3	42.9	72.8	67.3	82.5
	OSNet [38]	256 × 128	2028	84.9	94.8	52.9	78.7	73.5	88.6
	SAN [39]	256 × 128	290	88.0	**96.1**	55.7	79.2	75.7	87.9
	PGFA [40]	256 × 128	263	76.8	91.2	-	-	65.5	82.6
	HOReID [41]	256 × 128	310	84.9	94.2	-	-	75.6	86.9
	ISP [42]	256 × 128	315	**88.6**	95.3	-	-	**80.0**	89.6
	MGN [43]	384 × 128	287	86.9	95.7	52.1	76.9	78.4	88.7
	SCSN [8]	384 × 128	267	88.5	95.7	58.5	**83.8**	79.0	**91.0**
	ABDNet [9]	384 × 128	223	88.3	95.6	**60.8**	82.3	78.6	89.0
PVT	Basic	256 × 128	359	86.3	94.9	60.2	81.2	77.8	87.9
	PVTReID	256 × 128	341	87.8	95.3	63.2	82.3	80.5	90.0

- denotes absence of data.

Although CNN-based ReID methods mostly use the ResNet50 backbone to extract person features, they may contain several branches (e.g., attention modules, pose estimation models, and other modules) which increase computational cost. We conducted a fair comparison of inference speed between PVTReID and CNN-based ReID on the same hardware. OSNet does not use ResNet50 as the backbone, instead using a self-designed CNN. Compared to state-of-the-art methods using ResNet50, PVTReID is 0.8% slower than ISP in terms of mAP and 8% faster than ISP in terms of inference speed on the Market-1501 dataset; on MSMT17, PVTReID shows an improvement of 2.4% in mAP compared to ABDNet and a 52% improvement in inference speed, while on the DukeMTMC-reID dataset PVTReID shows an improvement of 0.5% in mAP and 8% in inference speed compared to ISP. The reasons for these results are as follows: first, most CNN-based methods use complex branches to improve their fine-grained discrimination ability, which generates additional resource consumption; second, several ReID methods (e.g., ABDNet) use 384 × 128 size images, which increases resource consumption during feature extraction;

third, PVTReID is based PVT, which has inherent disadvantages on small-scale datasets; fourth, while PVTReID has additional branches, these branches are not complex and the added resource consumption is not large. Therefore, PVTReID can achieve faster inference speed with accuracy comparable to the majority of CNN-based methods.

5. Conclusions

In this article, we use PVT for ReID and propose two modules: a local feature clustering (LFC) module and a side information embeddings (SIE) module. The proposed PVTReID model achieves 87.8%, 63.2%, and 80.5% mAP, respectively, on the popular ReID datasets Market1501, MSMT17, and DukeMTMC-reID, with fast inference speed. From our theoretical analysis and experimental results, the following conclusions can be drawn: (1) PVT used for ReID can provide improved inference speed under the premise of extracting effective person features; (2) clustering the discrete local features can improve the robustness of the resulting features; (3) encoding the camera information to participate in the feature extraction process can reduce the influence of nonvisual information on the extracted features; and (4) using the Transformer structure for ReID has inherent problems with poor performance on small datasets.

Based on the good results achieved by the proposed PVTReID model, we believe that PVT has great potential for further development on the ReID task. It is possible to use unsupervised learning to pretrain the PVT model on a large ReID dataset to solve the cross-domain problem that exists between the current pretraining ImageNet dataset and the ReID datasets used for testing. The PVT network outputs four different feature maps, and PVT feature maps with different resolutions can be used in the ReID task. The current PVTReID does not run fast enough to be reliably used in practice; thus a future direction is to simplify the proposed PVTReID model using knowledge distillation to generate a lighter version.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are provided in the paper.

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

Appendix A. Additional Experimental Results

Study of Basic PVT-Based Network

Section 3.1 of this paper describes the underlying PVT-based network with important improvements. In this Appendix, the hyperparameter settings for training such a network are analyzed in detail. Table A1 shows the results of the ablation study on the Market-1501 database with different training settings.

Table A1. Ablation study on the Market-1501 database used for training hyperparameter settings. The first line corresponds to the default configuration used for the basic PVT-based network (with PVT-V2-B5 serving as the default backbone). The ✔ indicates that the corresponding settings are included. The abbreviations LR, DR, ADR, DPR, and LS denote the learning rate, drop rate, attention drop rate, drop path rate, and label smoothing, respectively.

Method	LR	DR	ADR	DPR	LR	Market1501	
						mAP	R1
Basic	0.0008	0.1	0.1	0.3	✔	86.3	94.9
Learning Rate	0.008	✔	✔	✔	✔	86.0 (−0.3)	94.1 (−0.8)
	0.004	✔	✔	✔	✔	86.1 (−0.2)	94.2 (−0.7)
	0.002	✔	✔	✔	✔	86.2 (−0.1)	94.5 (−0.4)
	0.001	✔	✔	✔	✔	86.3 (0.0)	94.7 (−0.2)
	0.0006	✔	✔	✔	✔	86.2 (−0.1)	94.6 (−0.3)
	0.0004	✔	✔	✔	✔	86.0 (−0.3)	94.2 (−0.7)
Drop Rate	✔	0.0	✔	✔	✔	86.1 (−0.2)	94.4 (−0.6)
	✔	0.2	✔	✔	✔	85.5 (−0.8)	94.0 (−0.9)
Attention Drop	✔	✔	0.0	✔	✔	86.2 (−0.1)	94.2 (−0.7)
	✔	✔	0.2	✔	✔	85.2 (−0.1)	94.3 (−0.6)
Drop Path	✔	✔	✔	0.0	✔	84.9 (−1.4)	93.6 (−1.3)
	✔	✔	✔	0.1	✔	85.7 (−0.6)	94.1 (−0.8)
	✔	✔	✔	0.2	✔	86.1 (−0.2)	94.6 (−0.3)
	✔	✔	✔	0.4	✔	86.0 (−0.3)	94.4 (−0.5)
Loss Function	✔	✔	✔	✔	✘	86.0 (−0.2)	94.2 (−0.7)

References

1. Luo, H.; Jiang, W.; Fan, X.; Zhang, S. A survey on deep learning based person re-identification. *Acta Autom. Sin.* **2019**, *45*, 2032–2049.
2. Zhuang, Z.; Wei, L.; Xie, L.; Zhang, T.; Zhang, H.; Wu, H.; Ai, H.; Tian, Q. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XII; Springer: Cham, Switzerland, 2020; pp. 140–157.
3. Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; Yang, Y. Improving person re-identification by attribute and identity learning. *Pattern Recognit.* **2019**, *95*, 151–161. [CrossRef]
4. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part VII; Springer: Cham, Switzerland, 2016; pp. 499–515.
5. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 480–496.
6. Zheng, Z.; Zheng, L.; Yang, Y. A discriminatively learned cnn embedding for person reidentification. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2017**, *14*, 1–20. [CrossRef]
7. Zhang, Z.; Lan, C.; Zeng, W.; Jin, X.; Chen, Z. Relation-aware global attention for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3186–3195.
8. Chen, X.; Fu, C.; Zhao, Y.; Zheng, F.; Song, J.; Ji, R.; Yang, Y. Salience-guided cascaded suppression network for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3300–3310.
9. Chen, T.; Ding, S.; Xie, J.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; Wang, Z. Abd-net: Attentive but diverse person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8351–8361.
10. 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 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
11. 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.
12. Peng, Z.; Huang, W.; Gu, S.; Xie, L.; Wang, Y.; Jiao, J.; Ye, Q. Conformer: Local features coupling global representations for visual recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 367–376.

13. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; 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.
14. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
15. Islam, M.A.; Jia, S.; Bruce, N.D. How much position information do convolutional neural networks encode? *arXiv* **2020**, arXiv:2001.08248.
16. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
17. Luo, H.; Jiang, W.; Zhang, X.; Fan, X.; Qian, J.; Zhang, C. Alignedreid++: Dynamically matching local information for person re-identification. *Pattern Recognit.* **2019**, *94*, 53–61. [CrossRef]
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
19. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person re-identification: Past, present and future. *arXiv* **2016**, arXiv:1610.02984.
20. Matsukawa, T.; Suzuki, E. Person re-identification using CNN features learned from combination of attributes. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 2428–2433.
21. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
22. Yuan, Y.; Chen, W.; Yang, Y.; Wang, Z. In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 354–355.
23. Luo, H.; Gu, Y.; Liao, X.; Lai, S.; Jiang, W. 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, Long Beach, CA, USA, 16–17 June 2019.
24. Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; Wei, Y. Circle loss: A unified perspective of pair similarity optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6398–6407.
25. Wei, L.; Zhang, S.; Yao, H.; Gao, W.; Tian, Q. Glad: Global-local-alignment descriptor for pedestrian retrieval. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 420–428.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–15.
27. 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 vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [PubMed]
28. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM Comput. Surv. (CSUR)* **2022**, *54*, 1–41.
29. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pvt v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media* **2022**, *8*, 415–424.
30. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
31. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, arXiv:1802.03426.
32. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1116–1124.
33. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 79–88.
34. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–10 October 2016; pp. 17–35.
35. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13001–13008.
36. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
37. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. Resnest: Split-attention networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 2736–2746.
38. 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.
39. Jin, X.; Lan, C.; Zeng, W.; Wei, G.; Chen, Z. Semantics-aligned representation learning for person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11173–11180.

40. Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; Yang, Y. Pose-guided feature alignment for occluded person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 542–551.
41. Wang, G.; Yang, S.; Liu, H.; Wang, Z.; Yang, Y.; Wang, S.; Yu, G.; Zhou, E.; Sun, J. High-order information matters: Learning relation and topology for occluded person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6449–6458.
42. Zhu, K.; Guo, H.; Liu, Z.; Tang, M.; Wang, J. Identity-guided human semantic parsing for person re-identification. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part III; Springer: Cham, Switzerland, 2020; pp. 346–363.
43. Wang, G.; Yuan, Y.; Chen, X.; Li, J.; Zhou, X. Learning discriminative features with multiple granularities for person re-identification. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 274–282.