PVTReID: A Quick Person Re-Identification Based Pyramid Vision Transformer

Ke Han *, QianLong Wang, Mingming Zhu, Xiyan Zhang

Posted Date: 4 August 2023

doi: 10.20944/preprints202308.0373.v1

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

Article

PVTReID: A Quick Person Re-Identification Based Pyramid Vision Transformer

Ke Han [1,†], QianLong Wang [2], Mingming Zhu [3] and Xiyan Zhang [4]

[1] North China University of Water Resources and Electric Power; hanke@ncwu.edu.cn
[2] wql19971225@gmail.com
[3] 1071937012@qq.com
[4] 3072683664@qq.com
* Correspondence: wql19971225@gmail.com
† Current address: School of Software, North China University of Water Resources and Electric Power, ZhengZhou 450046 ,China.

Abstract: Due to the influence of background conditions, lighting conditions, occlusion issues and the image resolution, how to extract robust person features is one of the difficulties in ReID research. Vision in Transformers (ViT) has achieved significant results in the field of computer vision. However, the existing problems still limit its application in ReID due to slow extraction of person features and difficulty in utilizing local features of people. To solve the mentioned problems, we utilize Pyramid Vision Transformer (PVT) as the backbone of feature extraction and propose a PVT-based ReID method in conjunction with other studies. Firstly, some improvements suitable for ReID are used on the PVT backbone, and we establish a basic model by using powerful methods verified on CNN-based ReID. Secondly, in an effort to further promote the robustness of the person features extracted by the PVT backbone, two new modules are designed. (1) The local feature clustering (LFC) is recommend to enhance the robustness of person features by calculating the distance between local features and global feature to select the most discrete local features and clustering them. (2) The side information embeddings (SIE) are used to encode non-visual information and send it into the network for training to reduce its impact on person features. Finally, the experiments show that PVTReID has achieved excellent results in ReID datasets and are 20% faster on average than CNN-based ReID methods.

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

1. Introduction

In recent times, people have paid more and more attention to the public safety, and improving the ability to retrieve specific persons is particularly important. Person re-identification (ReID) is a technology for searching certain persons 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 also be used for person tracking in smart devices. Due to the influence of visual factors such as lighting, posture, occlusion, 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 effect of objective causes like background interference, person blockage, and posture misalignment makes it difficult for global feature to meet the requirements of person retrieval. The common practice of CNN-based ReID is to combine local features extraction to obtain fine-grained person features and make up for the shortcomings of global feature. With the development of vision attention, the application of attention to ReID [7–9] has also become popular. With the vast accomplishment of ViT [10] in the field of computer vision, a number of scholars started to explore the application of ViT to the ReID task [11] by extracting global person features through the Transformer.

By reviewing CNN-based methods and ViT-based ReID, we discover three significant matters which are not well solved in ReID. (1) The inference speed is slow. Most CNN-based ReID methods combine local features to advance the fine-grained ability to identify people, but these additional structures greatly increase the model complexity and computational consumption [1]. Due to the large amount of computing resources required by Transformer, the ViT-based ReID also has a slow feature extraction speed. (2) ViT is designed for image classification tasks, the network needs to use a class token for classification to reduce preferences for specific image patches. It learns the information of image patches and overly relies on global information, which reduces the use of local features and decreases fine-grained identification ability [12]. (3) ViT uses position encoding to process spatial information. Position encoding corresponds to the input image size and resolution. The position embedding requires retraining when dealing with images of different resolutions.

In order to solve the above problems, we present a ReID method using Pyramid Vision Transformer (PVT) based approach. PVT [13] has achieved high accuracy not only in image classification tasks but also in other downstream tasks. PVT uses pyramid feature structure [14] and zero-padding [15] to solve the mentioned problems very well. (1) Unlike ViT, PVT uses a pyramid structure, which greatly reduces the calculation consumption and boosts the efficiency of extracting person features. (2) PVT does not use an additional token as the output vector of the feature extraction, but aggregates all local features trained by Transformer to obtains a global feature representation, increasing the dependence of the output feature on local features while reducing model complexity and computational consumption and improving model training and inference efficiency. (3) 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, but it still needs to be modified on to adapt to the unique challenges in ReID (such as occlusion, pose changes, 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 to strengthen the robustness of person features. At the same time, the person features extracted by the CNN-based methods are different from those extracted by the PVT-based method, and their semantic information also differ. Furthermore, considering side information such as camera can reduce the impact of non-visual factors on the robustness of person features. If complex information constructed on CNN is directly used for PVT, the encoding ability and pyramid structure of PVT cannot be fully utilized. To successfully solve these problems, we should design modules specifically for PVT.

For these reasons, we present a novel ReID method called PVTReID for obtaining robust person features. Firstly, we have built a strong basic network based on PVT with a few crucial adaptations. Secondly, in view of solving the problem of some local features of person 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 in accompanying with the global feature brance. Thirdly, to further improve the robustness of person features, we employ side information embeddings (SIE) to embed side information into the four encoding stages of the PVT.

The purposes of this paper are summarized as follow:

- We recommend a strong basic network that utilizes PVT for ReID and carry out performance which are able to be compared with CNN-based methods.
- We introduce a local feature clustering (LFC) module, consisting of calculating and optimizing operation, which makes feature representation of person more robust.
- We encode the camera information by SIE and send it to the feature extraction network to improve the robustness of the person features, and verify the effect of camera information in different stages of PVT.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0373.v1

- The final framework PVTReID achieves comparable performance on ReID benchmarks including Market-1501, MSMT17, DukeMTMC-reID, and has faster speed on inference compared to CNN-based methods.

2. Related Work

2.1. Person Re-identification

Many CNN-based ReID methods have been proposed and proved to have good performance in recent years. A popular pipeline is to build on the 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 a image classification task, regards each person ID as a category, and uses the global feature 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 image clustering and it aims to find out the distance between two images in the feature space by learning. In 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 ID. 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] came up with BNNeck and Sun *et al*. [24] submitted Circle Loss, both of which have made good explanations on how to use ID Loss and Triplet Loss.

Representation learning and metric learning use the global feature of people. In the case of misaligned posture, occlusion of person images, and only local details are dissimilar, global feature is prone to making mistakes. The local feature methods can solve these problems to some extent by mining person 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, and then uses 1×1 convolution to obtain independent local features, and then predicts classification based on these local features. Local feature methods usually add a branch on the basis of global feature, which can extract rich person features but increase model inference time.

2.2. Vision Transformer

Ashish Vaswani *et al*. [26] proposes the Transformer model to process sequence data in natural language processing (NLP). Inspired by Transformer, numerous researchers have explored its application in computer vision. Han [27] and Salman [28] investigate the application of Transformer in computer vision and showed its effectiveness in different tasks. ViT [10] divides the image into patches and flattens them into a sequence of one-dimensional vectors as input to the Transformer. The uniqueness of ViT lies in that a learnable embedding is added to extract global feature, reduce the preference for a certain image patch when classifying images, but reduce 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 removed position encoding through 3×3 convolution, making it flexible to handle inputs of various resolutions.

2.3. Side Information

In the process of ReID, it is very common for people to change their posture and the resolution of images due to different camera angles. In order to resolve these problems, some works use side information such as camera information and viewpoint information to enhance the robustness of the learned features. CBN [2] transforms person images under different cameras into the same subspace,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0373.v1

and the distribution difference of person images taken by different cameras has been effectively improved. TransReID [11] sends side information encoding into the ViT for training, reduces the influence of camera factors on person features, and improves the robustness of person features.

3. Methods

Our ReID framework is designed on the basis of PVT-based image classification network, and using some useful improvements [23] to get the robust person feature. To further enhance the robustness of person feature training in many challenges, in Section 3.2 and Section 3.3, a local feature clustering module (LFC) and a side information embeddings (SIE) module are cautiously designed . An end-to-end network uses both modules at the same time and displas on Figure 2.

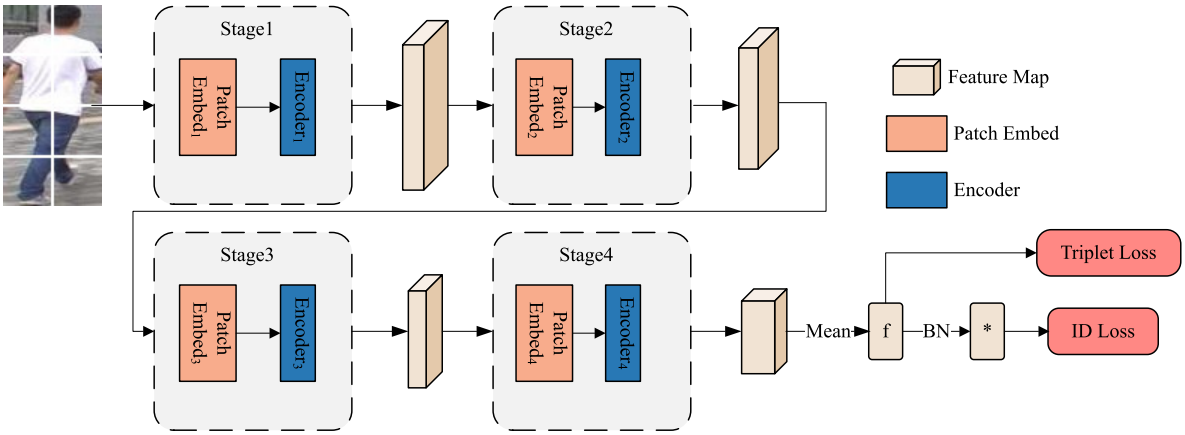


Figure 1. PVT-based basic framework. * is worked as the feature using Batch Normalization. Inspired by [23], we bring in the BNNeck before the ID Loss.

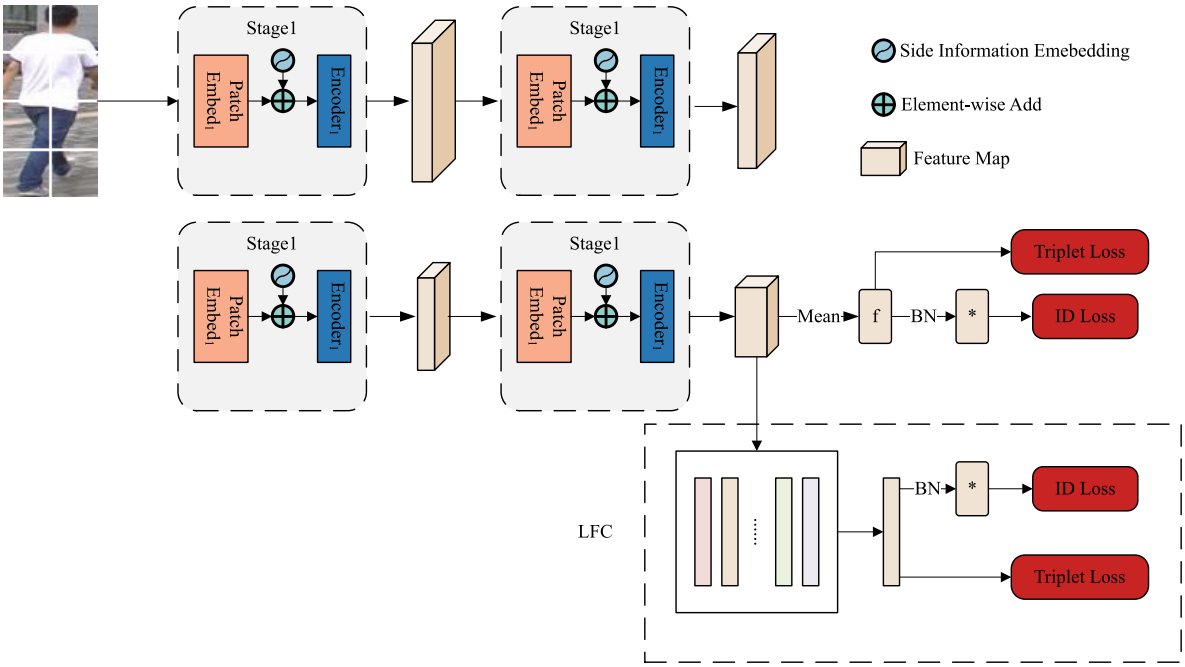


Figure 2. PVTReID Framework. Non-visual information such as camera is encoding into side information embeddings. Image patch embeddings with them are input into transformer encoder. Supervised learning consists of two separate branches. One branch is a common branch that acquires global features. The other branch uses the local feature clustering (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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0373.v1

3.1. PVT-based Basic Network

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

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

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

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

Given an image $x \in \mathbb{R}^{H \times W \times C}$ as input, where H, W, C respectively represent its height, width, and number of channels, we split the image into N fixed-sized patches by PatchEmbed. I_i represents the input vector of each part of PVT. The image x is the first part input vector of the PVT backbone network, and 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; F_i is calculated by $Encoder_i$ to get the output vector O_i of each part. And Repeating the above operations four times constitutes a PVT-based ReID backbone.

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

3.1.1. Patch Embed

Early Transformer models (such as ViT) divided images into non-overlapping image patches, which destroyed the structural features of local region. The PVT uses sliding windows to generate overlapping pixel image patches, which better preserves the integrity of the local features of person images. The PVT uses a pyramid structure to divide the entire feature extraction process into four parts, each of these requires the encoding of a feature map. Assuming the size of the image patch is P x P, and the step size is S, and S is less than P, then the area where the two image patches overlap is (P-S) x P. If an input feature map which has a resolution of H x W , it will be divided into N patches by using suitable P and S.

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

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

3.1.2. Feature Using

We optimize the PVT-ReID basic network by using classification loss and triplet loss for 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0373.v1

the total count of person IDs in the training set. We employ the cross-entropy loss for the ID loss L_{ID} along with label smoothing [30].

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

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

The sum of person IDs in training set is N, p_i is the predicted probability of person ID, and y represents the true ID of the predicted person. The equation 6 represents the label smoothing operation on the ID. ε is a hyper parameter. 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 soft-margin [21] is illustrated as follows:

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

f_a, f_p and f_n denote the feature representations of the persong 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 Triplet Lose.

3.2. Local Feature Clustering Module

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, only using global feature as the standard for person distance measurement cannot meet the discrimination needs of difficult samples. Therefore, we need to learn local features of people to improve fine-grained discrimination ability, such as stripe features and posture estimation have been widely used in CNN-based methods to extract 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 then the global feature f is obtained through mean operation. In order to obtain person local fine-grained features, straightforward approach is to discriminate ability for each local feature, that is, to cluster all local features which have same person ID in the feature space. The global feature f obtained after aggregating all local features will also be clustered, but it cannot take into account all local features and there wobtainill be situations that some local features cannot be distinguished, as shown in Figure 3(a) and Figure 3(c).

In order to solve the problem that some 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 3(a) and 3(c). The process of selecting the farthest local feature f_m from the global feature f is shown following:

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

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 discriminating power.

As illustrated in Figure 2, another global branch which parallels to the LFC branch, obtains f by mean operation. f is the global feature using the CNN method. Finally, the loss is computed by ID

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0373.v1

Loss and Triplet Loss for the global feature f and the most discrete local feature f_m , respectively, and then they are added together by a certain factor. The total training loss is:

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

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

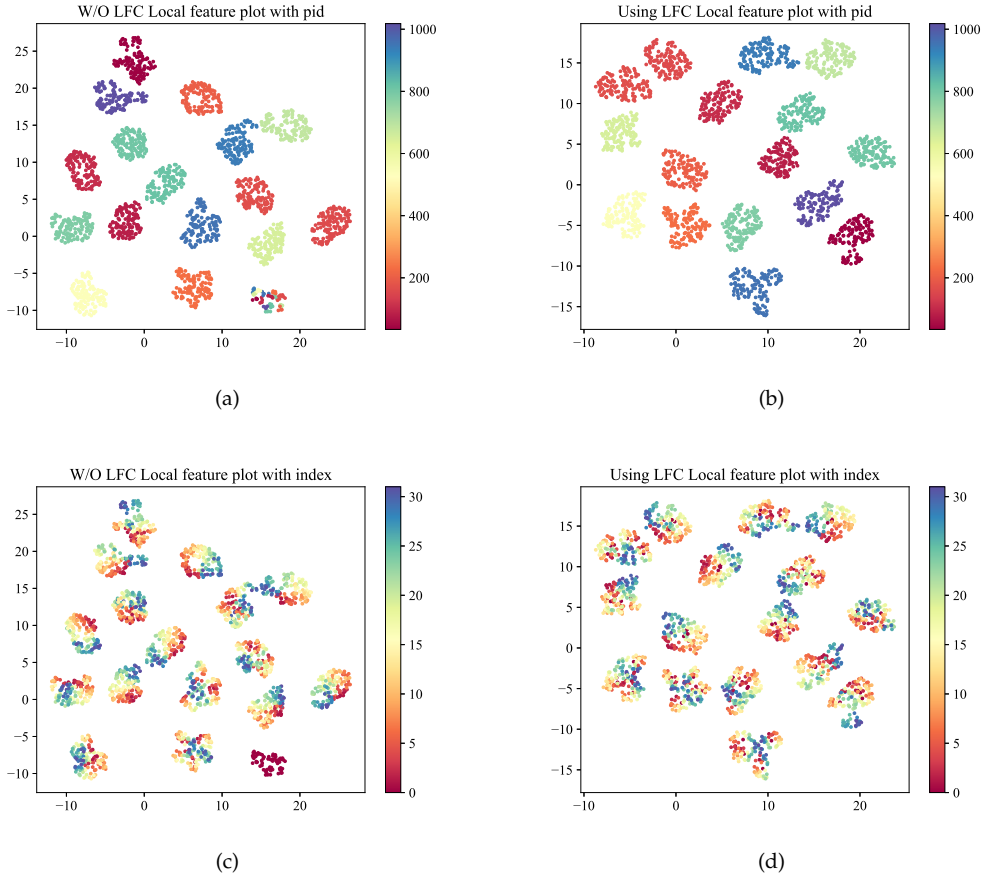


Figure 3. Local feature distribution. We use UMAP(n_neighbors=10, min_dist=1) [31] to reduce the dimension of local features. 3(a) w/o LFC and plot with id. 3(b) use LFC and plot with id. 3(c) w/o LFC and plot with local index. 3(d) use LFC and plot with index. By comparing the distribution of local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

3.3. Side Information Embeddings

Although the person feature representations are obtained through PVT network, these features are still affected by non-visual information like cameras, angles, etc. In other words, due to changes in the camera, a well-trained 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 non-visual information on person feature, we will use side information embedding (SIE) to embed non-visual information into the feature extraction network to extract more robust person features.

Specifically, assuming a 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0373.v1

How to embed SIE into the network for training is 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 Enconder as show in follow:

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

I is the input sequence of each encoding stage, and λ is a hyper parameter used to balance SIE. This process after PatchEmbed Eq.2 and before Encoder Eq.3.

4. Experiments

4.1. Datasets

We conducted an evaluation on three person ReID datasets for our proposed methods, Market-1501 [32], MSMT17 [33], DukeMTMC-reID [34]. Each image of above datasets is provided with camera ID, and Table 1 shows the above datasets detailed information.

Table 1. Statistics of commonly used ReID datasets.

Dataset	#camera	#image	#ID
Market-1501	6	32668	1501
MSMT17	15	126441	4101
DukeMTMC-reID	8	36441	1404

4.2. Implementation

Except for some special datasets, the image resolution of common ReID datasets is 256×128, and there is no need to process the image size. During the image processing, we use random erasing [35], random cropping random, padding and horizontal flipping [36] to process the training set images. There are 128 images per mini-batch and 4 images per person ID. In the training process of the model, we optimize the model using AdamW with momentum of 0.9 and weight decay of 5×10^{-2}. The initial learning rate is set to 8×10^{-3} and decreases following the cosine schedule. All the experiments are performed with one Nvidia RTX 3090 GPU. The initial weights of PVT are pre-trained on ImageNet-1K.

4.3. Results of PVT-based Basic Network

In this section, we compare the performance of different backbone networks on ReID in Table 2. Several different backbone networks are chosen as the feature extraction network for ReID to display the trade-off between computation consumption and performance. ViT-Base, PVT-V2-b2, PVT-V2-b5 denoted as ViT-B, PVT2-b2, PVT2-b5, respectively. To comprehensively compare the different backbone networks, we take into account the parameters, inference time and performance.

Between ResNet and PVT, we can observe a huge gap in the ability of the model to extract person features. Compared with ResNet50, PVT2-B2 uses more parameters to perform a little bit better in terms of inference speed and accuracy. PVT2-B5 fetches a similar performance to ResNest50 [37] backbone, but it employs significantly less inference time (1.22× vs 1.86×). PVT2-B5 uses fewer parameters and less inference time (1.22× vs 1.79×) less inference time, achieving results comparable to ViT-B.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0373.v1

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

Backbone	Params(M)	Inference Speed	MSMT17 mAP	MSMT17 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
PVT2-B2	25.4	0.82×	54.1	77.3
PVT2-B5	82.0	1.22×	60.2	81.2

4.4. Ablation Study of LFC

The effectiveness of the proposed LFC module is validated in Table 3. LFC respectively improves +1.3% mAP on Market1501, +1.9% mAP on MSMT17 and +1.5% mAP on DukeMTMC-reID compared to basic network. Comparing LFC and LFC w/o local, we can observe that, if both are trained using LFC in the training stage, using only global features ("w/o local") in the inference stage gives a slightly worse performance than the full version with similar inference times. The local features visualized in Figure 3 indicates that by clustering the local features of person images with the same ID, it can help the model learn more fine-grained features that are discriminatory, which makes the model more robust in the face of difficult samples.

Table 3. Local feature clustering ablation study. "w/o local" indicates that we evaluate global features and do not use local features.

Backbone	Market1501 mAP	Market1501 R1	MSMT17 mAP	MSMT17 R1	DukeMTMC-reID mAP	DukeMTMC-reID 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

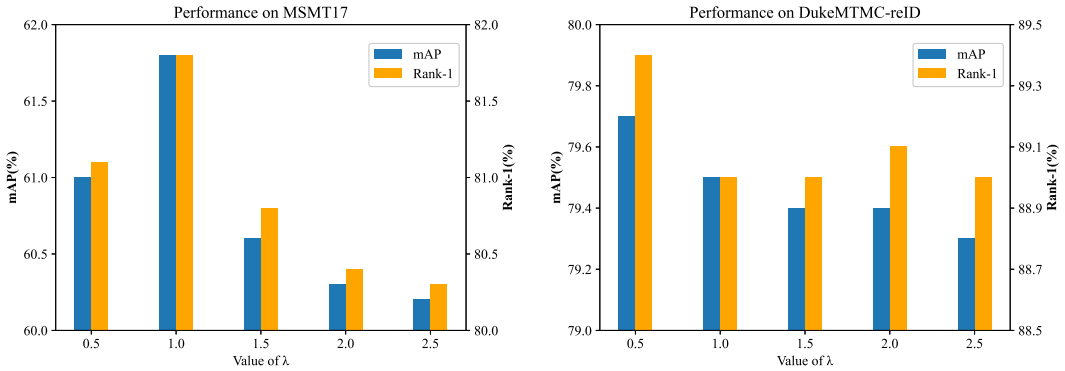
4.5. Ablation Study of Camera Information

By studying the side information in the three datasets, only the camera information can be effectively applied as SIE information. In Table 4, we compared the performance of the camera SIE on Market1501, MSMT17 and DukeMTMC-reID. Simultaneously we investigated the effects of adding SIE at four different stages on accuracy for PVT-based ReID. In Figure 4, we evaluated the impact of the camera information weights λ by the performance on MSMT17 and DukeMTMC-reID.

Table 4. Ablation study of SIE. Since PVT has 4 embed stages, we try to add SIE in different stages. λ is 1.0 in Eq.10.

Method	Embed Stage 1	2	3	4	Market1501 mAP	Market1501 R1	MSMT17 mAP	MSMT17 R1	DukeMTMC-reID mAP	DukeMTMC-reID 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0373.v1



(a) MSMT17 (b) DukeMTMC-reID

Figure 4. Influence of the λ in SIE.

4.5.1. Performance Analysis

Through Table 4, it can be concluded that applying SIE in the first and second stages is much better than 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 accuracy the model gets.

When SIE encodes only in the first stage, the Basic+SIE improves 0.7% rank-1 accuracy and 1.6% mAP on MSMT17 comparing to the bsic network. About the same result can be appeared on DukeMTMC-reID. Basic+SIE improves 0.7% rank-1 accuracy and 1.6% mAP. But Market1501 only obtains 0.4% mAP improvement. This could be due to the small size of the datasets or the limited number of cameras, which results in poor performance of SIE on Market1501.

4.5.2. Ablation Study of λ

In Figure 4, When = 0, Basic achieves 60.2% mAP on MSMT17 and 77.8% mAP on DukeMTMC-reID, respectively. With the increasing of λ , DukeMTMC-reID achieves 79.7% mAP when $\lambda = 0.5$, MSMT17 gets 61.8% mAP when $\lambda = 1.0$. The great performance means that SIE helps reduce the impact of environmental factors on the robustness of person features and helps learn invariant features. Continuing to increase the value of λ, the model's performance will decrease. This is due to when λ is too large, SIE dominates person features and the role of vision information is weakened.

4.6. Ablation Study of PVTReID

The effectiveness of the two proposed modules was evaluated in Table 5. Comparing to Basic network, LFC module and SIE module increase the performance by +1.3%/+1.9%/+1.5 mAP and +0.4%/+1.6%/+1.9% mAP on Market1501/MSMT17/DukeMTMC-reID, respectively. 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 Market1501, MSMT17 and DukeMTMC-reID, respectively. The effectiveness of our proposed ReID method and modules is further demonstrated by these experimental results.

Table 5. The ablation study of PVTReID.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0373.v1

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

In Table 6, we compared our PVTReID to state-of-the-art methods on three benchmarks, 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 achieve 2.4% mAP improvement. On DukeMTMC-reID, PVTReID obtains 0.5% mAP improvement. But in small datasets such as Market-1501, PVTReID slightly lags behind some 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]	384×129	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 the 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 but a self-designed CNN. Compared with CNN-based ReID using ResNet50, PVTReID has an average 20% faster inference speed . Therefore, PVTReID can achieve faster inference speed with accuracy comparable to the majority of CNN-based methods.

5. Conclusion

In this article, we use the PVT for ReID and proposed two modules, local feature clustering (LFC) module and the side information embeddings (SIE). The ultimate PVTReID performs better than other state-of-the-art methods on several popular ReID datasets consisting of Market1501, MSMT17 and DukeMTMC-reID whth faster inference speed. Since PVTReID has achieved good results, we believe that PVT has a lot of room for improvement in ReID. In the future we will focus on how PVT feature maps of different resolutions can be used for ReID.

References

1. Luo, H.; Jiang, W.; Fan, X.; Zhang, S. A survey on deep learning based person re-identification. *Acta Automatica Sinica* **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, August 23–28, 2020, Proceedings, Part XII 16. Springer, 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 recognition* **2019**, *95*, 151–161.
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, October 11–14, 2016, Proceedings, Part VII 14. Springer, 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 Proceedings of the European conference on computer vision (ECCV), 2018, pp. 480–496.
6. Zheng, Z.; Zheng, L.; Yang, Y. A discriminatively learned cnn embedding for person reidentification. *ACM transactions on multimedia computing, communications, and applications (TOMM)* **2017**, *14*, 1–20.
7. Zhang, Z.; Lan, C.; Zeng, W.; Jin, X.; Chen, Z. Relation-aware global attention for person re-identification. In Proceedings of the Proceedings of the ieee/cvf conference on computer vision and pattern recognition, 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 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 Proceedings of the IEEE/CVF international conference on computer vision, 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 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
11. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. Transreid: Transformer-based object re-identification. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 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 Proceedings of the IEEE/CVF international conference on computer vision, 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 Proceedings of the IEEE/CVF international conference on computer vision, 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
15. Islam, M.A.; Jia, S.; Bruce, N.D. How much position information do convolutional neural networks encode? *arXiv preprint arXiv:2001.08248* **2020**.
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 preprint arXiv:1704.04861* **2017**.
17. Luo, H.; Jiang, W.; Zhang, X.; Fan, X.; Qian, J.; Zhang, C. Alignedreid++: Dynamically matching local information for person re-identification. *Pattern Recognition* **2019**, *94*, 53–61.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
19. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984* **2016**.
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). IEEE, 2016, pp. 2428–2433.
21. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* **2017**.
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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2019, pp. 0–0.

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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 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 Proceedings of the 25th ACM international conference on Multimedia, 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. *Advances in neural information processing systems* **2017**, *30*.
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 transactions on pattern analysis and machine intelligence* **2022**, *45*, 87–110.
28. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM computing surveys (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. *Computational Visual 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.
31. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* **2018**.
32. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 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. Springer, 2016, pp. 17–35.
35. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2020, Vol. 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 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 Proceedings of the IEEE/CVF international conference on computer vision, 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 Proceedings of the AAAI Conference on Artificial Intelligence, 2020, Vol. 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 Proceedings of the IEEE/CVF international conference on computer vision, 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 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, August 23–28, 2020, Proceedings, Part III 16. Springer, 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 Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 274–282.

 doi:10.20944/preprints202308.0373.v1