

Article

A Patch Information Supplement Transformer for Person Re-Identification

Li Zhu ^{1,2,†}, Chenglong Jiang ^{2,†} and Minghu Wu ^{1,2,†*}

¹ Hubei Key Laboratory for High-Efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei University of Technology, Wuhan 430068, China

² Hubei Engineering Research Center for Safety Monitoring of New Energy and Power Grid Equipment, Hubei University of Technology, Wuhan 430068, China

* Correspondence: 102000215@hbut.edu.cn; Tel.: +86-131-6538-7373

† These authors contributed equally to this work.

Abstract: Extracting fine-grained features from person images has proven crucial in person re-identification (re-ID). Although the research of convolutional neural networks (CNN) has been very successful in person re-ID, due to the small receptive field and downsampling operation, the existing CNNs cannot solve the problem of information loss. The multi-head attention modules in transformer can solve the above problems well. However, since dicing operations destroy the spatial correlation between patches, transformer still loses some local features. In this paper, we propose the scheme of the patch information supplement transformer (PIT) to extract fine-grained features in the dicing stage. Patch pyramid network (PPN) is introduced to solve the problem of local information loss. This is accomplished by dividing the image into different scales through the dicing operation and adding them together from top to bottom according to the pyramid structure. In addition, we insert a learnable identity information-embedding module (IDE) to reduce the feature bias of clothing and camera perspective. Experiments verify the superiority and effectiveness of PIT compared to state-of-the-art methods.

Keywords: person re-identification; fine-grained features; transformer; patch pyramid network; identity information embedding module



Citation: Zhu, L.; Jiang, C.; Wu, M. A Patch Information Supplement Transformer for Person Re-Identification. *Electronics* **2023**, *12*, 1997. <https://doi.org/10.3390/electronics12091997>

Academic Editor: Gemma Piella

Received: 1 March 2023

Revised: 12 April 2023

Accepted: 20 April 2023

Published: 25 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Person re-identification [1] technology is a crucial but challenging task in the fields of multimedia and computer vision, which has been applied in various applications. The purpose of person re-identification is to associate a particular person quickly across different scenes and camera perspectives. The main challenge comes from the difficulty of accurately distinguishing a person with similar appearances. For example, as shown in Figure 1, the clothes of the two women are of the same colour, the postures of the two men have changed, and the occlusions cover many key pedestrian features. Therefore, extracting fine-grained features is crucial to person re-ID [2].

According to the authors in [3,4], CNN-based methods are the very intriguing in current research. However, the research methods based on CNN still face two challenges: (1) obtaining rich contextual information in global feature extraction is crucial for person re-ID [5]. Due to the convolution and the small receptive field, CNN methods only focus on local small feature information [6]. (2) Extracting detailed fine-grained features with detailed information in the person image is also important. However, the convolution and pooling operations of CNN reduce the resolution of the image, leading to feature information losses. This greatly affects the accuracy of fine-grained feature extraction, such as similar appearance [7].



Figure 1. Person re-identification is about identifying the correct person, which is extremely challenging due to similarities in appearance, changes in posture, and occlusion. Failure to consider the identity information of the person can lead to recognition errors.

The multi-head attention modules in transformer [8] and the operations of removing convolution and pooling can better solve the above-mentioned critical problems facing CNNs. The reasons are as follows. (1) The multi-head attention modules can capture the relationship between long distances and make the model concentrate more on the relationship between different regions. (2) Removing operations such as convolution and pooling to reduce feature loss allows the model to save more detailed feature information [9]. Combining the above advantages, we chose the transformer network as the baseline for person re-ID.

Although transformer has the above-mentioned significant advantages, it still has tremendous challenges in the face of person re-ID tasks, such as occlusion, lighting, camera perspective, pose diversity, etc. Therefore, considering the above problems, the transformer should be improved. Many attempts have been made to face this problem when researching person re-ID based upon the CNN. As for extracting fine-grained features, local part features [10–12] and auxiliary features [13,14] have been shown to be essential and effective. CNN-based methods extract features from a complete image, while the slicing operation in transformer cuts the image into independent patches, directly migrating CNN-based methods into transformer results in suboptimal recognition [15]. In addition, directly adding the CNN-based side semantic information module to the transformer network cannot improve the use of its encoding ability. Therefore, it is also a big challenge to embed the module based on the specific design of CNN into the transformer model.

We proposed a scheme of the PIT to solve the above-mentioned critical problems. Firstly, we proposed a PPN to solve the problem that the correlation between the image regions is destroyed due to the dicing operation. This divides the image into blocks of different scales by dicing, which are connected and added from top to bottom according to a pyramid structure for further fine-grained feature learning. The PPN structure can make the most miniature scale patch to obtain the feature information of other scale images to strengthen its fine-grained features. Therefore, the network can extract global perturbation-invariant and fine-grained features. To the best of our knowledge, the issue of transformer-based identity information encoding has not been investigated [16]. Different from utilizing semantic information embedding in CNN-based methods, we designed the IDE for Transformer, effectively integrating identity semantic information through learnable

embeddings to reduce feature bias of peripheral information. For instance, the proposed IDE can solve the problem of matching similarity caused by similar clothing.

In summary, this is the first study to design a PPN structure and apply it to person re-ID. The main contributions of this paper are as follows:

1. We designed a PPN structure and applied it with transformer for person re-ID, solving the global perturbation caused by spatial dimension segmentation and poor fine-grained features.
2. We provided an approach to identity information embedding, encoding identity information through learnable embeddings. Thus, effectively addressing the problem of learned feature bias.
3. The proposed PIT had better performance improvement in Market-1501 [17], Duke-MTMC [18], Occluded-Duke [19] and MSMT17 [20].

2. Related Work

2.1. Auxiliary Feature Representation Learning

The information interference is the most challenging problem in person re-ID [16]. Auxiliary feature representations have been adopted to cultivate consistent person re-ID to deal this problem. Some common methods are adopted to enhance auxiliary feature representation such as additional annotation information [13,14,21–23], timing information [24,25] and generated/augmented training samples [26]. Furthermore, some studies have fused representations from multiple levels [4,27,28], requiring extra methods, such as body-part detection [4,27] or attention mechanisms [28]. However, while most of these methods have been designed based on CNNs, the slicing operation in transformer cuts the image into independent patches, making most of these methods unsuitable for transformer for direct application [15]. Thus, an auxiliary feature representation method of the transformer for person re-ID is needed.

2.2. Visual Transformer

Transformer [9] was originally proposed and applied in the field of natural language processing (NLP) and has gradually become the mainstream method for NLP tasks. Recently, transformer has been shown to outperform traditional methods in many vision tasks, such as object detection [29], generative adversarial networks [30], action recognition [31]. Visual transformers (ViT) [8] are the first to transformers to be applied to image vision. They first divide the input image into image patches, then use linear projection to map it to a 1D vector and add learnable ‘class labels (CLS)’, and finally pass the merged 1D vector to the transformer encoder. TransReID [32] is the first application of ViT in the field of object re-ID. The patch shuffle operation and side information module were designed to verify the advantages of the transformer on the re-ID task. PAT [27] integrates transformer architectures into CNNs. PAT [27] designs part diversity part discriminability to diversify part discovery. However, it failed solve the problem of fine-grained feature loss caused by transformer dicing operations. Thus, it is still an urgent problem to obtain high correlations from scattered patch blocks in transformer. This paper aims to propose a transformer-based patch information supplementation method to address the problems in existing transformer fine-grained feature extraction schemes.

3. Methodology

In this section, we will present the designed techniques to improve the transformer baseline. As demonstrated in Figure 2, the proposed PIT consists of the baseline, a patch feature extract network PPN, and the semantic information encoder IDE. The first part is the baseline, we use the ViT model as the baseline for person re-identification task. The details will be introduced in Section 3.1. The second part is the PPN, which obtains fine-grained features that are invariant to global disturbances by solving the local feature information loss caused by dicing operations. The details of which are introduced in Section 3.2. The third part is the IDE, which aims to avoid interfering with the generation

of information encoding. The same person should have the same identity information encoding. The details will be introduced in Section 3.3.

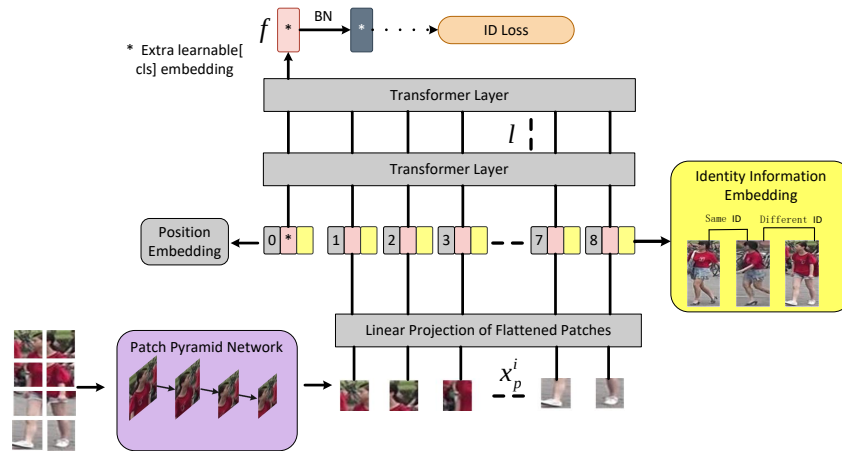


Figure 2. Framework of the proposed PIT. The PPN (light purple) cuts the image into patches and passes it into the encoder. The identity information is encoded as an embedded representation by the IDE (light yellow). The transformer encoder uniformly encodes the identity information embedding, patch embeddings, and position embeddings.

3.1. Baseline

As shown in Figure 3, input an image $x \in R^{C,H,W}$, while C , H , and W represent the channel size, height and width, respectively. We divide it into N fixed-size patches $\{x_p^i \mid i = 1, 2, \dots, N\}$ through the slicing operation. These patches are flattened into one-dimensional tensors and then projected onto a lower-dimensional space D using a linear transformation. The resulting feature vectors for each patch are used as inputs for the subsequent layers. In addition, a learnable embedding token denoted as x_{cls} is pre-loaded into the input sequence, with its output serving as the global feature denoted by f . Learnable position embeddings are also incorporated to capture spatial information. The input sequence to the transformer layers can be represented as:

$$Z_0 = [x_{cls}; F(x_p^1); F(x_p^2); \dots; F(x_p^N)] + P \quad (1)$$

where Z_0 is the input sequence embeddings of the baseline, and $P \in R^{(N+1) \times D}$ is the location information embedding. F is the linear mapping of the diced patches x_p^i to D dimensions. Additionally, l transformer layers is used to learn feature representations.

Patch Slicing. Patch splitting refers to the process of dividing an input image into several small, equally-sized blocks called “patches”. We use a sliding window approach to divide an image of size $H \times W$ into N fixed-size patches. The stride is denoted as S , and the patch size is denoted as P .

$$N = N_H \times N_W = \left\lfloor \frac{H + S - P}{S} \right\rfloor \times \left\lfloor \frac{W + S - P}{S} \right\rfloor \quad (2)$$

where $\lfloor \cdot \rfloor$ is the floor function and S is set same to P . N_H and N_W represent the number of splitting patches in height and width, respectively.

Position Embeddings. The image resolution of person re-ID is different from the original resolution of ViT images. The pre-trained position embeddings in ImageNet cannot be directly loaded into the baseline. Therefore, we introduce bilinear 2D interpolation to handle the input image resolution.

Supervised Learning. We use ID loss to optimize the baseline. ID loss is cross-entropy loss without label smoothing and can be expressed as:

$$L(y, \hat{y}) = \sum_{i=1}^M -q_i \log(p_i) \quad \begin{cases} q_i=0, y \neq i \\ q_i=1, y=i \end{cases} \quad (3)$$

where M represents the total number of people, y represents the truth ID label, q_i represents the target probability, and p_i represents the ID prediction logits of class i . The smaller the cross-entropy loss, the closer the predicted result is to the true result.

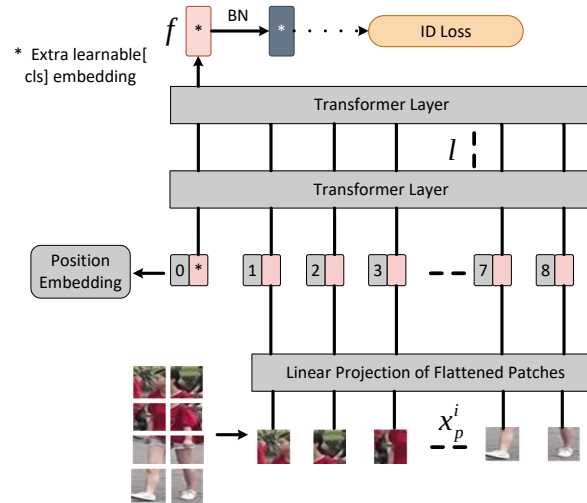


Figure 3. Framework of the baseline. The output [cls] token marked with * serves as the global feature f . Inspired by [7], we introduce the BNNeck after f .

3.2. Patch Pyramid Network

Although encoding the transformer provides a larger receptive field, the dicing operation reduces the correlation of the patch in the spatial dimension, leading to an inability to effectively extract fine-grained features.

As shown in Figure 4, the proposed PPN is a multi-scale patch fine-grained feature extraction network. Given an image $x \in R^{C,H,W}$, C , H , and W represent the channel size, height and width, respectively. Then, we split it into N fixed-sized patches $\{x_p^i \mid i = 1, 2, \dots, N\}$ by PPN. We explain the details of PPN below.

Based on convolution operations with four different kernel sizes (2×2 , 4×4 , 8×8 , 16×16), x is divided into various patch sizes $\{P_0, P_1, P_2, P_3\} \in R^{C,H,W}$, which only alter the size of the patches. The size of H , and W among the patches are expressed as

$$\begin{aligned} H_3 &= 2H_2 = 8H_1 = 16H_0 \\ W_3 &= 2W_2 = 8W_1 = 16W_0 \end{aligned} \quad (4)$$

Then, P_0, P_1, P_2 , and P_3 are arranged in a pyramid structure, where the top of the pyramid is the largest patch scale P_3 and the bottom is the smallest P_0 . Starting from the top of the pyramid, the downsampling operator F_{2D} is applied to scale the patches. After downsampling, adjacent patches are summed together. The above two operations are repeated until the patches reach the smallest scale, which is the target patch. Accordingly, the object patches $x_p^i \in R^{C,H,W}$ can be obtained. Moreover, the proposed feature extraction module F_E is used before each downsampling operation to obtain $\{P_1'', P_2'', P_3''\} \in R^{C,H,W}$. It is worth noting that the minimum size of a patch does not perform feature extraction, having an impact on model accuracy due to the shallow feature extraction operation. Finally, the output x_p^i can be deduced, given by

$$\begin{aligned}
P_3'' &= F_E(P_3) \\
P_2'' &= F_E(F_{2D}(P_3'') + P_2) \\
P_1'' &= F_E(F_{2D}(P_2'') + P_1) \\
x_p^i &= F_{2D}(P_1'') + P_0
\end{aligned} \tag{5}$$

where F_{2D} represents the downsampling operation with a convolution kernel of 3×3 .

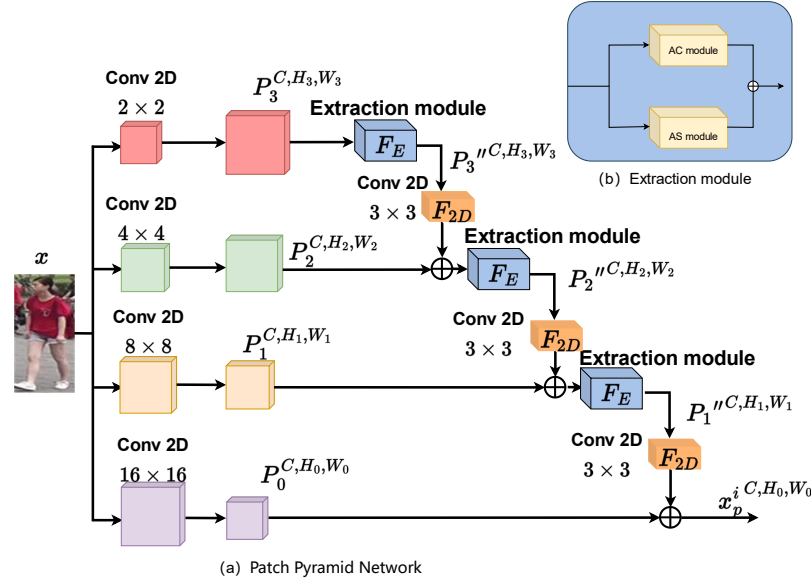


Figure 4. (a) Framework of the patch pyramid network; (b) framework of the extraction module.

As shown in Figure 5, the extraction module consists of four parts: Input $P \in R^{C,H,W}$, AC module, AS module and output $P'' \in R^{C,H,W}$. The AC module is an attention mechanism module in the channel dimension, and the AS module is an attention mechanism module in the spatial dimension. The two modules work in parallel in the extraction module by being connected together. They, respectively, extract channel features and spatial features in the patch block, and then add them together to obtain the final fine-grained feature.

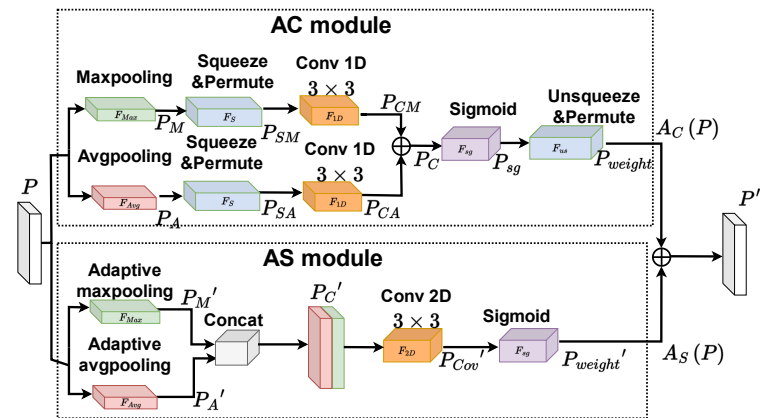


Figure 5. Extraction module.

AC module. In the channel dimension, the AC module is an adaptive feature extraction module. Two 2D maps $P_M \in R^{C,1,1}$ and $P_A \in R^{C,1,1}$ can be generated by the parallel maximum-pooling F_{Max} and average-pooling F_{Avg} , which resize the input P from C, H, W to $C, 1, 1$. Then, in order to avoid overcomplicating the model, the obtained 2D feature maps are processed by the squeeze and permute operation F_S to obtain the $P_{SM} \in R^{1,C}$ and $P_{SA} \in R^{1,C}$. After this, the 1D convolution module F_{1D} with a kernel size of 3×3 is

used to process P_{SM} and P_{SA} separately. The local cross-channel interaction can be realized, leading to $P_{CM} \in R^{1,C}$ and $P_{CA} \in R^{1,C}$. Then, both are added together to obtain $P_C \in R^{1,C}$. The $P_{sg} \in R^{1,C}$ is acquired by applying a sigmoid activation function F_{sg} to P_C . Finally, the weight of the channel dimension $P_{weight} \in R^{C,1,1}$ is obtained after a dimensional upgrade operation F_{us} . The operation of the AC module can be deduced by

$$P_{weight} = F_{us} [F_{sg} [F_{1D}(F_S(F_{Max}(p))) + F_{1D}(F_S(F_{Avg}(P)))]]) \quad (6)$$

The adaptive channel feature $A_C(P)$ can be acquired by multiplying P_{weight} by the input patch P , giving

$$A_C(P) = P_{weight} \otimes P \quad (7)$$

AS module. The AS module is an adaptive feature extraction module in the spatial dimension. Two 2D maps $P_M' \in R^{1,H,W}$ and $P_A' \in R^{1,H,W}$ can be generated through the adaptive maximum-pooling F_M and adaptive average-pooling F_A , producing two $1 \times H \times W$ feature maps from the input P . After this, the $P_C' \in R^{2,H,W}$ is obtained by concatenating P_M and P_A through a concatenation operation $\oplus$. Then, $P_{Cov}' \in R^{1,H,W}$ is obtained by a convolution F_{2D} with a kernel size of 3×3 that turns the number of channels in P_C to 1. The adaptive weight P_{weight}' of the spatial dimension is acquired by applying a sigmoid activation function F_{sg} to the P_{Cov}' . Accordingly, the operation of the AS module can be computed as

$$P_{weight}' = F_{sg} [F_{2D}(F_M(P) \oplus F_A(P))] \quad (8)$$

Similarly, the adaptive spatial feature $A_S(P)$ can be defined as

$$A_S(P) = P_{weight}' \otimes P \quad (9)$$

Based on the AC and AS operation, the output P'' of the extraction module can be computed as

$$P'' = w_1 * A_C(P) + w_2 * A_S(P) \quad (10)$$

where w_1 and w_2 are the weights that change with the gradient whose sum is 1.

3.3. Identity Information Embedding Module

Despite fine-grained feature being obtained, the impact of clothing changes cannot be ignored. In other words, the model may not be able to discriminate different objects from the same angle due to the bias of clothing information. Therefore, the proposed IDE incorporates the identity information of people into the embedding representation to obtain robust features.

Similar to positional embeddings employing learnable embeddings to encode positional information, we insert learnable one-dimensional embeddings to preserve identity information. In particular, as shown in Figure 1, we insert the identity information embedding into the transformer encoder with the patch and position embeddings. Specifically, assuming that there is a total of M person IDs, we initialize the learnable identity information embedding as $G_{ID} \in R^{M \times D}$. If the ID of a person is k , the identity information embedding can be expressed as $G_{ID}(k)$. Unlike positional embeddings which vary between patches, identity information embeddings $G_{ID}(k)$ are the same for all patches of an image.

As the identity information embedding, patch embedding, and position embedding are all linearly mapped to a D -dimensional space, their input sequences can be directly added for information integration. The IDE can be written as

$$Z_0' = Z_0 + \lambda G_{ID}(k) \quad (11)$$

where Z_0' is the input sequence after adding the identity information embedding, Z_0 is the original input sequence in the baseline, and λ is the parameter of the identity information embedding. The transformer encoding layer can encode embeddings of different distribution types and add these features directly.

4. Experiments

4.1. Datasets

In order to demonstrate the effectiveness of our methods, the experiments were conducted on the three widely used holistic re-ID datasets (Market-1501 [17], Duke-MTMC [18], MSMT17 [20]), and one occluded re-ID dataset (Occluded-Duke [19]). Table 1 provides a brief explanation of the datasets, explained in detail as follows.

Table 1. Dataset statistics used in this article

Dataset	#ID	#Cam	#Image	#Query	#Gallery
Market-1501	1501	6	32,668	3368	15,913
Duke-MTMC	1812	8	36,411	2228	17,661
MSMT17	4101	15	126,441	11,659	82,161
Occluded-Duke	1812	8	15,618	2210	17,661

Market-1501 [17] has a total of 32,668 images, consisting of 1501 identities under 6 cameras. It is split into a training set of 12,936 images of 751 identities and a test set of 19,732 images of 750 identities.

Duke-MTMC [18] has a total of 36,411 images from 1812 people captured by 8 cameras. The 16,522 images of 702 people are randomly selected as the training set, the 2228 remaining images are used as query images and the 17,661 gallery images are used as a test set.

MSMT17 [20] is the largest dataset with 4101 IDs, with 126,441 images captured from 15 cameras. The training set has 32,621 images of 1041 identities, and the test set has 93,820 images of 3060 identities. In the test set, the 11,659 images are randomly selected as query images.

Occluded-Duke [19] is a split of Duke-MTMC [18] that preserves occluded images and removes overlapping images. It contains 15,618 training images, 17,661 gallery images, and 2210 occlusion query images.

4.2. Experimental Strategy and Experimental Environment

The size of all images were adjusted to 256×128 for training and testing. As for the data augmentation, we employed random horizontal flipping, padding, random cropping and random erasing. The SGD optimizer was adopted with a momentum of 0.9 and a weight decay set as 1×10^{-4} . The initial learning rate and the batch size are set to 0.008 and 64, respectively.

Limited by experimental equipment, all experiments used a Nvidia TITANX GPU, which were trained on the PyTorch platform with PyTorch = 1.7.0, torchvision = 0.8.0, CUDA = 11.1 and CUDNN = 8.0.

Evaluation Protocols. Cumulative matching characteristic (CMC) curves and mean average precision (mAP) were adopted to evaluate the quality of different the re-ID models. The re-ranking was not taken to refine the matching results further in the experiments that are performed in a single query mode.

4.3. Results of Backbones

In this section, we compare the performances of the backbones. We chose several backbones, including SqueezeNet, ShuffleNet, DenseNet, ResNet50, ResNet101, SEResNet50, SEResNet101 and VIT-B/16, where VIT-B/16 represents the VIT baseline with patch size 16.

As shown in Table 2, there is a big gap in performance between the VIT-B/16- and CNN-based backbones. On the re-ID benchmark, the VIT-B/16 greatly outperforms the

CNN-based backbones, the ViT-B/16 achieved at least 11.6% mAP and 8.2% rank-1 over SEResNet101. This suggests that the multi-head attention mechanism of transformer can obtain a larger receptive field. Thus, we choose the ViT-B/16 as the baseline.

Table 2. Comparison of different backbones. ViT-B/16 is the baseline model of this paper, referred to as baseline for short.

Backbone	Market-1501		Duke-MTMC		Occluded-Duke	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
SqueezeNet	30.7	49.1	21	36.3	6.7	11.2
ShuffleNet	34.1	56.9	25.2	41.1	9.5	14.8
DenseNet	62.5	83.9	61.6	73.6	28.8	42.1
ResNet50	67.9	84.5	54.3	73	31.1	42.7
SEResNet101	68.7	84.9	59.7	76.8	38.4	41
SEResNet50	73.7	88.4	60.1	77.6	32.7	42.8
SEResNet101	75.1	89.3	63	80.2	35.9	45.3
ViT-B/16	86.7	94.5	79.4	88.4	52.6	60.2

4.4. Ablation Study of PPN

As shown in Table 3, we evaluate the superiority of our proposed PPN. Compared to the baseline, the PPN improves by +1.6% mAP and +1.1% rank-1. Increasing number of P in the PPN, the performance of the PPN gradually increases, suggesting that adding larger patches to the PPN is helpful to extract spatial features in the dicing stage. Furthermore, by comparing the PPN and PPN (without the extraction module), we can find that the extraction module operation can help the model to improve by +0.8% mAP and +0.6% rank-1, suggesting that the extraction module in our PPN helps the model to effectively obtain fine-grained features. The visual attention map in Figure 6 shows that the model obtains more context-aware information and better fine-grained features through the PPN operation, improving the anti-interference ability of the model.

Table 3. The ablation study of patch pyramid network on the MSMT17.

Backbone	P_0	P_1	P_2	P_3	Extraction Module	mAP	Rank-1	Rank-5	Rank-10
Baseline	✓					60.9	81	90.2	92.9
+PPN	✓	✓			✓	61.2	81.2	90.2	92.9
+PPN	✓	✓	✓		✓	61.9	81.8	90.3	93.0
+PPN	✓	✓	✓	✓		61.7	81.5	90.2	92.9
+PPN	✓	✓	✓	✓	✓	62.5	82.1	90.5	93.1

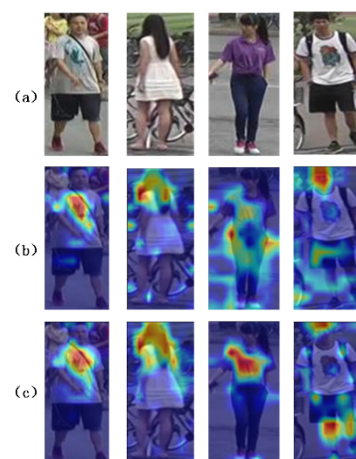


Figure 6. Grad-CAM visualization of attention maps. (a) Input images, (b) baseline, (c) PPN.

4.5. Ablation Study of IDE

Performance Analysis. As shown in Table 4, we assess the utility of the proposed IDE. The experimental results show that our method achieves 1.7% mAP and 0.8% rank-1 improvement with slight improvements of the both rank-5 and rank-10 at extremely high performances, showing that the proposed IDE can work well on the baseline. The IDE can separate invariant ID features from semantic features, significant for person re-ID to eliminate interference.

Table 4. Ablation study of IDE on the Market-1501.

Method	mAP	Rank-1	Rank-5	Rank-10
Baseline	86.7	94.5	98.2	99.0
Baseline + IDE	88.4	95.3	98.5	99.1

Ablation study of λ . As demonstrated in Figure 7, we analysed the effect of parameter λ on the IDE. Particularly, the baseline reached 86.7% mAP and 94.5% rank-1 with λ set as 0. When λ increases, the performance of the IDE improves by 88.4% mAP and 95.3% rank-1 ($\lambda = 2.0$), indicating that our methods can learn invariant features well. With a further increase in λ , the feature embedding and position embedding gradually decrease to set the performance of the system.

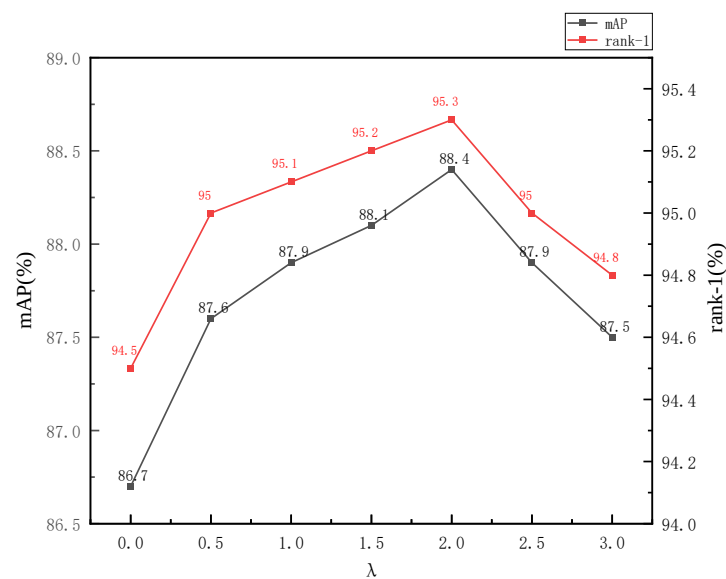


Figure 7. Parameter analysis of λ on Market-1501.

4.6. Ablation Study of PIT

As shown in Table 5, compared to the baseline, our proposed PPN and IDE improve the performance by +1.6% mAP / +1.1% rank-1 and +1.9% mAP / +0.9% rank-1, respectively. The ablation studies illustrate that the proposed PIT achieves a better performance of 63.1% (+2.2%) mAP and 82.3% (+1.3%) rank-1 than the baseline. The experimental results demonstrate that the PPN and IDE complement each other in person re-ID.

Table 5. The ablation study of PIT on the MSMT17.

Backbone	PPN	IDE	mAP	Rank-1	Rank-5	Rank-10
Baseline	×	×	60.9	81	90.2	92.9
	✓	×	62.5	82.1	90.5	93.1
	×	✓	62.8	81.9	90.4	92.9
PIT	✓	✓	63.1	82.3	90.8	93.1

4.7. Comparison with State-of-the-Art Approaches

As shown in Table 6, we compare our model with the state-of-the-art methods on three holistic benchmarks and one occluded benchmark to show the effectiveness of the proposed PIT. In addition, we also compare the results from the baseline.

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

Methods	Market-1501		Duke-MTMC		MSMT17		Occluded-Duke	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
PCB(2018) [4]	81.6	93.8	69.2	83.3	-	-	33.7	42.6
AlignedReID (2017) [12]	79.3	91.2	-	-	-	-	37.3	51.4
PGFA (2019) [19]	76.8	91.2	65.5	82.6	-	-	43.8	55.1
HONet(2020) [33]	84.9	94.2	75.6	86.9	-	-	52.3	62.8
ICE (2021) [34]	86.6	95.1	76.5	88.2	50.4	76.4	-	-
Transreid (2021) [32]	88.0	94.7	81.2	90.1	63.9	82.7	55.6	62.8
PAT (2021) [27]	88.0	95.4	78.2	88.8	-	-	53.6	64.5
AAformer (2021) [35]	87.7	95.4	80.0	90.1	63.2	83.6	58.2	67.0
DRL-Net (2022) [36]	86.9	94.7	76.6	88.1	55.3	78.4	50.8	65.0
SpCL (2022) [37]	76.7	90.3	68.8	82.9	26.8	53.7	-	-
FED (2022) [38]	85.9	94.8	78.0	89.4	-	-	55.8	67.4
ISE (2022) [39]	87.8	95.6	-	-	51.0	76.8	-	-
Baseline	86.7	94.5	79.4	88.4	60.9	81	52.6	60.2
Ours	88.8	95.6	81.3	90.4	63.1	82.3	54.8	61.7

Market1501. Compared with the state-of-the-art methods, the proposed PIT acquires outstanding performances. Specifically, when compared with other transformer-based networks [17,18,33,38], our method receives the best result of 88.8% (+0.8%) mAP. There is little difference in the results between these methods because the exploration of Market1501 is saturated.

Duke-MTMC. Although the proposed PIT performs close to the other methods, our model still derives first-rate performances, showing the robustness of the proposed methods. Furthermore, compared to the baseline, our method exhibits better results (+1.9% mAP/+2.0% rank-1), which validates our work on the Transformer baseline.

MSMT17. Being the largest person re-ID dataset means that robust feature extraction in complex environments is more difficult. Due to the use of a greater number of cameras to capture people in the MSMT17 dataset, there is a greater diversity of clothing and posture changes among the individuals, leading to a suboptimal recognition performance of the models. However, the proposed PIT proved its reliability with good experimental results, outperforming the ISE [39] when integrating identity information.

Occluded re-ID. From the experimental results, the introduced PIT still achieves a good score of 54.8% mAP in the face of severe occlusion and fewer features. This is because occlusions generate interfering features, resulting in a suboptimal recognition performance of the models. Additionally, compared with the PAT [27] with aligned body parts, our method achieved better results (+1.2% in front of PAT [27]) without aligning body parts.

4.8. The Matching Visualization Results

We compare the ranking results of the baseline and PIT in Figure 8, and the results indicate that PIT can effectively address the recognition inaccuracy caused by similar clothing, human pose variations, and partial occlusions. The identity information provides useful assistance to the model.

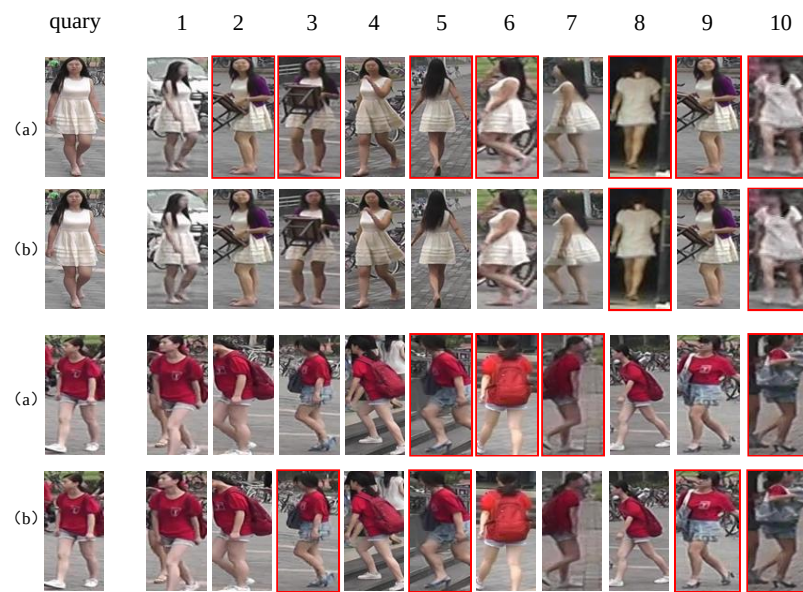


Figure 8. The matching visualization results. (a) Baseline; (b) PIT.

5. Conclusions

In this paper, we proposed a patch information supplement method (PIT) that can handle local information loss. By using the introduced PPN, PIT can solve decreasing spatial correlations caused by dicing operations and obtain the fine-grained features. Furthermore, we designed the IDE to eliminate the influence of clothing changes to separate ID-related features from the semantic features. Extensive experimental results for re-ID verify the superiority of PIT and the effectiveness of its components.

However, our model does not perform optimally on occluded re-ID data. It cannot offer a high generalization performance in a randomly occluded environment. We will concentrate on how to better utilize structural information to handle the interference of occlusions in future research.

Author Contributions: L.Z., C.J. and M.W. conceived the experiments, C.J. conducted the experiments. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Research and Development Plan of Hubei Province, China (2021BGD013); the Knowledge Innovation Special Project of Wuhan Science and Technology Bureau, China (2022010801010255); Natural Science Foundation of Hubei Province, China (2022CFA007).

Data Availability Statement: Code and data are available.

Conflicts of Interest: This authors have no conflict of interest.

References

1. Zheng, W.S.; Gong, S.; Xiang, T. Associating Groups of People. In Proceedings of the British Machine Vision Conference, London, UK, 7–10 September 2009. [[CrossRef](#)]
2. Hirzer, M.; Beleznaï, C.; Roth, P.M.; Bischof, H. Person re-identification by descriptive and discriminative classification. In Proceedings of the Image Analysis: 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, 9 May 2011; pp. 91–102.
3. Khorramshahi, P.; Peri, N.; Chen, J.C.; Chellappa, R. The devil is in the details: Self-supervised attention for vehicle re-identification. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 369–386.
4. 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.
5. Zhang, Z.; Lan, C.; Zeng, W.; Jin, X.; Chen, Z. Relation-aware global attention for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 3186–3195.
6. Luo, W.; Li, Y.; Urtasun, R.; Zeme, T. Understanding the effective receptive field in deep convolutional neural networks. *arXiv* **2016**, arXiv:1701.04128.

7. 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.
8. 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.
9. Wang, C.; Zhang, Q.; Huang, C.; Liu, W.; Wang, X. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 365–381.
10. Yao, H.; Zhang, S.; Hong, R.; Zhang, Y.; Xu, C.; Tian, Q. Deep representation learning with part loss for person re-identification. *IEEE Trans. Image Process.* **2019**, *28*, 2860–2871. [[CrossRef](#)] [[PubMed](#)]
11. 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](#)]
12. Zhang, X.; Luo, H.; Fan, X.; Xiang, W.; Sun, Y.; Xiao, W.; Jiang, W.; Zhang, C.; Sun, J. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv* **2017**, arXiv:1711.08184.
13. 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 European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 140–157.
14. Sun, X.; Zheng, L. Dissecting person re-identification from the viewpoint of viewpoint. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 608–617.
15. Wang, G.; Lai, J.; Huang, P.; Xie, X. Spatial-temporal person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8933–8940.
16. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 2872–2893. [[CrossRef](#)] [[PubMed](#)]
17. 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.
18. 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; 15–16 October 2016; pp. 17–35.
19. 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.
20. 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.
21. Chang, X.; Hospedales, T.M.; Xiang, T. Multi-level factorisation net 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. 2109–2118.
22. Lin, J.; Ren, L.; Lu, J.; Feng, J.; Zhou, J. Consistent-aware deep learning for person re-identification in a camera network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5771–5780.
23. Sarfraz, M.S.; Schumann, A.; Eberle, A.; Stiefel, R. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5771–5780.
24. Liu, Y.; Li, D.; Wan, S.; Wang, F.; Dou, W.; Xu, X.; Li, S.; Ma, R.; Qi, L. A long short-term memory-based model for greenhouse climate prediction. *Int. J. Intell. Syst.* **2022**, *37*, 135–151. [[CrossRef](#)]
25. Liu, Y.; Song, Z.; Xu, X.; Rafique, W.; Zhang, X.; Shen, J. Bidirectional GRU networks-based next POI category prediction for healthcare. *Int. J. Intell. Syst.* **2022**, *37*, 4020–4040. [[CrossRef](#)]
26. 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; pp. 13001–13008.
27. Li, Y.; He, J.; Zhang, T.; Liu, X.; Zhang, Y.; Wu, F. Diverse part discovery: Occluded person re-identification with part-aware transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2898–2907.
28. Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; Wang, X. Hydraplus-net: Attentive deep features for pedestrian analysis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 350–359.
29. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on COMPUTER Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
30. Jiang, Y.; Chang, S.; Wang, Z. Transgan: Two transformers can make one strong gan. *arXiv* **2021**, arXiv:2102.07074.
31. Li, X.; Hou, Y.; Wang, P.; Gao, Z.; Xu, M.; Li, W. Trear: Transformer-based rgb-d egocentric action recognition. *IEEE Trans. Cogn. Dev. Syst.* **2021**, *14*, 246–252. [[CrossRef](#)]
32. 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, BC, Canada, 11–17 October 2021; pp. 15013–15022.

33. Wang, G.A.; 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.
34. Chen, H.; Lagadec, B.; Bremond, F. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 14960–14969.
35. Zhu, K.; Guo, H.; Zhang, S.; Wang, Y.; Huang, G.; Qiao, H.; Liu, J.; Wang, J.; Tang, M. Aaformer: Auto-aligned transformer for person re-identification. *arXiv* **2021**, arXiv:2104.00921 .
36. Jia, M.; Cheng, X.; Lu, S.; Zhang, J. Learning disentangled representation implicitly via transformer for occluded person re-identification. *arXiv* **2022**, arXiv:2107.02380.
37. Ge, Y.; Zhu, F.; Chen, D.; Zhao, R. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 11309–11321.
38. Wang, Z.; Zhu, F.; Tang, S.; Zhao, R.; He, L.; Song, J. Feature Erasing and Diffusion Network for Occluded Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4754–4763.
39. Zhang, X.; Li, D.; Wang, Z.; Wang, J.; Ding, E.; Shi, J.Q.; Zhang, Z.; Wang, J. Implicit Sample Extension for Unsupervised Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7369–7378.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.