Part-Aware Self-Supervised Pre-Training for Person Re-Identification

Kuan Zhu[1,2], Haiyun Guo[1], Tianyi Yan[1,2], Yousong Zhu[1]
Jinqiao Wang[1,2], and Ming Tang[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
{kuan.zhu, haiyun.guo, tianyi.yan, yousong.zhu, jqwang,
tangm}@nlpr.ia.ac.cn

Abstract. In person re-identification (ReID), very recent researches have validated pre-training the models on unlabelled person images is much better than on ImageNet. However, these researches directly apply the existing self-supervised learning (SSL) methods designed for image classification to ReID without any adaption in the framework. These SSL methods match the outputs of *local* views (e.g., red T-shirt, blue shorts) to those of the *global* views at the same time, losing lots of details. In this paper, we propose a ReID-specific pre-training method, Part-Aware Self-Supervised pre-training (PASS), which can generate part-level features to offer fine-grained information and is more suitable for ReID. PASS divides the images into several local areas, and the *local* views randomly cropped from each area are assigned with a specific learnable [PART] token. On the other hand, the [PART]s of all local areas are also appended to the *global* views. PASS learns to match the outputs of the *local* views and *global* views on the same [PART]. That is, the learned [PART] of the *local* views from a local area is only matched with the corresponding [PART] learned from the *global* views. As a result, each [PART] can focus on a specific local area of the image and extracts fine-grained information of this area. Experiments show PASS sets the new state-of-the-art performances on Market1501 and MSMT17 on various ReID tasks, e.g., vanilla ViT-S/16 pre-trained by PASS achieves 92.2%/90.2%/88.5% mAP accuracy on Market1501 for supervised/UDA/USL ReID. Our codes are available at https://github.com/CASIA-IVA-Lab/PASS-reID

Keywords: person re-identification, self-supervised pre-training, local representations

1 Introduction

Person re-identification (ReID) aims to associate the person images captured by different cameras. Limited by the scale of the labeled ReID datasets, most existing methods first pre-train the backbone networks (e.g., ResNet [16], ViT [10]) on ImageNet [8] and then fine-tune them on person ReID datasets to boost
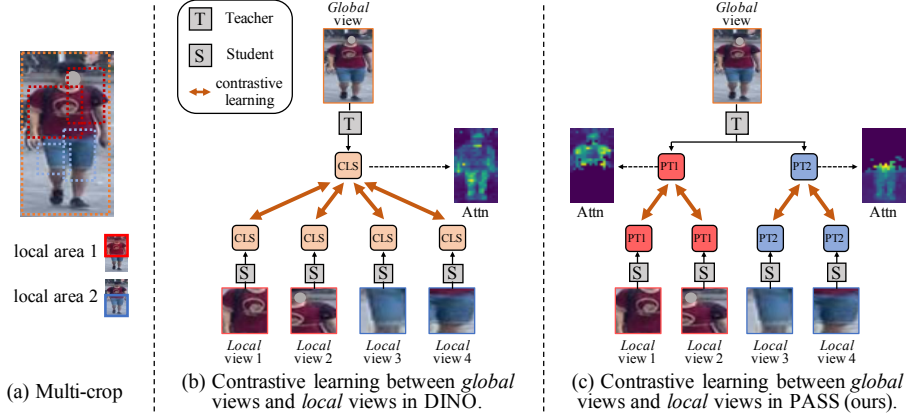
Fig. 1. (a) Multi-crop [1] is a data augmentation strategy which uses a mix of views with different resolutions. (b) DINO [2] learns to match the outputs of all the *local* views and *global* views, losing lots of details. (c) In PASS, the *local* views cropped from each local area are assigned with a specific [PART], e.g., the *local* views cropped from the upper local area, *local* view 1 and 2, are assigned with [PART]$_1$; the *local* views cropped from the lower local area, *local* view 3 and 4, are assigned with [PART]$_2$. The *global* views are appended with all the [PART]s. PASS learns to match the predictions on the same [PART] (i.e., [PART]$_1$ or [PART]$_2$) of *local* views and *global* views, **respectively**

the performance [24, 21, 17]. However, it is arguable whether using ImageNet for pre-training is optimal as there exist large domain gaps between ImageNet and person ReID data, e.g., 1) ImageNet-1K contains a thousand kinds of objects while ReID datasets only contain persons; 2) the model pre-trained on ImageNet will focus on category-level differences, losing lots of rich visual information. Therefore, the fine-grained identity information, which is preferred by ReID, can not be provided by pre-training on ImageNet [11, 22].

To bridge the gap between pre-training and fine-tuning datasets for better ReID models, Fu et al. [11] propose the first large scale unlabeled person ReID dataset "LUPerson" and demonstrate that unsupervised pre-training the models on LUPerson is quite better than supervised pre-training on ImageNet. Luo et al. [22] further investigate that DINO [2] algorithm with Transformer architecture (ViT [10]) obtains the best ReID performance among the existing self-supervised learning (SSL) methods and network architectures. However, these works [11, 22] directly apply the existing SSL methods, that are proposed for image classification, to ReID and do not make any adaption in the SSL framework. Another gap will appear when employing these SSL methods to ReID, which is shown in Figure 1(b). DINO matches the outputs of all the *local* views with the *global* views in the same feature space. It is unreasonable to make the outputs of red T-shirt and blue shorts both match the output of the global image at the same time. To compromise on this matching, only the shared features of different views are retrained and lots of local detailed information are removed in the learned

features. However, as fine-grained information has been verified to be crucial to describe a person image [24, 27, 20], it is necessary to design a pre-training method that can extract as many fine-grained clues as possible for ReID.

In this paper, we propose a ReID-specific pre-training method based on Transformer, Part-Aware Self-Supervised pre-training (PASS), by which part-level features can be automatically extracted to offer fine-grained information for person ReID. PASS uses the learning paradigm of knowledge distillation to match the outputs of the teacher network and student network. It first divides the image into several fixed overlapping local areas and randomly crops *local* views from these local areas. The *global* views are randomly cropped from the whole image with higher resolution. All views are passed through the student while only *global* views are passed through the teacher. For simplicity, we only illustrate the comparison between the outputs of *local* views passed through student and *global* views passed through teacher in Figure 1(c). Before passing through the student, the *local* views cropped from each local area are assigned with a specific learnable [PART] token, which is used to learn the local representation. All these [PART]s are also appended to the *global* views and fed to the teacher to learn local features from the whole image. PASS learns to match the corresponding [PART]s of *local* views and *global* views, where the [PART]s of different areas are not compared. Take Figure 1(c) as an example, the *local* views cropped from the upper local area, *local* views 1 and 2, are assigned with [PART]$_1$; the *local* views cropped from the lower local area, *local* views 3 and 4, are assigned with [PART]$_2$. The predictions on the [PART]$_1$s of *local* views 1 and 2 are only compared with those on the [PART]$_1$s of the *global* views and so does the [PART]$_2$s.

In the student, the *local* views assigned to each [PART] are cropped from a specific local area, thus the [PART]s can focus on different areas. PASS uses the student to update the teacher, which can guarantee [PART]s in teacher also focus on different local areas and learns fine-grained information. In pre-training, the student learns to match the output of the teacher on the same [PART], which can guarantee each [PART] learns a robust local representation from the *local* views cropped in its corresponding local area. In fine-tuning, all [PART]s are appended to the input image and each [PART] automatically learns the local representation for a specific area.

We summarize the contributions of this work as: (i) In this paper, we propose the ReID-specific pre-training method, Part-Aware Self-Supervised pre-training (PASS), which is more suitable for ReID with part-level features offering fine-grained information. It is worth noting that we do not add any complex module to extract part-level features but only use several learnable tokens. (ii) The pre-trained ViT backbone can be fine-tuned on various ReID downstream tasks, i.e., supervised learning, unsupervised domain adaptation (UDA), and unsupervised learning (USL). PASS helps the ViTs set the new state-of-the-art performance on Market-1501 [33] and MSMT17 [28], e.g., vanilla ViT-S pre-trained by PASS achieves 92.2%/90.2%/88.5% Rank-1 and 69.1%/49.1%/41.0% mAP accuracy on Market1501 and MSMT17 for supervised/UDA/USL ReID, respectively.

2   Related Work

2.1   Self-Supervised Learning

Self-supervised learning (SSL) methods aim to learn discriminative representations from large scale unlabeled data. Recently, contrastive learning methods have made remarkable achievements [15, 4, 5, 14, 2, 1, 30], significantly reducing the gap with supervised pre-training. MoCo series [4, 5], which are developed from Momentum Contrast, treat the augmentations of a sample as positive pairs and all other samples as negative pairs. Ge et al. [14] propose a new paradigm, BYOL, where the online network predicts the output representation of the target network on the same image under a different augmented view, removing the need of negative pairs. DINO [2] further improves BYOL by using a centering and sharpening of the momentum teacher outputs to avoid model collapse. Besides, DINO adopts the augmentation strategy of multi-crop [1] to conduct the comparison between the representations of *global* views and *local* views. Xie et al. [30] propose MoBY which is a Transformer-specific method and combines MoCo with BYOL. Among these methods, DINO algorithm, plus the ViT architecture, can achieve the best performance on ReID [22]. Therefore, we aim to improve the framework of DINO to better adapt to the ReID task and obtain higher performance.

2.2   Person Re-Identification

Part-based person ReID. Employing part-level features for person image description can offer fine-grained information and has been verified as beneficial for person re-identification. Many efforts have been made to develop the part-based person re-ID to boost the state-of-the-art performance. PCB [24] first directly partitions the person images into fixed horizontal stripes and extract stripe-based part features. MGN [27] enhances the robustness by dividing images into stripes of different granularities and designs overlap between stripes. SPReID [20] uses a pre-trained human semantic parsing model to provide the mask of body parts to extract part features. ISP [34] proposes to automatically locate both human parts and non-human ones at pixel-level by iterative clustering. TransReID [18], which is a Transformer-specific method, also designs to extract the part-level features by re-arranging the patch embeddings and re-grouping them. All these methods show extracting part-level features to offer fine-grained information is of vital importance for person ReID.

Unsupervised pre-training for person ReID. Recently, more and more works find that there exists large domain gaps between ImageNet and person ReID data [11, 22], which makes pre-training ReID models on unlabelled person images is much better than pre-training on ImageNet. Fu et al. [11] first focus on this problem and propose a large scale unlabeled person ReID dataset "LUPerson". They also validate that unsupervised pre-training on LUPerson can improve the ReID performance compared with ImageNet-1k pre-training. Luo et

al. [22] further investigate several self-supervised learning methods and backbone networks. The results show that DINO [2] with ViT [10] obtains the best performance. [22] also proposes a data filtering method and IBN-based convolution stem for ViT architecture. However, apart from some adaptions in data augmentation or hyper-parameters, they do not specifically improve these pre-training methods for person ReID. The gaps will appear when applying the existing SSL methods proposed for image classification to ReID as shown in Fig. 1. In this paper, we propose a ReID-specific self-supervised pre-training method, PASS, which makes the model learn to extract part-level features in pre-training and automatically extract part-level features in fine-tuning.

3   Method

3.1   Preliminaries

Vision Transformer. We briefly describe the mechanism of the Vision Transformer (ViT) [10, 25] here, and please refer to [25] for details about Transformers and to [10] for its adaptation to images. The ViT architecture takes a grid of non-overlapping contiguous image patches of resolution $N \times N$ as input. Typically, we use $N = 16$ ("/16") in this paper. The patches are then mapped to a sequence of patch embeddings by a trainable linear projection. Some extra learnable tokens are appended to the sequence, e.g., class token [CLS] [10, 9] and part token [PART] [35]. The role of these tokens is to aggregate information from the patch sequence. [CLS] is proposed to learn a global representation for the input image and [PART] aims to extract part-level feature. There is no structural difference between [CLS] and [PART], while PASS will make them learn different-level features during training.

Student network & Teacher network. The framework used for this work, PASS, shares a similar overall structure as the popular self-supervised approach, DINO [2]. The overview of PASS is illustrated in Figure 2, where the learning paradigm of knowledge distillation is used. PASS contains a student network and a teacher network, and they share the same architecture. PASS trains the student f_{θ_s} to match the output of the teacher f_{θ_t}, which are parameterized by θ_s and θ_t respectively. Given an input image x, both networks predict probability distributions of K dimensions on the appended learnable tokens (i.e., [PART] and [CLS]) by projection heads. The probability P is obtained by normalizing the output of the network f with a softmax function. Specifically,

$$P_s^g(x)^{(t)} = \frac{\exp\left(f_{\theta_s}^{cls}(x)^{(t)}/\tau_s\right)}{\sum_{k=1}^{K} \exp\left(f_{\theta_s}^{cls}(x)^{(k)}/\tau_s\right)}, \tag{1}$$

where $f_{\theta_s}^{cls}(x)$ is the predicted distribution on [CLS], and t, k are the indexes of vector components. $\tau_s > 0$ is a temperature parameter that controls the
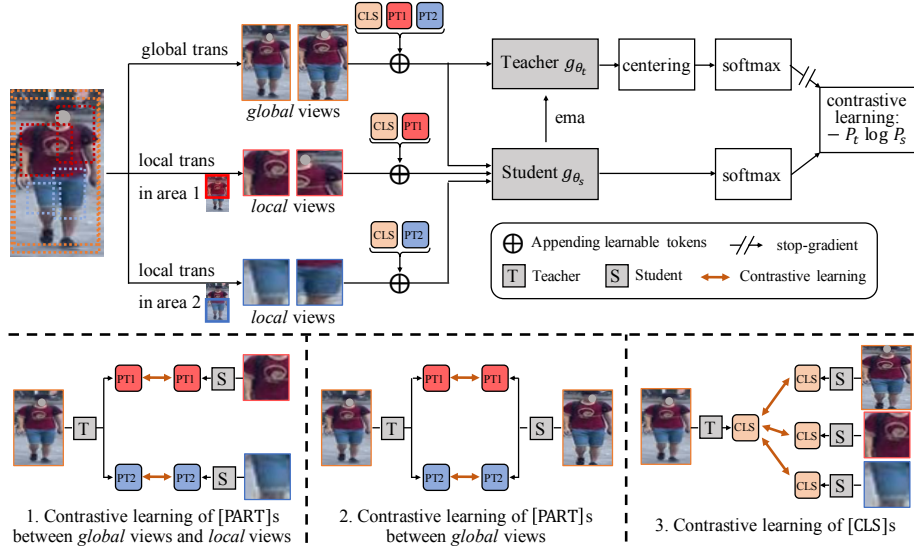
Fig. 2. The overview of PASS. We illustrate PASS in the case of $L = 2$ for simplicity. The global random transformation (global trans) crops out the *global* views from the whole images and the two local random transformations (local trans) crop out *local* views from two different local areas, respectively. The *local* views from each local area are assigned with a specific `[PART]`. The `[PART]`s of all local areas are also appended to the *global* views. `[CLS]` is also appended to all the views. In pre-training, all views are passed through the student while only the *global* views are passed through the teacher. Each network predicts several K dimensional features on `[CLS]` and `[PART]`s by the projection heads, and the features are normalized with a temperature softmax over the feature dimension. Then the similarities between the same `[PART]`/`[CLS]` output by student and teacher are measured with cross-entropy losses. We apply a stop-gradient operator on the teacher to propagate gradients only through the student. The teacher parameters are updated with an exponential moving average (ema) of the student parameters

sharpness of the output distribution. The similar formulas hold for P_t^g, $P_s^{l_i}$, $P_t^{l_i}$, where l_i means predicting on the ith `[PART]`.

The teacher f_{θ_t} is not pre-defined but built by means of the past iterations of the student f_{θ_s}. Following [2, 15], we freeze the teacher network over an epoch and use an exponential moving average (EMA) on the student weights. The update rule is:

$$\theta_t \leftarrow \lambda\theta_t + (1 - \lambda)\theta_s, \tag{2}$$

with λ following a cosine schedule from 0.996 to 1 during training. The output of the teacher network is centered with a mean calculated over the batch to avoid collapse [2].

3.2   Part-Aware Self-Supervised Pre-training

Given an input image x, PASS constructs a set of different views which includes different distorted views, or crops, of x. This set contains M *global* views, x_m^g, randomly cropped from the whole image, and $L \times J$ *local* views of smaller resolution randomly cropped from L local areas of the image, where the jth cropped view of the ith local area is denoted as $x_j^{l_i}, (i \in \{1, ..., L\}, j \in \{1, ..., J\})$.

All views are passed through the student while only the *global* views are passed through the teacher. For a *global* view, all learnable tokens, i.e., [CLS] and [PART]s, are appended to the sequence of patch embeddings. While for a *local* view cropped from the ith local area, only [CLS] and the ith part token [PART]$_i$ are appended to the patch embeddings. The projection heads are added to these extra learnable tokens in both teacher and student, and output the predicted probability distributions P. Specifically, given an input image x, its mth *global* view x_m^g is appended with [CLS] and all the [PART]s, and is passed through both the teacher and student. The predicted distributions by teacher and student are $P_t^g(x_m^g)$, $\{P_t^{l_i}(x_m^g)|i \in \{1, ..., L\}\}$ and $P_s^g(x_m^g)$, $\{P_s^{l_i}(x_m^g)|i \in \{1, ..., L\}\}$, respectively. Its *local* view $x_j^{l_i}$ is only appended with [CLS] and [PART]$_i$, and then is only passed through the student. The student predicts two probability distributions on [CLS] and [PART]$_i$: $P_s^g(x_j^{l_i})$ and $P_s^{l_i}(x_j^{l_i})$.

PASS learns to match the predictions of the same learnable tokens by cross-entropy loss. Specifically, **for** [PART]$_i$, the student outputs two types of predictions: (1) the prediction on [PART]$_i$ appended to $x_j^{l_i}$: $P_s^{l_i}(x_j^{l_i})$. (2) the prediction on [PART]$_i$ appended to x_m^g: $P_s^{l_i}(x_m^g)$. While the teacher only predicts the latter one and outputs $P_t^{l_i}(x_m^g)$. The student learns to match the output distributions of teacher by minimizing the cross-entropy loss w.r.t. the parameters of the student θ_s:

$$
\min_{\theta_s} \left\{ \sum_{m=1}^{M} \sum_{j=1}^{J} H\left(P_t^{l_i}\left(x_m^g\right), P_s^{l_i}\left(x_j^{l_i}\right)\right) + \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} H\left(P_t^{l_i}\left(x_{m_1}^g\right), P_s^{l_i}\left(x_{m_2}^g\right)\right) \right\}
\tag{3}
$$

where $m_1 \neq m_2$ and $H(a, b) = -a \log b$. The two terms are corresponding to the contrastive learning 1 and 2 in the second row of Figure 2. PASS applies this optimization to all the [PART]s. That is, any i in $\{1, ..., L\}$ is applicable in Eq. 3.

In the student, the *local* views assigned to each [PART] are randomly cropped from a specific local area, thus the [PART]s can focus on different areas. PASS uses the student to update the teacher, which can guarantee [PART]s in teacher also focus on different local areas and learns fine-grained information. In pre-training, the student learns to match the output of the teacher on the same [PART], which can guarantee each [PART] learns a robust local representation from the *local* views cropped in its corresponding local area.

For the [CLS], the predicted distributions from all the views are matched by:

$$\min_{\theta_s} \left\{ \sum_{m=1}^{M} \sum_{i=1}^{L} \sum_{j=1}^{J} H\left(P_t^g\left(x_m^g\right), P_s^g\left(x_j^{l_i}\right)\right) + \sum_{\substack{m_1 \\ =1}}^{M} \sum_{\substack{m_2 \\ =1}}^{M} H\left(P_t^g\left(x_{m_1}^g\right), P_s^g\left(x_{m_2}^g\right)\right) \right\} \tag{4}$$

where $m_1 \neq m_2$. This formula makes the [CLS] learn the "local-to-global" correspondences and robust to the difficult scenarios, e.g., occlusion and incorrect detection, which are also why we still append [CLS] to the *local* views.

3.3   Fine-tuning

After being pre-trained by PASS, the [CLS] of ViT backbone has the capability to learn the global description of the input image, and the [PART]s can automatically focus on local areas and extract local representations. Therefore, in fine-tuning, we only need to append the [CLS] and all the [PART]s to the embedding sequence of the input image, and they can automatically focus on different-level features. As commonly used by [2, 22], the teacher network is used for fine-tuning.

Supervised ReID. We concatenate the output [CLS] and the mean of output [PART]s, denoted by [$\overline{\texttt{Part}}$], as the representation of the input image. The ReID head [21] is attached to the concatenated feature. Most of the training hyper-parameters are borrowed from the baseline of TransReID [18]. That is, **none** of the overlapping patch embedding, jigsaw patch module, or side information embedding is used in our fine-tuning. The commonly used cross-entropy loss and triplet loss with hard sample mining [19] are adopted to train our model, of which the cross-entropy loss is calculated as:

$$L_{cls} = -\log \mathcal{P}(\texttt{[CLS]} \copyright \texttt{[}\overline{\texttt{Part}}\texttt{]}), \tag{5}$$

where $\mathcal{P}(\mathbf{x})$ is the probability of $\mathbf{x}$ belonging to its ground truth identity and $\copyright$ means concatenating.

The triplet loss is also calculated as:

$$L_{tri} = [d_p - d_n + \alpha]_+, \tag{6}$$

where d_p and d_n are feature distances from positive pair and negative pair, respectively. α is the margin of triplet loss. $[\cdot]_+$ equals to $max(\cdot, 0)$. Therefore, the overall objective function for our model is:

$$L = L_{cls} + L_{tri}. \tag{7}$$

In the testing phase, [CLS] and [$\overline{\texttt{Part}}$] are concatenated to represent a person image.

UDA/USL ReID. We follow most of the settings in C-Contrast [7] to conduct our experiments on UDA/USL ReID. Before training in an unsupervised manner on target datasets, UDA ReID need to be first pre-trained on source datasets while USL ReID does not need. Apart from this, UDA ReID and USL ReID share the same training settings. First, all the training images pass through the network to obtain the training data features. Then, clustering is conducted on these features to generate pseudo labels. By averaging the features with the same pseudo labels, the cluster prototypes are obtained. Next in training, the contrastive loss between the output features and the cluster prototypes are computed to optimize the network. More details can be found in [7].

4   Experiments

4.1   Implementation Details

Datasets. LUPerson [11], which contains 4.18M unlabeled human images collected from $50,534$ online videos, are used for pre-training. To evaluate the performance on ReID tasks, we conduct the fine-tuning experiments on two widely used benchmarks, i.e., Market-1501 [33] and MSMT17 [28], which contain $32,668$ images of 1501 identities and $126,441$ images of $4,101$ identities, respectively. In fine-tuning, the images are resized to 256×128 unless mentioned otherwise. Following common practices, we use the cumulative matching characteristics (CMC) and the mean average precision (mAP) as the evaluation protocols.

Pre-training. For pre-training on LUPerson with PASS, the model is trained on $8\times$A100 GPUs for 100 epochs, which costs about 60 hours for ViT-S and 120 hours for ViT-B. The divided local areas are with overlap to guarantee that the *local* view can be cropped from anywhere in the image. The *global* views are resized to 256×128 and the *local* views are resized to 96×48. Similar to DINO, we set $M = 2$ and $J = \lceil \frac{9}{L} \rceil$, e.g., $J = 3$ when $L = 3$ and $J = 5$ when $L = 2$.

Supervised ReID. To fine-tune the pre-trained Transformer, we use most of the training strategies of the baseline in TransReID [18], which means none of the overlapping patch embedding, jigsaw patch module, or side information embedding is used here. We set the learning rate to $lr = 0.0004 \times \frac{batchsize}{64}$ and warm up the model by 20 epochs [22]. The α in triplet loss is set to 0.3.

UDA/USL ReID. SGD optimizer is used and the ViTs are trained for 50 epochs. The initial learning rate is set to 3.5e-4 and is multiplied by 1/10 every 20 epochs. Each mini-batch contains 256 images of 32 person IDs, i.e., each ID contains 8 images.

4.2   Comparison with State-of-the-Art Methods

Supervised ReID. We compare PASS to some outstanding state-of-the-art methods on supervised ReID in Table 1. Compared with the methods pre-trained on ImageNet, our method significantly outperforms the existing best

Table 1. Comparison with the state-of-the-art methods of supervised ReID. Methods in the 1st group are pre-trained on ImageNet. Methods in the 2nd group are self-supervised methods pre-trained on LUPerson. Most of the self-supervised methods share the same fine-tuning settings as ours. The last group is our method. The results with underline are the best in their groups. "TransReID$^-$" means side information and overlapping patches are removed for a fair comparison

Methods	Backbone	Market1501		MSMT17	
		mAP	Rank-1	mAP	Rank-1
BOT [21]	R50	85.9	94.5	50.2	74.1
MGN [27]	R50↑384	87.5	95.1	63.7	85.1
SCSN [6]	R50↑384	88.5	95.7	58.5	83.8
ABDNet [3]	R50↑384	88.3	95.6	60.8	82.3
AAformer [35]	ViT-B↑384	87.7	95.4	63.2	83.6
TransReID$^-$ [18]	ViT-B	87.4	94.6	63.6	82.5
TransReID$^-$ [18]	ViT-B↑384	87.6	94.6	65.8	84.4
MoCoV2 [4]	ViT-S	72.1	87.6	27.8	47.4
MoCoV2 [11]	MGN↑384	91.0	96.4	65.7	85.5
MoCoV3 [5]	ViT-S	82.2	92.1	47.4	70.3
MoBY [30]	ViT-S	84.0	92.9	50.0	73.2
DINO [2]	ViT-S	90.3	95.4	64.2	83.4
DINO + CFS [22]	ViT-S	91.0	96.0	66.1	84.6
DINO + CFS [22]	ViT-S↑384	91.5	96.0	68.8	86.1
PASS (ours)	ViT-S	92.2	96.3	69.1	86.5
PASS (ours)	ViT-S↑384	92.6	96.8	71.7	87.9
PASS (ours)	ViT-B	93.0	96.8	71.8	88.2
PASS (ours)	ViT-B↑384	**93.3**	**96.9**	**74.3**	**89.7**

Table 2. Comparison with the state-of-the-art methods of UDA ReID. Methods in the 1st group are pre-trained on ImageNet. Methods in the 2nd group are self-supervised methods pre-trained on LUPerson. All the self-supervised methods are fine-tuned with C-Contrast [7]

Methods	Backbone	MSMT2Market		Market2MSMT	
		mAP	Rank-1	mAP	Rank-1
DG-Net++ [36]	R50	64.6	83.1	22.1	48.4
MMT [12]	R50	75.6	83.9	24.0	50.1
SPCL [13]	R50	77.5	89.7	26.8	53.7
MCRN [29]	R50	-	-	32.8	64.4
C-Contrast [7]	R50	82.4	92.5	33.4	60.5
MoCoV2 [4]	R50	85.1	94.4	28.3	53.8
DINO [2]	ViT-S	88.5	95.0	43.9	67.7
DINO + CFS [22]	ViT-S	89.4	95.4	47.4	70.8
PASS (ours)	ViT-S	**90.2**	**95.8**	**49.1**	**72.7**

methods without adding any complex module to the backbone network, e.g., our ViT-B↑384 obtains 93.3%/74.3% mAP and 96.9%/89.7% Rank-1 accuracy on Market1501/MSMT17, outperforming the state-of-the-art performance by 4.8%/8.5% and 1.2%/4.6%, respectively. It should be noted that AAformer [35] and TransReID are pre-trained on $ImageNet$-$21K$ which is much larger than LUPerson.

Compared with the self-supervised methods pre-trained on LUPerson, PASS also shows considerable superiority, e.g., our ViT-S↑384 obtains 92.6%/71.7% mAP and 96.8%/87.9% Rank-1 accuracy on Market1501/MSMT17 datasets, surpassing the existing best method (CFS [22]) by 1.1%/2.9% and 0.8%/1.8%, respectively. DINO can be regarded as the baseline of our method, and the results validate the remarkable effectiveness of using [PART] to offer fine-grained local information in PASS. Besides, Table 1 also shows that self-supervised pre-training on LUPerson is much more effective than supervised pre-training on ImageNet. The ICS module in [22] adds extra IBN [23] layers to the ViT backbone and increases the overall complexity, thus is not compared here for fairness.

UDA ReID. Some latest UDA-ReID methods are compared in the Table 2. Our ViT-S outperforms the existing best method by considerable margins. Specifically, ViT-S obtains 90.2%/49.1% (+0.8%/+1.7%) mAP and 95.8%/72.7% (+ 0.4%/+1.9%) Rank-1 accuracy on MSMT17 → Market1501 and Market1501 → MSMT17 respectively, which are already comparable to many supervised methods. Besides, PASS also surpasses its baseline method (DINO) by large margins, which validates local details are also vital for unsupervised learning in ReID.

Table 3. Comparison with the state-of-the-art methods of USL ReID. Methods in the 1st group are pre-trained on ImageNet. Methods in the 2nd group are self-supervised methods pre-trained on LUPerson. All the self-supervised methods are fine-tuned with C-Contrast [7]

Methods	Backbone	Market1501 mAP	Market1501 Rank-1	MSMT17 mAP	MSMT17 Rank-1
MMCL [26]	R50	45.5	80.3	11.2	35.4
HCT [32]	R50	56.4	80.0	-	-
IICS [31]	R50	72.9	89.5	26.9	52.4
MCRN [29]	R50	80.8	92.5	31.2	_63.6_
C-Contrast [7]	R50	_82.6_	_93.0_	_33.1_	63.3
MoCoV2 [4]	R50	84.0	93.4	31.4	58.8
DINO [2]	ViT-S	87.8	_94.4_	38.4	63.8
DINO + CFS [22]	ViT-S	_88.2_	94.2	_40.9_	_66.4_
PASS (ours)	ViT-S	**88.5**	**94.9**	**41.0**	**67.0**

USL ReID. We list some of the latest USL-ReID methods in Table 3, which shows our method outperforms the existing best method by 0.7% (94.9% *vs.* 94.2%) and 0.6% (67.0% *vs.* 66.4%) on Rank-1 accuracy on Market and MSMT17, respectively. The results validate PASS can provide a better initialization to the ViT model on USL ReID.

4.3   Ablation Studies

The choices of L. We first investigate the most suitable division strategy for PASS to divide the person images into several local areas. A *local* view accounts for up to 40% of the whole image, thus when $L = 2$, each local area should occupy 70% of the image (the same width as the original image and 70% height) to guarantee that the *local* view can be cropped from anywhere in the image. For the same reason, when $L = 3$, each local area should occupy 50% of the image. We also follow MGN [27] to conduct the ablation experiments which use $L = \{2, 3\}$, where two kinds of division granularity are used at the same time. The results in Table 4 show that increasing the number of local areas does not always bring positive feedback, and PASS with $L = 3$ obtains the best performance. Therefore, we recommend dividing the image into 3 local areas with each area occupying 50% of the image. The 3 local areas are evenly arranged from top to bottom of an image and *local* views are cropped from each local area.

Table 4. Ablation studies about the number of divided local areas L on supervised ReID

L	Backbone	Market1501		MSMT17	
		mAP	Rank-1	mAP	Rank-1
2	ViT-S	91.7	96.1	67.7	85.6
3	ViT-S	**92.2**	**96.3**	**69.1**	86.5
2,3	ViT-S	92.0	96.3	68.7	**86.6**

Feature fusion. Next, we conduct experiments to investigate how to effectively integrate the output global feature and local features on supervised/UDA/USL ReID, which are shown in Table 5 and Table 6, respectively. One way is directly concatenating all these features, which is used by TransReID [18]. However, this way increases the dimension of feature embedding several times and cannot obtain the best performance in our experiments. We also conduct the ablation study which uses the mean feature or concatenating [CLS] and [$\overline{\text{Part}}$]. The results show concatenating [CLS] and [$\overline{\text{Part}}$] can obtain the best performance on supervised ReID. It is also worth noting that the strategy of "mean feature", which does not increase the feature dimension, already outperforms the existing state-of-the-art methods, some of which increase the feature dimension several times [27, 18], by considerable margins. On UDA/USL ReID, MSMT17 dataset prefers the "mean feature" while Market1501 prefers concatenating [CLS] and [$\overline{\text{Part}}$], which may be related to their data distributions.

Table 5. Ablation studies about feature fusion on supervised ReID. ⓒ means concatenating. C is the number of channels. The first row is concatenating all these features. The second row uses the mean feature. The third row is concatenating `[CLS]` and `[Part]`

Method	Dim	Backbone	Market1501		MSMT17	
			mAP	Rank-1	mAP	Rank-1
`[CLS]`ⓒ$\frac{[\texttt{PART}]_1}{L}$ⓒ...$\frac{[\texttt{PART}]_L}{L}$	(L+1)×C	ViT-S	90.8	96.1	67.7	85.2
		ViT-B	92.2	96.3	71.3	87.4
$\frac{1}{2}$(`[CLS]`+$[\overline{\texttt{Part}}]$)	C	ViT-S	92.0	**96.3**	68.7	86.3
		ViT-B	92.5	96.7	71.5	87.8
`[CLS]`ⓒ$[\overline{\texttt{Part}}]$	2C	ViT-S	**92.2**	**96.3**	**69.1**	**86.5**
		ViT-B	**93.0**	**96.8**	**71.8**	**88.2**

Table 6. Ablation studies about feature fusion on UDA/USL ReID. The backbone is ViT-S and ⓒ means concatenating. The first row is concatenating all these features. The second row uses the mean feature. The third row is concatenating `[CLS]` and `[Part]`

Method	UDA reID				USL reID			
	MS2MA		MA2MS		Market1501		MSMT17	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
`[CLS]`ⓒ$\frac{[\texttt{PART}]_1}{L}$ⓒ...$\frac{[\texttt{PART}]_L}{L}$	90.0	95.6	47.1	71.1	88.4	94.6	39.4	64.9
$\frac{1}{2}$(`[CLS]`+$[\overline{\texttt{Part}}]$)	89.8	95.5	**49.1**	**72.7**	**88.6**	94.6	**41.0**	**67.0**
`[CLS]`ⓒ$[\overline{\texttt{Part}}]$	**90.2**	**95.8**	44.9	69.4	88.5	**94.9**	36.6	62.5

Input   [CLS]   [PART]₁   [PART]₂   [PART]₃   Input   [CLS]   [PART]₁   [PART]₂   [PART]₃

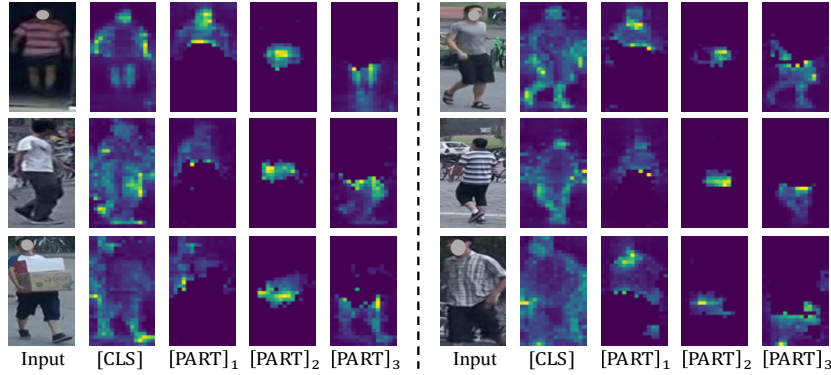**Fig. 3.** Visualization of attention maps of `[CLS]` and `[PART]`s in the last self-attention layer of ViT pre-trained by PASS. For each `[PART]`, the patches that have the maximal similarity with it among all the `[PART]`s are visualized. The examples in the last row show the visualization on occlusion and bad detection scenarios

Visualization. To further give an intuitive illustration of the effectiveness of PASS, we conduct visualization experiments to show the focus area of `[CLS]` and `[PART]`s in the ViT backbone pre-trained by PASS. As illustrated in Figure 3, given an input image, `[CLS]` focuses on the whole images to extract the global feature, and the `[PART]`s focus on different local areas, i.e., upper-body, waist, and legs, to extract part-level features, which can contain more fine-grained information. More satisfactory, `[PART]`s can locate on specific semantic parts rather than simply locating different positions. Take the bottom-left sample in Figure 3 as an example, when a man carries a box and the box blocks his upper body, the first `[PART]` does not focus on the box but focus on the visible part of his upper body. This makes the learned feature more robust as the man may carry boxes with different colors in other images. This visualization also validates PASS can well handle the occlusion scenarios. It is worth noting that we do not add any extra module to the ViT backbone to achieve this, but only pre-train the ViT backbone by PASS.

Fig. 4. The ranking lists of models pre-trained by DINO and PASS. For each query, the ranking list of the first row is from DINO and that of the second row is from PASS. The results show the PASS is largely superior to DINO in finding fine-grained information of local details

Ranking list. Finally, we compare the ranking results of ViT pre-trained by PASS with DINO (baseline) in Figure 4. *The first example shows even the person takes different boxes with different colors, our model can still find him in the massive images, which validates our model is not affected by the obstructions and focuses on the details of visible parts.* The other samples show the strong ability of our model in discriminating the extremely similar samples through the

identifiable tiny clues, e.g., patterns on clothes. All these samples validate the superiority of PASS over the baseline method DINO.

5   Conclusion

In this paper, we propose a ReID-specific self-supervised pre-training method, Part-Aware Self-Supervised pre-training (PASS). Pre-trained by PASS, the models can automatically extract part-level features from the input images and offer more fine-grained information. Experimental results validate the powerful performance of PASS in both supervised ReID and UDA/USL ReID.

References

1. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 (2020)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294 (2021)
3. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abdnet: Attentive but diverse person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
4. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
5. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9640–9649 (2021)
6. 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. pp. 3300–3310 (2020)
7. Dai, Z., Wang, G., Zhu, S., Yuan, W., Tan, P.: Cluster contrast for unsupervised person re-identification. arxiv 2021. arXiv preprint arXiv:2103.11568 (2021)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186 (2019)
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. Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., Li, H., Chen, D.: Unsupervised pre-training for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14750–14759 (2021)

12. Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. arXiv preprint arXiv:2001.01526 (2020)
13. Ge, Y., Zhu, F., Chen, D., Zhao, R., Li, H.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. arXiv preprint arXiv:2006.02713 (2020)
14. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
16. 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. pp. 770–778 (2016)
17. He, L., Liao, X., Liu, W., Liu, X., Cheng, P., Mei, T.: Fastreid: A pytorch toolbox for general instance re-identification. arXiv preprint arXiv:2006.02631 (2020)
18. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (2021)
19. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
20. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1062–1071 (2018)
21. 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 Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
22. Luo, H., Wang, P., Xu, Y., Ding, F., Zhou, Y., Wang, F., Li, H., Jin, R.: Self-supervised pre-training for transformer-based person re-identification. arXiv preprint arXiv:2111.12084 (2021)
23. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 464–479 (2018)
24. 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). pp. 480–496 (2018)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6000–6010 (2017)
26. Wang, D., Zhang, S.: Unsupervised person re-identification via multi-label classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10981–10990 (2020)
27. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: 2018 ACM Multimedia Conference on Multimedia Conference. pp. 274–282. ACM (2018)
28. 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. pp. 79–88 (2018)

29. Wu, Y., Huang, T., Yao, H., Zhang, C., Shao, Y., Han, C., Gao, C., Sang, N.: Multi-centroid representation network for domain adaptive person re-id. In: Proceedings of the AAAI Conference on Artificial Intelligence (2022)
30. Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., Hu, H.: Self-supervised learning with swin transformers. arXiv preprint arXiv:2105.04553 (2021)
31. Xuan, S., Zhang, S.: Intra-inter camera similarity for unsupervised person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11926–11935 (2021)
32. Zeng, K., Ning, M., Wang, Y., Guo, Y.: Hierarchical clustering with hard-batch triplet loss for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13657–13665 (2020)
33. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Computer Vision, IEEE International Conference on Computer Vision. pp. 1116–1124 (2015)
34. Zhu, K., Guo, H., Liu, Z., Tang, M., Wang, J.: Identity-guided human semantic parsing for person re-identification pp. 346–363 (2020)
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 preprint arXiv:2104.00921 (2021)
36. Zou, Y., Yang, X., Yu, Z., Kumar, B., Kautz, J.: Joint disentangling and adaptation for cross-domain person re-identification. In: European Conference on Computer Vision. pp. 87–104. Springer (2020)