
PersonViT: Large-scale Self-supervised Vision Transformer for Person Re-Identification

Bin Hu

Institute of Artificial Intelligence, Huazhong University of Science and Technology
Wuhan, China
hubin@hust.edu.cn

Xinggang Wang

School of EIC, Huazhong University of Science and Technology
Wuhan, China
xgwang@hust.edu.cn

Wenyu Liu*

School of EIC, Huazhong University of Science and Technology
Wuhan, China
liuwu@hust.edu.cn

Abstract

Person Re-Identification (ReID) aims to retrieve relevant individuals in non-overlapping camera images and has a wide range of applications in the field of public safety. In recent years, with the development of Vision Transformer (ViT) and self-supervised learning techniques, the performance of person ReID based on self-supervised pre-training has been greatly improved. Person ReID requires extracting highly discriminative local fine-grained features of the human body, while traditional ViT is good at extracting context-related global features, making it difficult to focus on local human body features. To this end, this article introduces the recently emerged Masked Image Modeling (MIM) self-supervised learning method into person ReID, and effectively extracts high-quality global and local features through large-scale unsupervised pre-training by combining masked image modeling and discriminative contrastive learning, and then conducts supervised fine-tuning training in the person ReID task. This person feature extraction method based on ViT with masked image modeling (PersonViT) has the good characteristics of unsupervised, scalable, and strong generalization capabilities, overcoming the problem of difficult annotation in supervised person ReID, and achieves state-of-the-art results on publicly available benchmark datasets, including MSMT17, Market1501, DukeMTMC-reID, and Occluded-Duke. The code and pre-trained models of the PersonViT method are released at <https://github.com/hustvl/PersonViT> to promote further research in the person ReID field.

1 Introduction

Person re-identification (ReID) aims to learn visual features from human images that can distinguish different individual identities. It is an important and challenging computer vision problem that requires overcoming severe occlusion, appearance changes, shape changes, and viewpoint changes. Person

*Corresponding Author.

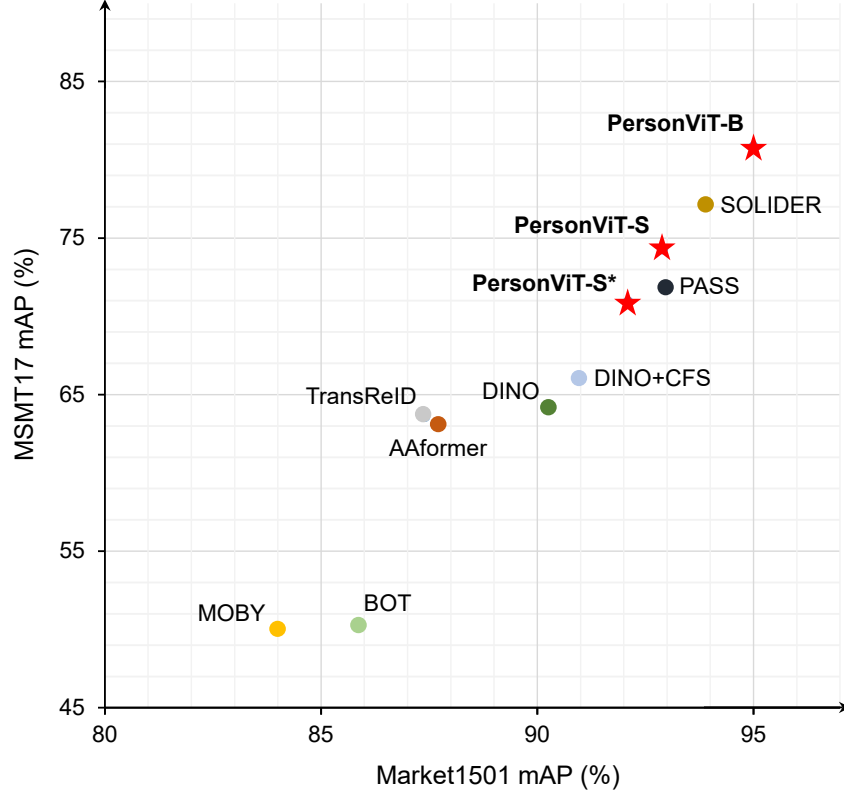


Figure 1: Person ReID performance on both MSMT17 and Market1501. The proposed PersonViT method obtains SOTA results and significantly outperforms previous methods.

ReID technology enables effective cross-camera retrieval of people in contactless and uncooperative scenarios and is widely used in public safety, video surveillance, and other fields, offering significant application value.

There have been many studies related to person ReID, most of them involve metric learning (for example, Triplet Loss [1]) on features extracted by backbone networks (such as ResNet [2] or ViT [3]). Constrained by the small size of labeled ReID datasets, most of the backbone networks in the current methods are pre-trained on ImageNet [4] and then fine-tuned on ReID datasets to achieve better performance. However, there is a significant difference between ImageNet and person ReID datasets, as ImageNet contains images of a thousand categories while ReID has only one person category. Therefore, models pre-trained on ImageNet can better extract category-level features but struggle to effectively extract fine-grained individual features, making them less suitable for the person ReID problem [5].

To bridge the gap between pre-training and fine-tuning datasets for better ReID backbones, [6] collected an unlabeled person dataset called LUPerson and demonstrated for the first time that CNN-based self-supervised pre-trained models perform well in ReID tasks. Subsequently, TransReID-SSL [7] conducted experiments using ViT-based self-supervised learning algorithms on the LUPerson dataset and found that models pre-trained with the DINO [8] algorithm were most effective for person ReID. However, these self-supervised learning algorithms all utilize the classification consistency between local and global images generated from different data augmentations of the same image to perform contrastive self-supervised learning. They can capture the inter-class differences in multi-class datasets such as ImageNet relatively well, but lose the fine-grained differences in single-class ReID datasets. For example, local images of different human bodies may be very similar and can belong to the same class, but DINO’s self-supervised contrast will separate them into different classes. Similarly, local images of the same person at different angles should belong to the same class, but DINO’s contrast will separate them. Therefore, the PASS method [5] proposed a self-supervised pre-training algorithm designed specifically for person ReID tasks. Based on DINO, it divides

person images into several (L) blocks according to the human body structure from top to bottom. It performs contrastive self-learning between blocks and between blocks and the whole image, thus enhancing the pre-trained model’s ability to express local features and achieving new state-of-the-art at the time. However, PASS has two problems: 1) like the previous block-based ReID such as PCB [9], it has an alignment dependency problem, i.e., the block division method is too mechanical in complex backgrounds, non-uniform alignments, and even occluded or incomplete human images, making it prone to misclassification; 2) a smaller number of blocks still struggle to fully express fine-grained local features, i.e., in the PASS method, the optimal performance experiment corresponds to a partition count of only three, which is evidently insufficient in granularity.

Recently, inspired by the monumental success of Masked Language Modeling (MLM), i.e. BERT [10], in the Natural Language Processing (NLP) field, Masked Image Modeling (MIM) techniques such as BEiT [11], MAE [12], and SimMIM [13] has also made breakthroughs in image classification, detection, and segmentation tasks in natural images through self-supervised learning. MIM employs random masking of local pixels for reconstruction learning, reinforcing the learning and extraction of local fine-grained features without manual image partitioning. This effectively compensates for the limitations of the PASS algorithm, indicating that the introduction of MIM should enhance the accuracy of person ReID.

In this paper, we incorporate a MIM self-supervised learning module based on DINO, conducting large-scale, unlabeled pre-training on the LUPerson [6] dataset, and subsequently fine-tuning supervised ReID on four datasets: MSMT17 [14], Market1501 [15], DukeMTMC-reID [16], and Occluded-Duke [17]. We refer to this method of utilizing masked image modeling and DINO contrastive learning for large-scale unsupervised pre-training based on vanilla ViT and fine-tuning on downstream datasets as the PersonViT method. Experimental results reveal that PersonViT achieves state-of-the-art results, particularly excelling on the challenging dataset Occluded-Duke. Furthermore, visualization analysis of the pre-trained models on MSMT17 discloses that PersonViT can automatically discover key human body parts, clothing patterns, and the correspondence between local body parts without annotations.

The innovations of this paper encompass the following aspects: 1) To reinforce the algorithm’s robustness to human image alignment and better extract the local fine-grained features of the human body, we pioneer the utilization of masked image modeling techniques in unsupervised feature learning for person ReID. This effectively addresses the challenge of acquiring local fine-grained visual features in human images with substantial occlusions and misalignments. 2) We propose a highly efficient, large-scale self-supervised human feature learning method based on the vanilla ViT, referred to as PersonViT. 3) Our approach achieves state-of-the-art results on several mainstream person ReID datasets.

The main significance of this research is that it achieves industry-leading ReID results through efficient and accurate self-supervised feature learning, greatly addressing the dilemma of insufficient labeled training data in person ReID. The proposed method is highly scalable and provides strong technical support for large-scale practical applications of ReID technology. Experimental results show that our method achieves industry-leading accuracy, significantly surpassing previous approaches.

2 Related Work

2.1 Self-Supervised Learning

Contrastive Learning Self-supervised learning (SSL) methods aim to learn discriminative features from large-scale unlabeled data [18]. In recent years, contrastive learning methods have flourished in the field of computer vision [19; 20; 21; 22; 23; 24; 8], significantly narrowing the gap with supervised pretraining. MOCO [19] first proposed momentum contrast, in which a pair of enhancements of a sample is regarded as positive, while other samples and their enhancements as negative, thereby conducting unlabeled contrastive learning training. This approach has been followed by a series of improvements such as MOCOv2 [20], MOCOv3 [21], and SimCLR [22]. BYOL [23] proposed a new contrastive learning paradigm that uses two network models to predict the representations of different enhanced images of the same sample, eliminating the dependence on large batches of negative samples. DINO [8] is an improved version of BYOL. It introduces centering and sharpening operations into the process of momentum smoothing update of the target model parameters, effectively preventing model collapse and improving the algorithm’s stability. In addition, DINO incorporates

extensive data enhancements, especially multiple local cropping enhancements, which enhance the model’s ability to learn local features through large-scale contrastive learning between local and global images.

DINO Framework DINO [8] serves as the inaugural self-distillation learning framework. Ahead of delving into DINO, let’s briefly touch upon the foundational principles behind Transformer. Viewed as an encoder, the Transformer translates images $I \in \mathbb{R}^{h \times w \times c}$, where $h \times w$ corresponds to the image resolution and c to the number of image channels, typically 3, denoting RGB, into target feature vectors. Drawing parallels with tokenization procedures in NLP, the image is bucketed into continuous entities. Transformers initiate the process through block projection tokenization of the images. If each $p \times p$ pixel-sized block is partitioned into $n = hw/p^2$ image patches, every segment can be depicted as a token, bearing similarity to individual words in NLP. Coupled with a learnable *cls_token*, these tokens can represent image tokenization results as X as indicated in Eq. (1). Here, the ";" symbol symbolizes the connection between stacked tokens. Following encoding via the ViT network, a similar representation can be achieved for feature vectors as per Eq. (2). Contrasting with the conventional ViT backbone network, the DINO encoder integrates a head module that channels the ViT output $z^{[cls]}$ through a multi-layer perception network (MLP) to map onto the targeted vector space $y^{[cls]}$ as illustrated in Eq. (3). To calculate the ultimate loss function, denoted as L_{dino} for differentiation from subsequent loss functions, DINO utilizes two isomorphic encoder networks - the student and teacher network, their respective outputs denoted as $Y^{[S]}$ and $Y^{[T]}$ as indicated in Eq. (4). The norm of gradient backpropagation updates the student network parameters, whereas the teacher network parameters evolve in tandem with the Exponential Moving Average (EMA) of the student network parameters, calculated as per $\theta_t = \lambda\theta_t + (1 - \lambda)\theta_s$. In an interesting twist, the DINO teacher network inputs two globally amplified data views while the student network calls upon corresponding global views as well as multiple local views. Consequently, L_{dino} encapsulates both global and local-global contrast calculations, thereby enhancing the discernment of local features.

$$\begin{aligned} X &= (x^{[cls]}; x^{[patches]}) \\ &= (x^{[cls]}; x_1; \dots; x_n) \in \mathbb{R}^{(n+1) \times d} \end{aligned} \quad (1)$$

$$\begin{aligned} Z &= (z^{[cls]}; z^{[patches]}) \\ &= (z^{[cls]}; z_1; \dots; z_n) \in \mathbb{R}^{(n+1) \times d} \end{aligned} \quad (2)$$

$$Y = \text{MLP}(z^{[cls]}) = y^{[cls]} \in \mathbb{R}^{1 \times d_{output}} \quad (3)$$

$$L_{dino} = -\text{Softmax}(Y^{[T]}) \log(\text{Softmax}(Y^{[S]})) \quad (4)$$

Masked Image Modeling Starting from 2018, the field of NLP has witnessed notable success in Masked Language Modeling (MLM), represented by models such as BERT [10] and GPT [25]. In recent years, several masked modeling attempts have also emerged in the field of vision, with representative works including [26; 13; 12; 27; 28]. MST[26], the first model to introduce masked image modeling, enhances the performance of DINO by adding masked predictive training to the DINO framework. Following closely is BEiT [11], which first uses a discrete Variational AutoEncoder (dVAE) to discretize image blocks and map them to corresponding vision tokens. It then masks the obstructed vision tokens to achieve prominent learning results; therefore, BEiT also operates as a two-stage training algorithm. MAE [12] is an end-to-end MIM autoencoder. It employs an encoder to code only visible image blocks and utilizes a relatively lightweight decoder to restore masked pixels. Interestingly, it has been shown to produce outstanding self-supervised learning results even when a majority of the input image (e.g., 75%) is masked, thereby significantly accelerating the pre-training process. As MAE’s self-supervised learning is to restore pixels, both its encoder and decoder learn certain feature expressions. Moreover, MAE restores the most primitive low-level pixel features, so the degree of feature abstraction learned by the encoder is not controlled. The relatively low accuracy of the k-NN and linear probing benchmarks in MAE’s ImageNet experiment results signifies that the encoder does not sufficiently extract distinguishing feature abstracts.

2.2 Part-based person ReID

Traditionally, feature expression learning predominantly employed an IDE model [29]. This approach involved global feature extraction followed by multi-classification training, where each individual was treated as a separate class. Gradually, the trend evolved to use ResNet50 [2] to extract global feature vectors, which were then subjected to metric learning, such as Triplet Loss [1]. However, extracting detailed local features remained a challenge for global feature representation learning. As a result, two primary forms of optimization emerged to address this. The first involved multi-scale fusion representation learning as presented by documents such as [30]. The second included the integration of attention mechanisms to strengthen local representation learning, such as harmonizing spatial and channel attention [31], and attenuating background in representation learning using segmentation attention mechanism [32]. To better capture local, fine-grained features of human images and improve the robustness of algorithms in handling alignment issues and obstructions, local feature-based person ReID methods were subsequently proposed. Early proposed local representation learning methods, such as those found in [33], [34], and PCB [9], partitioned human images into multiple horizontal stripes to extract local features. MGN [35] enhanced the robustness using different granularity stripes and overlaps between these stripes. TransReID [36], being the first Transformer-based person ReID algorithm, also involves local representation learning by rearranging and regrouping Transformer’s input local images. These methodologies highlight that the nuanced features extracted through local representation learning significantly bolster the precision of person ReID.

2.3 Self-supervised person ReID

Although research related to person ReID in self-supervised pre-training surfaced relatively late and is therefore scarcer, it showcases tremendous potential due to the avoidance of high-cost annotations in the pre-training phase. TransReID-SSL [7] was the first to compare the performance of the then mainstream self-supervised learning algorithms on the problem of ReID. In the comparison, DINO [8] demonstrated significant advantages over other algorithms. The researchers also studied the effect of pre-training data size on final accuracy and proposed an effective method for screening pre-training data. PASS [5], the first self-supervised pre-training algorithm geared towards the ReID task, segments human images. Building on the foundation of DINO, PASS introduced local-to-local and local-to-global contrast learning. It differentiates representations by using separate *cls_tokens* for different locals, and in addition to the global contrast classification feature space[CLS], also introduces additional segment classification feature space[PART]s, thereby enhancing the detailed representation capability of segments and improving local feature extraction ability of the pre-training model. The method set industry-leading benchmarks at the time in both supervised person ReID and cross-domain person ReID, thereby validating its effectiveness.

2.4 Summary of Research on person ReID

Reflecting on the evolution of person ReID over the past few years, we can summarize three major trends.

Firstly, the feature extraction backbone network has evolved from traditional convolutional neural networks to Transformers (ViT), primarily due to the advanced capabilities of ViT to express global contextual features and the significant developments in ViT-based self-supervised learning.

Secondly, the initial pre-training approaches based on ImageNet classification have moved towards self-supervised models using extensive human data. This shift is mainly because ReID challenges involve distinguishing minor differences between different individuals, unlike ImageNet classification tasks with varied and broad class differences. As the test individuals are typically not present in the training subsets, the generalization capability forms a pivotal component driving recognition accuracy. Nevertheless, the scarcity of public supervised training sets, resulting from challenges in data acquisition and labeling, falls short in fulfilling the requirements for algorithmic generalization capability.

Lastly, given the nuanced distinctions between different individuals—a much smaller scope compared to class variations in ImageNet—effective extraction of local fine-grained features proves crucial in enhancing ReID accuracy.

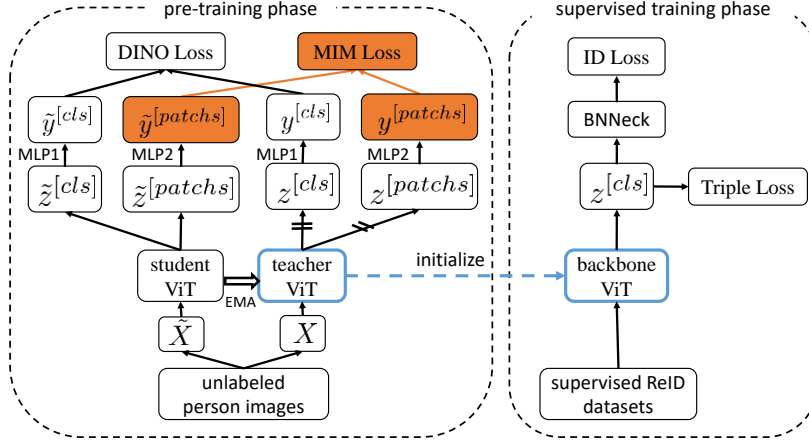


Figure 2: Overview of PersonViT framework.

The latest methodology, PASS [5], successfully aligns with these three trends and achieved industry-leading scores at the time of its inception. However, it highlights two critical limitations. Firstly, its reliance on manually designed segment partitioning restricts its robustness to the alignment of human images. Secondly, the finite number of partitions makes it challenging to adequately express local fine-grained features, thereby constraining the capability of the pre-training model to extract such features.

3 Method

Addressing the limitations identified in the PASS algorithm, a potential solution might be to introduce a classification feature space to each image patch input in the ViT model, akin to the [CLS] feature space, in the context of self-supervised pre-training. However, applying traditional data augmentation-based contrastive learning methods might pose challenges. Given the small size of each patch, the process of generating different data augmentations could easily lead to misplacements within local image patches, limiting comparability and not achieving the desired self-supervised effects. This is evidenced by the experimental results from the PASS method, which demonstrated that increasing the number of local segments or reducing the size of partitions does not consequently improve accuracy.

Inspired by the concept of masked image modeling, a potential solution might be to feed the same image into the system twice. One input would be the complete image, and the other would be the image with some patches masked. The corresponding output feature vectors from both inputs, denoted [PATCH] and represented as $\tilde{y}^{[patches]}$ in Eq. (6), would form a basis for comparison. This approach allows for feature representation at the most granular level, corresponding to the size of the patch, and could improve the pre-training model’s ability to extract fine-grained local features.

The overall algorithm framework proposed in this paper, depicted in Fig. 2, incorporates a contrast loss module at the patch granularity level (known as the MIM Loss module) into the DINO pre-training algorithm. Pre-training is executed with ViT-S and ViT-B as the backbone networks utilizing the largest publicly available dataset, LUPerson [6]. This is followed by supervised fine-tuning training on four mainstream public person ReID datasets: MSMT17 [14], Market1501 [15], DukeMTMC-reID [16], and Occluded-Duke [17]. The procedures for both pre-training and fine-tuning are detailed below.

3.1 Incorporation of Self-Supervised Pre-training MIM Loss Function

Inspired by the masked image modeling paradigm of BEiT [11], our methodology encompasses the random block-wise masking of images, akin to the introduction of a learnable token variable $x^{[mask]}$, analogous to $x^{[cls]}$. Consequently, the masked image may be delineated as $\tilde{X} = (x^{[cls]}; \tilde{x}_1; \dots; \tilde{x}_n)$

with \tilde{x}_i detailed in Eq. (5).

$$\begin{aligned}\tilde{X} &= (x^{[cls]}; \tilde{x}_1; \dots; \tilde{x}_n), \\ \tilde{x}_i &= (1 - m_i)x_i + m_i x_i^{[mask]}, i \in 1, \dots, n\end{aligned}\tag{5}$$

In this equation, $m_i \in 0, 1$ signifies the random image block mask, where 1 corresponds to masked and 0 to unmasked. The masked image \tilde{X} , when processed through the ViT encoder, yields \tilde{Z} , as depicted in Eq. (2).

$$\begin{aligned}\tilde{Y} &= (\text{MLP1}(\tilde{z}^{[cls]}); \text{MLP2}(\tilde{z}^{[patchs]})) \\ &= (\tilde{y}^{[cls]}; \tilde{y}^{[patchs]}) \in \mathbb{R}^{(1+n) \times d_{output}}\end{aligned}\tag{6}$$

$$\begin{aligned}L_{mim} &= \sum_{i=1}^n m_i \cdot P(y_i^{[patchs][T]}) \log(P(\tilde{y}_i^{[patchs][S]})), \\ P(x) &= \text{Softmax}(x)\end{aligned}\tag{7}$$

$$L = \lambda_1 L_{dino} + \lambda_2 L_{mim}\tag{8}$$

Strikingly different from DINO, here we also implement an MLP network transformation for $\tilde{z}^{[patchs]}$, thereby projecting it onto a d_{output} -dimensional vector space to produce $\tilde{y}^{[patchs]}$, prior to executing the reconstruction loss calculation. The target vector $\tilde{Y} = (\tilde{y}^{[cls]}; \tilde{y}_1; \dots; \tilde{y}_n)$ can then be articulated as shown in Eq. (6). To benchmark the reconstructed vector of the masked image blocks, masking is solely applied to the dual global view inputs of the student network, with the teacher network’s input comprising an intact global view for reference. The specific masked reconstruction loss function (abbreviated as L_{mim} is explicated in Eq. (7). This, in conjunction with the DINO loss function L_{dino} as evinced in Eq. (4), forms the final pretraining loss function L as outlined in Eq. (8). Here, L is conceived as the weighted sum of L_{dino} and L_{mim} , wherein $\lambda_1 = \lambda_2 = 1$ by default.

3.2 Supervised Fine-tuning

In this stage, the pre-training model produced from prior self-supervised learning undergoes fine-tuning for the specific task of person ReID, generating the final model for this task. Subsequently, we conduct tests to evaluate the ReID accuracy of the model. To ensure unbiased comparison of the efficacy of the self-supervised pre-training model, we continue utilising the BOT [37] framework during this stage. This approach aligns with the methodology employed in the baseline of TransReID-SSL, which consists of using a standard ViT network (i.e., ViT-S/16 and ViT-B/16) for the primary network, directly implementing $z^{[cls]}$ for feature aggregation, adopting Triplet Loss for metric learning, choosing Cross-entropy Loss for ID Loss, and inserting a BNNeck module between metric learning and ID Loss. For the primary standard ViT network, we initiate fine-tuning for person ReID training by using the pre-training model from the teacher network in the self-supervised learning stage as the starting reference point.

4 Experiments

4.1 Datasets

The main dataset used for self-supervised pre-training is LUPerson [6], which includes 4.18 million unlabeled human images. This dataset is four times the size of ImageNet, therefore, the computational resources required for pre-training are larger, and the same computational resources will result in four times the training time of the ImageNet dataset. To thoroughly verify the pre-training model’s ability to extract finer-grained local features, this research conducts supervised training on the four mainstream person ReID datasets, MSMT17 [14], Market1501 [15], DukeMTMC-reID [16], and Occluded-Duke [17], observing boosts in two key indicators: mAP (mean average precision) and Rank-1. The detailed information of the four datasets is shown in Table 1. Occluded-Duke, which is generated based on DukeMTMC-ReID, enhances the difficulty of ReID under occlusion situations.

Table 1: Statistics of some commonly used datasets for person ReID.

Dataset	Time	#ID	#image	#cam
Maket-1501	2015	1501	32668	6
DukeMTMC	2017	1404	36411	8
MSMT17	2018	4101	126441	15
Occluded-Duke	2019	1404	36411	8

4.2 Implementation Details

4.2.1 Self-Supervised Pre-Training Stage

Due to the large size of the LUPerson dataset and to reduce the experimental timeframe, we implemented large-batch training using $8 \times 8 \times A100$ GPUs to expedite our experiments. Nonetheless, to demonstrate the effectiveness of small batch training, we also carried out foundational experimentation using $4 \times RTX3090$ GPUs. To minimize the computational load during the pre-training phase, the number of training cycles (epochs) was set to 300. Similar to DINO and PASS, the teacher network accepted input with image dimensions of 256×128 , whereas the student network was designed to process global views with dimensions of 256×128 and only 6 local views with dimensions of 96×64 . Considering the extended time taken for pre-training, our experiments only trained two basic network models - ViT-S/16 and ViT-B/16. For the small batch experiment, the pre-training learning rate uniformly adopts $lr = 0.0005 * batch_size / 256$. For large-batch experiments, despite a significant reduction in experimental duration, the training process was rather unstable. Consequently, we adjusted the learning rate based on the specific batch size, making sure not to exceed 0.002, until stable training convergence was achieved. The specific parameters and training logs will be publicly released along with the code.

4.2.2 Supervised Training Stage

Based on the experimental setup of TransReID-SSL [7], the supervised fine-tuning was performed without adding any other optimization items, only using traditional ViT-S/16 and ViT-B/16 Transformer networks as the backbone for supervised training. The supervised training uniformly adopted stochastic gradient descent as the learning algorithm, and the learning rate was set as $lr = 0.0004 * batch_size / 64$. The batch size was composed of $4 * 16 = 64$, which means 16 different individuals per batch, with 4 images for each person. Following the same strategy as PASS [5], the first 20 cycles served as a warm-up phase, and the α parameter for Triplet Loss was set at 0.3.

4.3 Experimental Results

The comparison of different algorithms' experimental results is shown in Table 2 (the light blue background represents the highest accuracy). To make a fair comparison about the role of the pre-training parameters of the backbone network, TransReID- represents the ReID accuracy without the SIE and JPM modules from TransReID [36]. Furthermore, the last two rows represent the accuracy of the models pre-trained with $64 \times A100$ large batch size. The row indicated with a * represents the accuracy of models pre-trained with a small batch size on $4 \times RTX3090$. AAformer [38] and TransReID- are models pre-trained for classification on the ImageNet-21K dataset, which is larger than the LUPerson dataset and has classification labels. As seen, the ReID accuracy of the self-supervised pre-training models based on LUPerson far exceeds those based on ImageNet pre-training models.

Comparing with DINO+CFS [7] and PASS [5], the LUPerson pre-training models based on PersonViT proposed in this paper, and particularly the large batch size pre-training models, achieve significant improvements in accuracy, greatly surpassing the model pre-trained with the PART algorithm based on image partitioning self-supervised learning.

4.3.1 Hyper-parameter Ablation Experiment

The DINO study [8] demonstrated a notable influence of multi-crop techniques on performance. Considering the distinct characteristics of pedestrian images, it becomes relevant to investigate the implications of various hyper-parameters in global and local cropping. The person ReID pre-training

Table 2: Comparison with the state-of-the-art methods.

Methods	Backbone	MSMT17		Market1501		DukeMSMT		Occluded-Duke	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
BOT [37] <small>CVPRW2019</small>	R50	50.2	74.1	85.9	94.5	-	-	-	-
AAformer [38] <small>Arxiv</small>	ViT-B/16 \uparrow 384	63.2	83.6	87.7	95.4	-	-	-	-
TransReID ⁻ [36] <small>ICCV2021</small>	ViT-B/16	63.6	82.5	87.4	94.6	-	-	-	-
MOBY [24] <small>Arxiv</small>	ViT-S/16	50.0	73.2	84.0	92.9	-	-	-	-
DINO [8] <small>ICCV2021</small>	ViT-S/16	64.2	83.4	90.3	95.4	-	-	-	-
DINO+CFS [7] <small>Arxiv</small>	ViT-S/16	66.1	84.6	91.0	96.0	-	-	-	-
PASS [5] <small>ECCV2022</small>	ViT-S/16	69.1	86.5	92.2	96.3	82.5	90.7	60.2	70.4
PASS [5] <small>ECCV2022</small>	ViT-B/16	71.8	88.2	93.0	96.8	84.7	92.5	64.3	74.0
SOLIDER [39] <small>CVPR2023</small>	Swin-B	77.1	90.7	93.9	96.9	-	-	-	-
PersonMAE [40] <small>TMM2024</small>	ViT-S/16	75.2	89.1	92.5	96.7	-	-	65.2	72.0
PersonMAE [40] <small>TMM2024</small>	ViT-B/16	79.8	91.4	93.6	97.1	-	-	69.5	76
PersonViT(ours)*	ViT-S/16	70.9	87.3	92.1	96.4	83.6	91.6	61.5	70.8
PersonViT(ours)	ViT-S/16	74.3	89.2	92.9	96.8	84.7	91.9	65.2	73.2
PersonViT(ours)	ViT-B/16	80.8	92.0	95.0	97.6	88.1	93.8	72.2	79.8

dataset LUPerson varies from ImageNet in two primary aspects: 1) human images often possess lower resolution, leading to less pronounced local granular details; 2) human body crops typically exhibit a rectangular shape, therefore during pre-training, images are scaled down to 256x128 for network input. This contrasts with the ImageNet’s practice of using square images of 224x224 pixels as input. Acknowledging these differences, the ablation experiment primarily focuses on studying the influence of size distribution and aspect ratio variance between global and local crops. To save time in experimentation, we opted for the top 3% of the LUPerson sub-dataset, as ranked by the CFS in TransReID-SSL [7], as our pre-training dataset. Further, we used the ImageNet pre-training model as initial parameters. The experiment results are presented in Table 3. Here, "crop rate range" denotes the size range of local crops relative to global ones. For example, the first entry "0.1 0.6, 0.6 1.0" implies that the ratio of the local crops input for the student network within a random interval of [0.1, 0.6], while the ratio of global crops within the random interval [0.6, 1.0].

The findings from the ablation experiment suggest that overlapping between local and global crops during self-distillation fosters characteristics more fitting for person ReID. Moreover, maintaining the aspect ratio (height: weight) at 2:1 during random jittering (the default parameter is 1:1) proves to be more beneficial. This parameter adjustment significantly enhanced the mean average precision (mAP) score by two percentage points.

Table 3: Ablation studies about the Multi-Crop Augmentation.

Input size	Crop rate range	Aspect(width/height)	Crop size	mAP	Rank-1
192x96	0.1~0.6, 0.6~1.0	3/4~4/3	96x64	60.2	80.7
192x96	0.1~0.8, 0.8~1.0	3/4~4/3	96x64	59.9	80.8
192x96	0.1~0.8, 0.4~1.0	3/4~4/3	96x64	61.3	81.9
256x128	0.1~0.8, 0.4~1.0	3/4~4/3	96x64	62.5	82.4
256x128	0.1~0.8, 0.4~1.0	3/8~2/3	96x64	64.7	83.5
256x128	0.1~0.8, 0.4~1.0	3/8~2/3	96x48	64.2	83.2
256x128	0.1~0.8, 0.4~1.0	3/8~2/3	128x64	64.4	83.3

4.3.2 MIM Loss Function Ablation Experiment

To validate the role of masked image modeling, we carried out a simple ablation experiment. Specifically, under completely identical other experimental parameters, we compared the settings of $\lambda_2 = 1$ and $\lambda_2 = 0$ in Eq. (8). The experimental results are shown in Table 4. The results indicate that the introduction of the MIM loss function has led to a significant improvement compared to the DINO algorithm [8]. The increase in mAP for MSMT7 reaches an impressive 6.4, far exceeding the increase yielded by the PASS approach [5] when compared to DINO (3.0). The visualization analysis of local feature clustering in Fig. 4 more clearly demonstrates that our method extracts richer local fine-grained human body features, thereby validating our theoretical hypothesis.

Table 4: Ablation studies about the MIM Loss.

Methods	Backbone	MSMT7		Market501		DukeMSMT		Occluded-Duke	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
PersonViT	ViT-S/16	74.3	89.2	92.9	96.8	84.7	91.9	65.2	73.2
PersonViT w/o. MIM	ViT-S/16	67.9	85.6	91.4	96.3	82.1	90.6	58.5	66.9

4.3.3 Overfitting Problem in Pre-training

During the experiment, we found that the optimal model accuracy is not necessarily achieved at $epoch = 300$, corresponding to the end of the pre-training process. By monitoring the metrics during the self-supervised pre-training phase, we noticed an overfitting phenomenon occurring in the epoch range $[200, 300]$. To further investigate this overfitting issue in self-supervised learning, we undertook supervised training and tested the accuracy changes of models saved every 20 epochs under two conditions: pre-training from scratch and initializing parameters with the ImageNet self-supervised pre-training model. These results are shown in Fig. 3, where "w/ pt" denotes training with the full LUPerson dataset starting from ImageNet pre-training model initialization, while "w/o pt" signifies pre-training starting from zero. Two conclusions can be drawn from Figure 3: 1) Initializing with the ImageNet pre-training model gives a substantial advantage in the early stages, but this advantage is overtaken around $epoch = 160$ and the pre-training from zero shows a slight edge for later epochs in the range $epoch \in [160, 300]$. The early advantage is expected as the ImageNet pre-training model already has a good discriminative capability for pedestrians. 2) For the later training phase $epoch \in [160, 300]$, both mAP and Rank-1 accuracy metrics exhibit an overfitting trend of initial increase followed by a drop. This is particularly evident in the MSMT17 dataset, where optimum performance is achieved at $epoch = 200$ while in datasets optimal point occurs at $epoch = 240$. Therefore, the supervised training accuracy of the model at pre-training $epoch = 240$ is used for most of the experimental results presented in this paper. As the amount of training data is small (only 3% of the total), there is no clear overfitting issue, hence the results in Table 3 were obtained at $epoch = 300$.

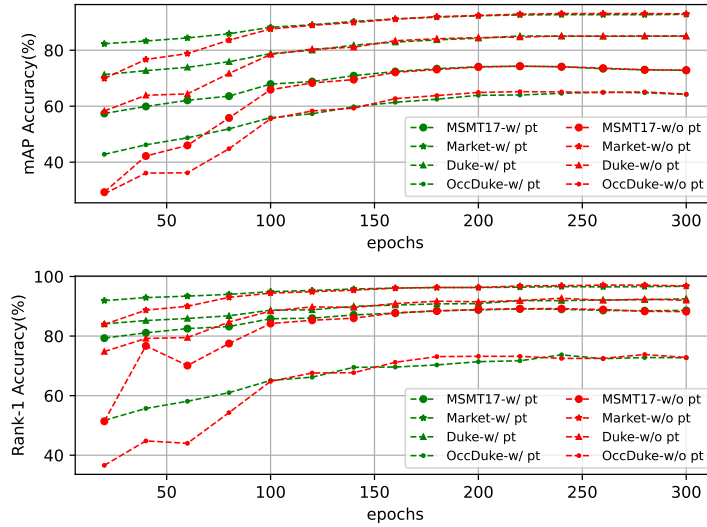


Figure 3: Supervised accuracy of Person Re-ID on the different pre-trained epochs.

4.3.4 Influence of Pre-training Data Volume

To explore further the impact of pre-training data scales on the supervised training outcome, we designed experiments with different samples from the LUPerson dataset. The results are shown in Table 5. Here, "w/ PT." denotes whether the ImageNet pre-training model was utilized to initialize parameters. "10%" represents randomly selecting 10% of the samples from LUPerson, and "10%+CFS" denotes selecting the top 10% of samples as per the CFS ranking in TransReID-SSL [7].

Analysis of Table 5 leads us to several conclusions: 1) The accuracy of person ReID through our self-supervised learning algorithm increases with the scale of pre-training data, implying that our method of learning self-supervised features is more suitable for large-scale unlabeled data. Furthermore, the larger scale of the data, the higher the ReID accuracy; 2) When the data scale is relatively small, initializing with ImageNet pre-training parameters offers significant improvement; 3) Selecting the top data according to the CFS ranking in TransReID-SSL [7] outperforms random selection, indicating a certain correlation between CFS and final accuracy. However, this selection method did not achieve the result stated in the TransReID-SSL paper, where CFS (50%) surpassed the accuracy of the full dataset.

Table 5: Ablation studies about the amount of pre-training datasets.

Backbone	w/ PT.	Scale	mAP	Rank-1
ViT-S/16	✗	3%	54.2	75.7
ViT-S/16	✗	10%	65.8	84.0
ViT-S/16	✗	25%	66.1	83.9
ViT-S/16	✗	50%	70.5	87.1
ViT-S/16	✗	100%	74.3	89.2
ViT-S/16	✓	10%	66.7	85.5
ViT-S/16	✓	10%+CFS [7]	67.5	84.9
ViT-S/16	✓	50%+CFS [7]	73.8	88.6

4.4 Visualization Analysis

To better understand the human features learned by the pre-trained model, we conducted a visualization analysis of the pre-trained ViT-S/16 model on the MSMT17 [14] dataset. This was pre-trained based on LUPerson, where we analyzed pattern layout for the patch tokens, visualized self-attention maps, and explored feature correlations.

Pattern Layout for Patch Tokens. Fig. 4 offers a visual representation of the unsupervised clustering results for the output features $y^{[patches]}$ from PersonViT. Each subgraph represents a cluster, with the red dots marking the position of the image patches that belong to the cluster. The left half of the figure portrays that the features $y^{[patches]}$ extracted by the PersonViT model can cluster key human body parts such as faces, feet, and knee joints. The right half of the figure demonstrates that the features $y^{[patches]}$ can cluster the neck of human body, backpacks, and their straps. These results substantiate that the PersonViT pre-training model can autonomously extract fine-grained features of crucial human body parts and nearby ancillary objects of the human body. This automatic extraction capability hugely contributes to improving the accuracy of person ReID. For instance, automated positional awareness of the neck could enhance the discriminability of personal ornaments like necklaces; auto-location of face and head can increase the recognition of hairstyles; auto-detection of the feet can significantly improve the recognizability of shoes.

Self-attention Visualizations. As depicted by Fig. 5, the self-attention view confirms that the pre-training model can effectively extract body contours, even in intricate situations such as occlusions, fragmentations, or misalignments of human body, by dismissing the impact of irrelevant backgrounds. This ability explains the significant improvement of our algorithm on the Occluded-Duke dataset [17].

Feature Correlation Analysis. The feature correlation assessment depicted in Fig. 6 indicates that our pre-training model can proficiently capture the relationships between features across various images of the same identity, even under significant deformations like turning around or cycling.

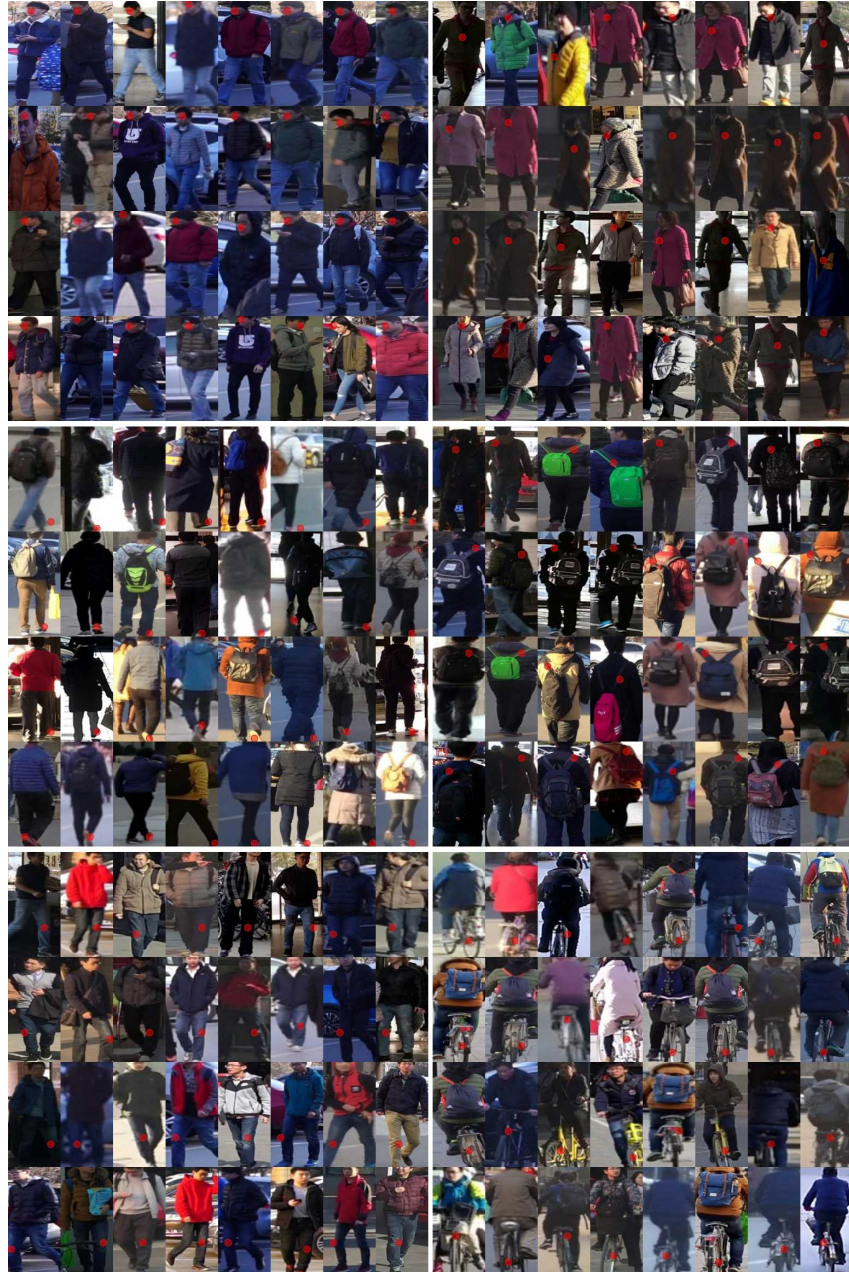


Figure 4: Visualization for pattern layout of patch tokens cluster.

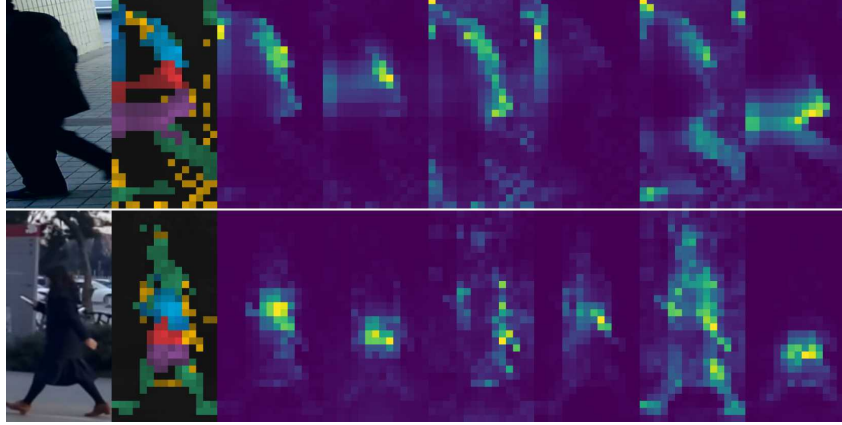


Figure 5: Visualization for self-attention map from complex background.



Figure 6: Visualization for sparse correspondence between two images of one person.

5 Conclusion

In this paper, we proposed a large-scale self-supervised person pre-training method, PersonViT, which introduces masked image modeling on the basis of contrastive learning. Through pre-training on a large-scale unlabeled human image dataset, PersonViT can effectively extract rich, highly discriminative, local fine-grained human features, and has achieved a significant accuracy improvement in person ReID tasks. Experiments show that even when the backbone network adopts a smaller vanilla ViT-S/16 model, the final recognition accuracy can further improve as the size of the pre-training dataset increases. Given that the cost of obtaining a large amount of unlabeled human pre-training data in real scenes is relatively low, this method has the potential for widespread application in practice, enhancing the effectiveness of re-identification algorithms in different scenarios.

However, like other self-supervised pre-training algorithms, PersonViT also faces the problem of high pre-training computational overhead, making the pre-training cycle longer in cases of limited computational resources. Although the frequency of updating the pre-training model for person ReID algorithms is lower than during supervised training, pre-training efficiency is still a key issue when PersonViT is extensively applied to vast amounts of unlabeled pre-training data from real-life situations. This problem can be addressed in several ways: 1) Choosing a more lightweight ViT model as the backbone network; 2) Adopting a masking method similar to MAE [12], where masked

tokens are discarded (MAE experiments demonstrate that 75% of tokens can be abandoned) and do not participate in the pre-training of the backbone network, which can improve pre-training efficiency; 3) Researching more efficient pre-training data filtration methods to reduce the size of the pre-training dataset without lowering accuracy, thus increasing pre-training efficiency; and 4) Incremental pre-training related research, meaning keeping the original features of the pre-training model and conducting online incremental pre-training learning for the additional pre-training data to update and evolve the pre-training model.

References

- [1] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” *arXiv preprint arXiv:1703.07737*, 2017.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.
- [5] K. Zhu, H. Guo, T. Yan, Y. Zhu, J. Wang, and M. Tang, “Pass: Part-aware self-supervised pre-training for person re-identification,” *arXiv preprint arXiv:2203.03931*, 2022.
- [6] D. Fu, D. Chen, J. Bao, H. Yang, L. Yuan, L. Zhang, H. Li, and D. Chen, “Unsupervised pre-training for person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14750–14759, 2021.
- [7] H. Luo, P. Wang, Y. Xu, F. Ding, Y. Zhou, F. Wang, H. Li, and R. Jin, “Self-supervised pre-training for transformer-based person re-identification,” *arXiv preprint arXiv:2111.12084*, 2021.
- [8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” *arXiv preprint arXiv:2104.14294*, 2021.
- [9] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “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.
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.
- [11] H. Bao, L. Dong, and F. Wei, “Beit: Bert pre-training of image transformers,” *arXiv preprint arXiv:2106.08254*, 2021.
- [12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.
- [13] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, “Simmim: A simple framework for masked image modeling,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.
- [14] L. Wei, S. Zhang, W. Gao, and Q. Tian, “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.
- [15] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Computer Vision, IEEE International Conference on Computer Vision*, pp. 1116–1124, 2015.
- [16] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by gan improve the person re-identification baseline in vitro,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3754–3762, 2017.
- [17] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, “Pose-guided feature alignment for occluded person re-identification,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 542–551, 2019.

- [18] L. Jing and Y. Tian, “Self-supervised visual feature learning with deep neural networks: A survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 4037–4058, 2020.
- [19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.
- [20] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” *arXiv preprint arXiv:2003.04297*, 2020.
- [21] X. Chen, S. Xie, and K. He, “An empirical study of training self-supervised vision transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021.
- [22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.
- [23] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, “Bootstrap your own latent—a new approach to self-supervised learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271–21284, 2020.
- [24] Z. Xie, Y. Lin, Z. Yao, Z. Zhang, Q. Dai, Y. Cao, and H. Hu, “Self-supervised learning with swin transformers,” *arXiv preprint arXiv:2105.04553*, 2021.
- [25] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, “Improving language understanding by generative pre-training,” 2018.
- [26] Z. Li, Z. Chen, F. Yang, W. Li, Y. Zhu, C. Zhao, R. Deng, L. Wu, R. Zhao, M. Tang, *et al.*, “Mst: Masked self-supervised transformer for visual representation,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 13165–13176, 2021.
- [27] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, “ibot: Image bert pre-training with online tokenizer,” *arXiv preprint arXiv:2111.07832*, 2021.
- [28] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabat, and N. Ballas, “Masked siamese networks for label-efficient learning,” *arXiv preprint arXiv:2204.07141*, 2022.
- [29] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, “Person re-identification in the wild,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1367–1376, 2017.
- [30] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, “Multi-scale deep learning architectures for person re-identification,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5399–5408, 2017.
- [31] W. Li, X. Zhu, and S. Gong, “Harmonious attention network for person re-identification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2285–2294, 2018.
- [32] C. Song, Y. Huang, W. Ouyang, and L. Wang, “Mask-guided contrastive attention model for person re-identification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1179–1188, 2018.
- [33] L. Zhao, X. Li, Y. Zhuang, and J. Wang, “Deeply-learned part-aligned representations for person re-identification,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3219–3228, 2017.
- [34] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, “Part-aligned bilinear representations for person re-identification,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 402–419, 2018.

- [35] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, “Learning discriminative features with multiple granularities for person re-identification,” in *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 274–282, ACM, 2018.
- [36] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, “Transreid: Transformer-based object re-identification,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2021.
- [37] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, “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.
- [38] K. Zhu, H. Guo, S. Zhang, Y. Wang, G. Huang, H. Qiao, J. Liu, J. Wang, and M. Tang, “AAformer: Auto-Aligned transformer for person re-identification,” *arXiv preprint arXiv:2104.00921*, 2021.
- [39] W. Chen, X. Xu, J. Jia, H. Luo, Y. Wang, F. Wang, R. Jin, and X. Sun, “Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15050–15061, 2023.
- [40] H. Hu, X. Dong, J. Bao, D. Chen, L. Yuan, D. Chen, and H. Li, “Personmae: Person re-identification pre-training with masked autoencoders,” *IEEE Transactions on Multimedia*, 2024.