Self-Guided Hard Negative Generation for Unsupervised Person Re-Identification

Dongdong Li[1,3] , **Zhigang Wang**[2] , **Jian Wang**[2] , **Xinyu Zhang**[2] , **Errui Ding**[2] ,
Jingdong Wang[2] , **Zhaoxiang Zhang**[1,3,4*]
[1]Institute of Automation, Chinese Academy of Sciences (CASIA), China
[2]Baidu VIS, China
[3]University of Chinese Academy of Sciences (UCAS), China
[4]Centre for Artificial Intelligence and Robotics, HKISI_CAS, China
{lidongdong2019, zhaoxiang.zhang}@ia.ac.cn
{wangzhigang05, wangjian33, zhangxinyu14, dingerrui, wangjingdong}@baidu.com

Abstract

Recent unsupervised person re-identification (re-ID) methods mostly apply pseudo labels from clustering algorithms as supervision signals. Despite great success, this fashion is very likely to aggregate different identities with similar appearances into the same cluster. As a result, the truly hard negative samples, playing an important role in training re-ID models, are significantly reduced. To alleviate this problem, we propose a self-guided hard negative generation method for unsupervised person re-ID. Specifically, a joint framework is developed which incorporates a hard negative generation network (HNGN) and a re-ID network. To continuously generate harder negative samples to provide effective supervisions in the contrastive learning, the two networks are alternately trained in an adversarial manner to improve each other. In detail, the re-ID network guides HNGN to generate challenging data, and HNGN in turn enforces the re-ID network to enhance discrimination ability. During inference, the performance of re-ID network is improved without introducing any extra parameters. Extensive experiments demonstrate that the proposed method significantly outperforms a strong baseline and also achieves better results than state-of-the-art methods.

1 Introduction

Person re-identification (re-ID) aims to retrieve images of a specific person from galleries captured by different cameras. With increasing numbers of large-scale person re-ID datasets, supervised methods [Sun *et al.*, 2018; Wang *et al.*, 2018] have achieved great success. Nevertheless, annotating person re-ID datasets across cameras is difficult and expensive. Therefore, unsupervised person re-ID methods have attracted much attention in recent years.

Without manually labeled annotations, most recent unsupervised person re-ID methods [Zeng *et al.*, 2020; Ge *et al.*, 2020; Chen *et al.*, 2021b] employ clustering algorithms to generate pseudo labels as supervision signals. However, they all suffer from the label noise caused by clustering procedure. Figure 1 illustrates such label noise, from which we can see that clustering algorithms, *e.g.* DBSCAN [Ester *et al.*, 1996], are very likely to aggregate different identities with similar appearances, *i.e.* the hard negative samples for each other. Consequently, for several identities, the truly useful hard negative samples are reduced significantly. The key target of training a re-ID model is to distinguish different identities, where easy negative samples satisfy various of losses easily, making them contribute little to the training procedure. By contrast, the hard negatives are more crucial. The decrease of hard negative samples will inevitably degrade the discrimination ability of a re-ID model.

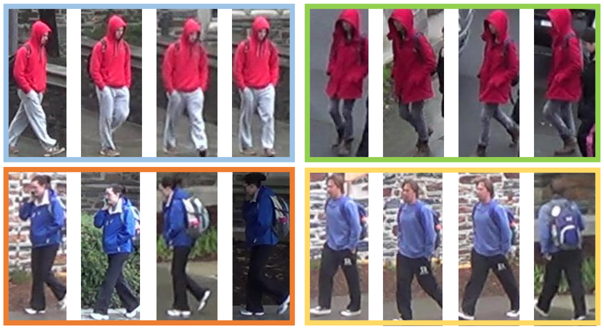

Figure 1: Illustration of noise generated by the widely used DBSCAN [Ester *et al.*, 1996] clustering algorithm. Each row denotes a cluster, while the clustered images with different color borders belong to different identities. Best viewed in color.

Several General Adversarial Network (GAN) based methods [Zhong *et al.*, 2018; Zhai *et al.*, 2020; Chen *et al.*, 2021b] attempt to enhance the intra-cluster diversity either by style transfer across cameras or by view generation for a specific person. These methods can be viewed as special data augmentations by which more positive samples are generated for anchor images. Although the lack of positives are alleviated in these methods, the hard negative samples are still lacking.

In addition, most re-ID methods adopt hard sample mining approaches [Hermans *et al.*, 2017; Yu *et al.*, 2018] to acceler-

[*] Corresponding author

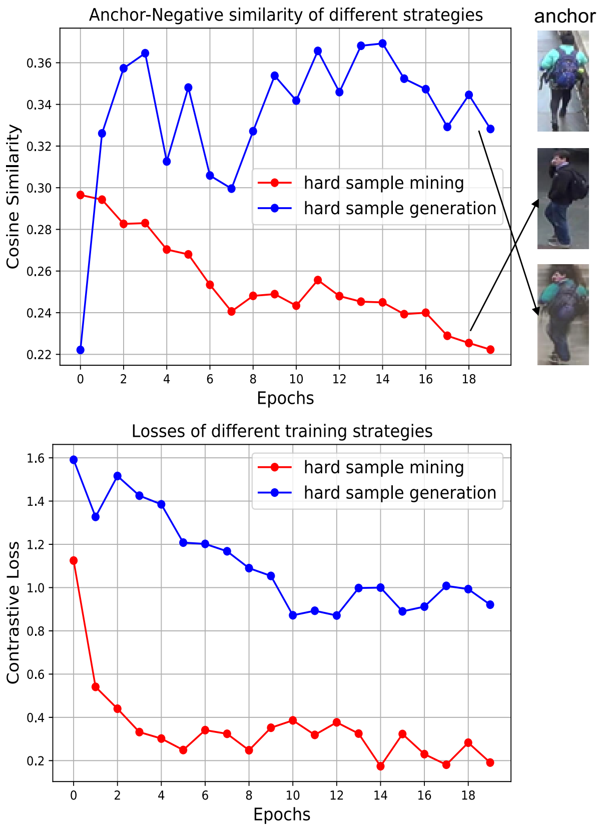

Figure 2: Changes of anchor-negative similarity (upper figure) and loss of different training strategies along with the training process. The three images are anchor, the selected hard negative and the generated negative from top to bottom.

ate metric learning during training. While hard sample mining selects samples from the existing data which may not be challenging enough. In Figure 2, it is clear that the similarity between anchor images and selected hard negatives continually decreases as the training progresses. Correspondingly, these hard negatives easily meet re-ID constraints, making loss decline rapidly. These observations indicate that the selected hard negatives may not constantly provide effective supervisions to improve the re-ID model, especially under the unsupervised setting.

To alleviate the aforementioned problems, we design a self-guided hard negative generation method for unsupervised person re-ID. Given a certain image as the anchor, to enrich its hard negatives, we propose to constantly generate hard samples rather than selecting from the existing data. To fulfill this goal, a joint framework is developed which incorporates a hard negative generation network (HNGN) and a re-ID network. These two networks are trained alternately to promote each other. Specifically, we design a novel adversarial scheme using contrastive learning, where the re-ID network guides HNGN to generate hard negatives by pulling anchor images and synthetics close in the re-ID embedding space, but conversely push them away during the training stage of the re-ID network. As for the structure of HNGN, a GAN is employed for implementation which follows an encoder-decoder pipeline. To constrain the synthetics to have proper hardness, *i.e.* the generated samples should be similar to both the anchor and the negative images, a part-level feature fusion strategy is designed to incorporate information from both of them. After then, the decoder generates hard negatives by the fused feature. As can be seen from Figure 2, the loss and anchor-negative similarity of the proposed hard negative generation method are maintained at a higher level, which can continually improve the discrimination ability of the re-ID model.

In this paper, we mainly focus on the fully unsupervised manner. There is thus no extra data or techniques, *e.g.* ID labels, attributes or pose estimation, for either hard negative generation or person re-ID training. During inference, the HNGN is discarded and the performance of the re-ID model can be improved without introducing any extra parameters.

The contributions of this paper can be summarized as follows. (1) We propose a self-guided hard negative generation concept for unsupervised person re-ID. To our knowledge, this is the first work to study hard negative generation in the unsupervised re-ID field. (2) We develop a joint framework where a hard negative generation network and a re-ID network can improve each other in an adversarial manner. (3) We propose a part-level feature fusion strategy to generate better hard negative samples.

2 Related Work

2.1 Unsupervised Person Re-ID

Unsupervised person re-ID methods can be mainly divided into two categories. The first category is fully unsupervised which does not use any ID labels. [Lin *et al.*, 2020] propose a SoftSim method to discard clustering procedure and use similarity-based soft labels to train the re-ID network. SpCL [Ge *et al.*, 2020] utilizes self-paced learning to progressively select reliable samples into clusters. Besides, [Chen *et al.*, 2021b] use 3D meshes as auxiliary information to generate more views of anchor images, which enhances the diversity of positives. [Zhang *et al.*, 2022] generate implicit samples to provide more complementary information to alleviate the clustering noise. However, they all ignore the lack of hard negative samples in the unsupervised re-ID task.

Another category is unsupervised domain adaptation (UDA) methods which aim to transfer knowledge from labeled source data to unlabeled target data. [Zhang *et al.*, 2019] provide a self-training with progressive augmentation strategy. [Bai *et al.*, 2021] introduce multiple source domains for re-ID learning. These UDA methods also do not lay emphasis on the hard-negative problem.

2.2 Hard Sample Mining and Generation

Hard sample mining methods have been studied for many years, which aim to boost machine learning performance by selecting and training hard samples. Given an anchor image, TriHard [Hermans *et al.*, 2017] selects hardest positive sample and negative sample within a batch to constitute a triplet for metric learning. HAPS [Yu *et al.*, 2018] develops a soft hard-mining scheme which will assign greater

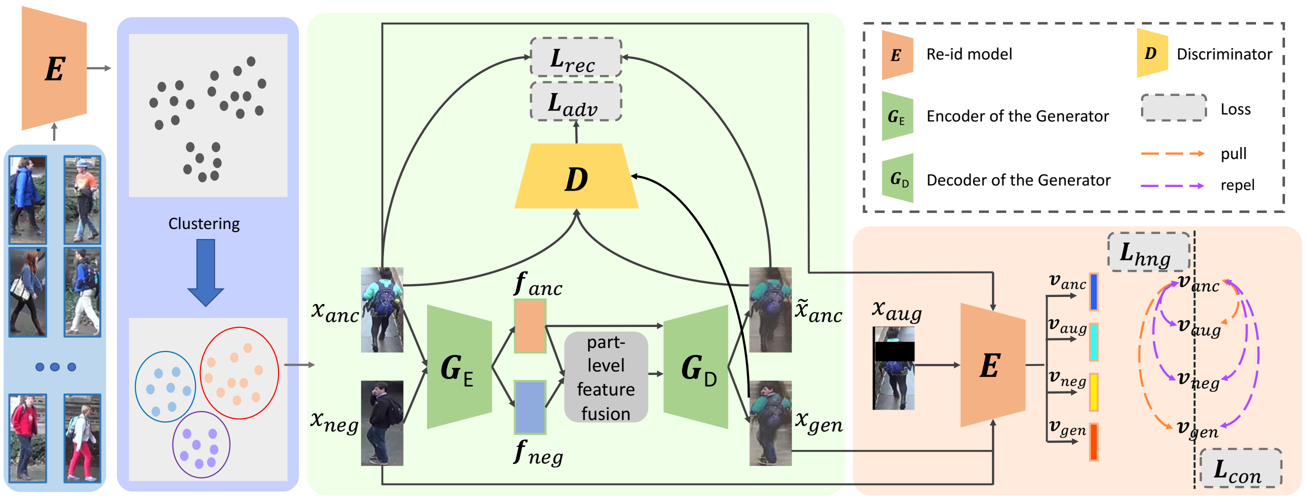

Figure 3: Illustration of the proposed framework. The colored lines are used to distinguish themselves from others. Best viewed in color. In this framework, the re-ID model first acts as a feature extractor to output features of the whole training set, and we use a clustering algorithm to assign pseudo labels to samples. These samples are fed into a generator to synthesize hard negatives, which are then used by re-ID model to learn more discriminative representations. The generator and re-ID model are optimized jointly in an adversarial manner, and the whole pipeline is trained iteratively.

weights to harder samples adaptively. All these methods rely on the hardness of the existing data. HDML [Zheng *et al.*, 2019] presents a hardness-aware framework to generate hard samples in embedding space. Different from our work, it is designed for supervised methods and trained in a non-adversarial manner.

2.3 Re-ID Related GANs

In re-ID field, many methods employ GANs for domain adaptation or data augmentation. SPGAN [Deng *et al.*, 2018] proposes a similarity-preserved GAN for style transfer between two domains. PTGAN [Wei *et al.*, 2018] uses semantic segmentation in GAN to retain the consistency of object regions during style transfer. JGCL [Chen *et al.*, 2021b] tries to disentangle appearance and structure of image, and reconstructs a synthetic one by fusing different components. None of them attempts to generate hard negative samples for re-ID.

3 Method

3.1 Baseline Revisit and Method Overview

Due to the lack of annotations, most unsupervised person re-ID methods adopt clustering algorithms to produce pseudo labels as supervisions before each epoch. The re-ID model is optimized iteratively with such a self-training scheme.

The proposed method also follows this basic pipeline as illustrated in Figure 3. While such a pipeline is difficult to distinguish different identities with similar appearances, thus causing the lack of hard negative samples. To alleviate this problem, we propose a re-ID guided generator (HNGN) to constantly synthesize hard negatives as a compensation or even enhancement. The re-ID model is then enforced to push hard negatives away from anchor images by contrastive learning. The generator and re-ID model can be optimized jointly in an adversarial manner to promote each other.

3.2 Hard Negative Generation Network

Let $\mathcal{X} = \{x^i\}_{i=1}^N$ denote a batch of training images, where the superscript i indicates the index and N is number of images in a batch. In Figure 3, the notation x_{anc} means the anchor image, x_{neg} denotes a randomly selected negative sample, $\boldsymbol{f}_{anc}$ and $\boldsymbol{f}_{neg}$ are latent embeddings of x_{anc} and x_{neg}. x_{aug} is the augmented view of the anchor. $\widetilde{x}_{anc}$ and x_{gen} denote the reconstructed anchor and the generated hard negative, respectively. $\mathcal{Y} = \{y^i\}_{i=1}^N$, y_{anc}, y_{neg}, y_{aug}, $\widetilde{y}_{anc}$ and y_{gen} indicate corresponding pseudo labels. In this paper, we usually use subscripts to indicate a type or a specific image.

The proposed hard negative generation network (generator) follows the encoder-decoder structure, where the encoder $\boldsymbol{G}_E$ maps an input image into a latent embedding and the decoder $\boldsymbol{G}_D$ is responsible for generating a synthetic image. We first use a reconstruction loss to constrain the generator can extract useful information and reconstruct the original image.

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_{i=1}^{N} \|\widetilde{x}_{anc}^i - x_{anc}^i\|_1, \tag{1}$$

where $\widetilde{x}_{anc}^i$ denotes the reconstructed image.

Then we consider that a good hard negative sample should meets two requirements. (1) It has similar appearance to the anchor. (2) It also has some negative's characters to control the hardness to be appropriate. Therefore, we attempt to feed the fused feature $h(\boldsymbol{f}_{anc}, \boldsymbol{f}_{neg})$ to the decoder $\boldsymbol{G}_D$ to generate desired hard negatives. $h(\cdot,\cdot)$ denotes a feature fusion strategy, *e.g.* concatenation or linear combination.

Specifically, in this paper, we propose a part-level feature fusion strategy for generating hard negatives. Given $\boldsymbol{f}_{neg} \in R^{C\times H\times W}$, as illustrated in Figure 4, we extract part features via a sliding window with size $(1/2H, W)$ and stride

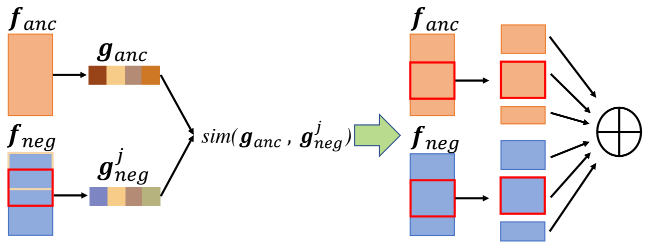

Figure 4: Illustration of our part-level feature fusion strategy.

d ($d = 2$ in our experiments). $\boldsymbol{f}_{anc}$ and all part features of $\boldsymbol{f}_{neg}$ are squeezed into feature vectors $\boldsymbol{g}_{anc}$ and $\{\boldsymbol{g}^{j}_{neg}\}^{M}_{j=1}$ by global average pooling, where M indicates the number of parts. Inspired by [Jetley *et al.*, 2018], we use cosine similarity $sim(\boldsymbol{g}_{anc}, \boldsymbol{g}^{j}_{neg})$ between a global feature vector and a part feature vector to measure which part of the negative is most similar to the anchor. Let $[h1 : h2, 0 : W]$ denotes the selected part, where $h1$ and $h2$ are the start and end vertical coordinates and W is the width of the feature map. The new feature is obtained by linearly combining part features.

$$\begin{aligned}\boldsymbol{f}_{new}[h1 : h2, 0 : W] &= \lambda \boldsymbol{f}_{anc}[h1 : h2, 0 : W] \\ &+ (1 - \lambda)\boldsymbol{f}_{neg}[h1 : h2, 0 : W],\end{aligned} \tag{2}$$

where $\boldsymbol{f}_{new}$ denotes the fused feature, $\lambda \in [0, 1]$ is a weight parameter. Other parts are fused by weights $(1 - \lambda)$ and λ conversely. We illustrate the fusion of the part $[0 : h1, 0 : W]$ as an example in Eq.(3), and the remaining part $[h2 : H, 0 : W]$ shares the same fusion format.

$$\begin{aligned}\boldsymbol{f}_{new}[0 : h1, 0 : W] &= (1 - \lambda)\boldsymbol{f}_{anc}[0 : h1, 0 : W] \\ &+ \lambda \boldsymbol{f}_{neg}[0 : h1, 0 : W].\end{aligned} \tag{3}$$

Intuitively, this strategy tries to generate hard negatives by remaining more anchor's characters on the similar parts and incorporating more negative information from other parts. We also verify the effectiveness of this strategy. See experiments for more details. For simplicity, we do not employ extra techniques, *e.g.* pose estimation, for strict feature alignment, yet the proposed strategy is also effective.

Subsequently, to enforce the synthetic image to be similar to the anchor, a hard-negative-generation loss L_{hng} is developed. The contrastive learning paradigm is employed to pull anchor image and synthetic image close in the feature space.

$$\mathcal{L}_{hng} = \frac{1}{N}\sum_{i=1}^{N} - \log \frac{\sum\limits_{p \in \mathcal{P}^i} exp(sim(\boldsymbol{v}^i, \boldsymbol{v}_p)/\tau)}{\sum\limits_{q \in \mathcal{P}^i \bigcup \mathcal{N}^i} exp(sim(\boldsymbol{v}^i, \boldsymbol{v}_q)/\tau)}, \tag{4}$$

where τ denotes a temperature parameter. $\mathcal{P}^i$ and $\mathcal{N}^i$ are the positive set and negative set for the image x^i_{anc}, where p and q indicate the image p in $\mathcal{P}^i$ and q in $\mathcal{N}^i$, respectively. $\boldsymbol{v}$ denotes a feature vector extracted by the re-ID network $\boldsymbol{E}$, *e.g.* $\boldsymbol{v}^i = \boldsymbol{E}(x^i_{anc})$, $\boldsymbol{v}_p = \boldsymbol{G}(p)$. When each image acts as an anchor, there will be an augmented image and a generated image correspondingly. Thus, we get an augmented batch $\mathcal{X}_{aug}$ and a generated batch $\mathcal{X}_{gen}$ corresponding to the original batch $\mathcal{X}$. Then, $\mathcal{P}^i$ and $\mathcal{N}^i$ can be defined as follows.

$$\begin{aligned}\mathcal{P}^i &= \{x^k_{gen} \mid y^k_{gen} = y^i_{anc}\}, \forall k \in [1, N] \\ \mathcal{N}^i &= \{\mathcal{X}_{gen} - \mathcal{P}^i, \mathcal{X}_{aug}, \mathcal{X}\},\end{aligned} \tag{5}$$

where $\mathcal{X}_{gen} - \mathcal{P}^i$ means the difference set of $\mathcal{X}_{gen}$ and $\mathcal{P}^i$. Note that, we treat the synthetic image of x^i_{gen} as positive, and treat the whole batch $\mathcal{X}$ and $\mathcal{X}_{aug}$ as negative set in $\mathcal{L}_{hng}$. This can be viewed as a re-ID guided strong constraint to make the synthetic image similar to the anchor.

Like many GAN-based methods, we employ a discriminator $\boldsymbol{D}$ to distinguish between real images and the synthetics, and use an adversarial loss L_{adv} to constrain the distribution of synthetics to be close to the real data as follows:

$$\mathcal{L}_{adv} = \frac{1}{3N}\sum_{i=1}^{N}((\boldsymbol{D}\left(x^i_{anc}\right) - 1)^2 + \boldsymbol{D}(\widetilde{x}^i_{anc})^2 + \boldsymbol{D}(x^i_{gen})^2), \tag{6}$$

where $D(.)$ measures the probability that an input image belongs to the real data.

3.3 Re-ID Network

After generating hard negative samples, the re-ID model is enforced to identify them with a contrastive loss.

$$\mathcal{L}_{con} = \frac{1}{N}\sum_{i=1}^{N} - \log \frac{\sum\limits_{p \in \mathcal{P}^i_{con}} exp(sim(\boldsymbol{v}^i, \boldsymbol{v}_p)/\tau)}{\sum\limits_{q \in \mathcal{P}^i_{con} \bigcup \mathcal{N}^i_{con}} exp(sim(\boldsymbol{v}^i, \boldsymbol{v}_q)/\tau)}, \tag{7}$$

where $\mathcal{P}^i_{con}$ and $\mathcal{N}^i_{con}$ are the positive set and negative set of image x^i_{anc}. They are defined as follows.

$$\begin{aligned}\mathcal{P}^i_{con} &= \{x^k \mid y^k = y^i_{anc},\ \ x^k_{aug} \mid y^k_{aug} = y^i_{anc}\}, \forall k \in [1, N] \\ \mathcal{N}^i_{con} &= \{\mathcal{X} + \mathcal{X}_{aug} - \mathcal{P}^i_{con}, \mathcal{X}_{gen}, \mathcal{X}_{pa}\},\end{aligned} \tag{8}$$

where $+$ denotes the merge operation between two sets. $\mathcal{X}_{pa}$ indicates a batch of synthetic image parts selected in Eq. 2. The contrastive learning between anchors and generated hard negatives (including negative parts) will constantly encourage the re-ID model to learn more discriminative features.

3.4 Optimization of the Whole Framework

The model is jointly trained with the following overall loss $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{con} + \mathcal{L}_{hng} + \mathcal{L}_{rec} + \mathcal{L}_{adv}. \tag{9}$$

In addition, the proposed method is Plug-and-Play. It can be applied to most unsupervised person re-ID works, and the original losses can also be retained.

Although this is a joint framework, we find that it is difficult to convergence when trained from scratch. This is because the two networks, *i.e.* the generator and the re-ID model, may confuse each other when they do not have any capability. Thus, we use a warm-up stage to improve their performance individually before joint training. During warm-up, the re-ID model is trained under the standard setting, and the generator is trained only using $\mathcal{L}_{rec}$ to make it have the basic reconstruction ability.

Methods		Market-1501				DukeMTMC-ReID			
		mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10
SoftSim [Lin *et al.*, 2020]	CVPR'20	37.8	71.7	83.8	87.4	28.6	52.5	63.5	68.9
MMCL [Wang and Zhang, 2020]	CVPR'20	45.5	80.3	89.4	92.3	40.2	65.2	75.9	80.0
JVTC+ [Li and Zhang, 2020]	ECCV'20	47.5	79.5	89.2	91.9	50.7	74.6	82.9	85.3
HCT [Zeng *et al.*, 2020]	CVPR'20	56.4	80.0	91.6	95.2	50.7	69.6	83.4	87.4
IICS [Xuan and Zhang, 2021]	CVPR'21	72.1	88.8	95.3	96.9	59.1	76.9	86.1	89.8
CycAs [Wang *et al.*, 2020]	ECCV'20	64.8	84.8	-	-	60.1	77.9	-	-
JGCL [Chen *et al.*, 2021b]	CVPR'21	66.8	87.3	93.5	95.5	62.8	82.9	87.1	88.5
CAP [Wang *et al.*, 2021]	AAAI'21	79.2	91.4	96.3	97.7	67.3	81.1	89.3	91.8
baseline	-	70.5	87.9	95.7	97.1	54.7	72.9	83.5	87.2
baseline+Ours	-	75.9	89.3	95.9	97.3	65.4	80.1	88.9	91.8
SpCL* [Ge *et al.*, 2020]	NeurIPS'20	76.0	89.5	96.2	97.5	67.1	82.4	90.8	93.0
SpCL*+Ours	-	78.3	91.1	96.8	98.0	70.0	82.8	**91.3**	**93.8**
ICE†[Chen *et al.*, 2021a]	ICCV'21	82.4	93.5	97.5	98.4	69.2	82.5	90.7	92.8
ICE†+Ours	-	**83.5**	**94.1**	**97.7**	**98.6**	**70.1**	**83.3**	91.1	93.2

Methods		MSMT17			
		mAP	top-1	top-5	top-10
JVTC+ [Li and Zhang, 2020]	ECCV'20	17.3	43.1	53.8	59.4
IICS [Xuan and Zhang, 2021]	CVPR'21	18.6	45.7	57.7	62.8
JGCL [Chen *et al.*, 2021b]	CVPR'21	21.3	45.7	58.6	64.5
CAP [Wang *et al.*, 2021]	AAAI'21	36.9	67.4	78.0	81.4
baseline	-	16.6	38.8	51.6	57.2
baseline+Ours	-	24.2	50.5	63.0	68.1
SpCL* [Ge *et al.*, 2020]	NeurIPS'20	20.5	44.4	57.3	63.0
SpCL*+Ours	-	23.6	48.7	60.9	66.4
ICE†[Chen *et al.*, 2021a]	ICCV'21	39.6	71.9	82.0	85.2
ICE†+Ours	-	**42.2**	**73.1**	**82.4**	**85.7**

Table 1: Comparison with state-of-the-art unsupervised person re-ID methods. (*) indicates the results of an enhanced version. †indicates the reproduced result by running the source code in our training environment, where we obtain a slightly worse result on Market-1501 and DukeMTMC-ReID, but achieve a slightly better result on MSMT17.

4 Experiment

4.1 Datasets and Evaluation Protocols

We conduct experiments on three widely-used person re-ID datasets, *i.e*. Market-1501 [Zheng *et al.*, 2015], DukeMTMC-ReID [Zheng *et al.*, 2017] and MSMT17 [Wei *et al.*, 2018] to validate our method. Following most re-ID methods, we use Cumulative Matching Characteristic (CMC) top-k accuracy and mean average precision (mAP) as evaluation metrics.

Market-1501 consists of 32,668 images of 1501 persons captured by 6 cameras, where 12,936 images of 751 persons constitute the training set. The remaining images are for test.

DukeMTMC-ReID contains 36,411 images of 1,404 persons from 8 cameras in total. Each person is captured by at least 2 cameras. The training set is comprised of 16,522 images of 702 persons. Other images constitute the test set.

MSMT17 is a challenging dataset containing 126,441 images of 4,101 persons. These images are captured by 15 cameras including 32,621 training images of 1,041 persons and 93,820 test images of 3,060 persons.

4.2 Implementation Details

We implement our method mainly based on the open-source codes released by [Ge *et al.*, 2020] and [Chen *et al.*, 2021a]. ResNet50 is adopted as the backbone of the re-ID network, which is initialized by ImageNet pre-trained parameters. The original fully connected layer is removed and a batch normalization (BN) layer is stacked onto the backbone to obtain normalized features. The encoder of our generator stacks 3 conv-norm-relu layers and 3 residual blocks sequentially. The detailed structure of a residual block is the same as CycleGAN [Zhu *et al.*, 2017]. The decoder stacks 3 residual blocks, 2 deconv-norm-relu layers and 1 convolution layer sequentially. $\boldsymbol{D}$ in this paper is a PatchGAN [Isola *et al.*, 2017] discriminator with 3 conv-norm-leakyrelu layers.

During training, images are resized to 256 × 128. Random flip, random erasing and image padding are adopted as data augmentations. Following most re-ID methods, we use identity-based sampling strategy in our training, which first randomly selects 16 identities and further randomly selects 4 instances for each identity to constitute a mini-batch of size 64. We use Adam optimizer to train our generator and discriminator. The optimizers for re-ID model follow the original settings of [Ge *et al.*, 2020] and [Chen *et al.*, 2021a]. The learning rate is initialized as 1×10^{-4} for Adam, and the weight decay is set as 5×10^{-4}. As aforementioned, the proposed method has two training stages, *i.e*. warm-up stage and

Method	DukeMTMC-ReID			
	mAP	Top-1	Top-5	Top-10
baseline	54.7	72.9	83.5	87.2
mixup	46.3	66.6	79.5	82.8
random noise	61.1	76.0	86.7	89.8
concatenation	62.3	76.5	87.2	90.2
global-fusion	63.1	78.1	88.1	90.9
part-fusion	**64.4**	**78.5**	**88.5**	**91.3**

Table 2: Ablation studies on hard negative generation strategies.

joint fine-tune stage. For the former, we train the re-ID model following the settings of a basic unsupervised re-ID method. Besides, we train the generator with only reconstruction loss for 50 epochs, where the learning rate is multiplied by 0.1 after 20 epochs. For the joint fine-tune stage, the whole framework is fine-tuned for 20 epochs, where the learning rate is multiplied by 0.1 after 10 epochs. The extra training procedure can be completed in a few hours by modern GPUs, *e.g.* NVIDIA P40. The temperature τ in Eq. 4 and Eq. 7 is set to 0.05. The widely-used DBSCAN is employed as the clustering algorithm in our method.

4.3 Comparison with state-of-the-arts

To verify the effect of our method, we first integrate our method into a strong baseline provided by [Ge *et al.*, 2020] and compare with many recent works. The results are shown in Table 1. We can see that when equipped with our method, the performance of baseline is improved remarkably (DukeMTMC-ReID mAP +10.7%, top-1 +7.2%; Market-1501 mAP +5.4%, top-1 +1.4%; MSMT17 mAP +7.6%, top-1 +11.7%). It has already achieved comparable results to many recent methods. When applying our method to more powerful approaches [Ge *et al.*, 2020] and [Chen *et al.*, 2021a], we get better results, and outperform all the comparison unsupervised person re-ID works. This indicates the effectiveness and robustness of our method.

4.4 Ablation Studies

To further validate several components proposed in this paper, extensive ablation experiments are conducted.

The Effectiveness of Hard Negative Generation Strategy. We propose to fuse latent embeddings of the anchor and the negative image. The decoder then takes the fused feature as input and outputs a hard negative sample. To validate the effectiveness of this pipeline, we compare it with several feasible generation strategies. The results are shown in Table 2. *Mixup* means linearly fusing the re-ID feature vectors of the anchor and a negative sample, which is then employed as a new negative sample for unsupervised contrastive learning. Without constraint on the fused feature, the performance of *mixup* strategy declines drastically. *Random noise* denotes that we add Gaussian random noise to the latent embedding of an anchor and feed it to the decoder to generate hard negative samples. Under our joint framework, *random noise* strategy has already outperforms the baseline by a large margin. We also explore the *concatenation* feature fusion strategy where the latent embeddings of two images are concatenated along the channel axis. This strategy achieves a better result than *random noise*. The *global-fusion* and *part-fusion* strategies can control how much information of anchors and negative references are used, thus generating better hard negatives and further improving the re-ID performance.

Methods	DukeMTMC-ReID			
	mAP	top-1	top-5	top-10
w/o warm-up	42.8	62.9	74.3	78.0
w/o gen warm-up	63.1	77.5	87.3	90.7
w/o re-ID warm-up	56.2	72.8	83.4	86.1
warm-up	**65.4**	**80.1**	**88.9**	**91.8**

Methods	Market-1501			
	mAP	top-1	top-5	top-10
w/o warm-up	46.1	68.5	84.2	88.5
w/o gen warm-up	72.1	88.2	95.8	97.3
w/o re-ID warm-up	61.3	79.6	90.3	93.3
warm-up	**75.9**	**89.3**	**95.9**	**97.3**

Table 3: Ablation studies on the warm-up stage. 'gen' denotes the generator in this table.

The Necessity of the Warm-up Stage. Although the proposed method can be trained jointly, the training procedure is difficult to convergence when initializing the re-ID model and generator from scratch. Under this situation, the generator cannot synthesize samples with high quality, and the re-ID model cannot benefit from the hard negative generation. To address this problem, we use a warm-up stage to make the re-ID model and generator have a better initialization. Ablation study experiments are carried out on DukeMTMC-ReID and Market-1501 datasets to verify the effectiveness of the warm-up stage. Results are listed in Table 3. We can see that if both two networks do not have a warm-up stage, only an inferior result can be achieved. In this case, the generator synthesize some low-quality samples, which may harm the performance of a re-ID model. We gain the best result when two networks all go through the warm-up stage, where they can promote each other based on a better initialization state.

5 Conclusion

In this paper, we propose a self-guided hard negative generation method for unsupervised person re-ID to alleviate the lack of hard negative samples. The proposed method does not rely on any extra data or techniques. We further develop a joint framework which consists of a generator and a re-ID model. These two networks are trained in an adversarial manner to promote each other. The re-ID model guides the generator to synthesize hard negatives, and the hard negatives enforce the re-ID model to learn more discriminative features. Moreover, we propose a part-level feature fusion strategy to make a hard negative similar to both the anchor image and the negative reference. Extensive experiments demonstrate the effectiveness and robustness of our method.

Acknowledgements

This work was supported in part by the Major Project for New Generation of AI (No.2018AAA0100400), the National Natural Science Foundation of China (No. 61836014, No. U21B2042, No. 62072457, No. 62006231).

References

[Bai *et al.*, 2021] Zechen Bai, Zhigang Wang, Jian Wang, Di Hu, and Errui Ding. Unsupervised multi-source domain adaptation for person re-identification. In *CVPR*, 2021.

[Chen *et al.*, 2021a] Hao Chen, Benoit Lagadec, and François Bremond. Ice:inter-instance contrastive encoding for unsupervised person re-identification. In *ICCV*, 2021.

[Chen *et al.*, 2021b] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. *CVPR*, 2021.

[Deng *et al.*, 2018] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018.

[Ester *et al.*, 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

[Ge *et al.*, 2020] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, 2020.

[Hermans *et al.*, 2017] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[Jetley *et al.*, 2018] Saumya Jetley, Nicholas A. Lord, Namhoon Lee, and Philip H. S. Torr. Learn to pay attention. In *ICLR*, 2018.

[Li and Zhang, 2020] Jianing Li and Shiliang Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *ECCV*, 2020.

[Lin *et al.*, 2020] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *CVPR*, 2020.

[Sun *et al.*, 2018] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. In *ECCV*, 2018.

[Wang and Zhang, 2020] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *CVPR*, 2020.

[Wang *et al.*, 2018] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 2018.

[Wang *et al.*, 2020] Zhongdao Wang, Jingwei Zhang, Liang Zheng, Yixuan Liu, Yifan Sun, Yali Li, and Shengjin Wang. Cycas: Self-supervised cycle association for learning re-identifiable descriptions. In *ECCV*, 2020.

[Wang *et al.*, 2021] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. In *AAAI*, 2021.

[Wei *et al.*, 2018] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[Xuan and Zhang, 2021] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *CVPR*, 2021.

[Yu *et al.*, 2018] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification. In *ECCV*, 2018.

[Zeng *et al.*, 2020] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *CVPR*. IEEE, 2020.

[Zhai *et al.*, 2020] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *CVPR*, 2020.

[Zhang *et al.*, 2019] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *ICCV*, 2019.

[Zhang *et al.*, 2022] Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. Implicit sample extension for unsupervised person re-identification. *arXiv preprint arXiv:2204.06892*, 2022.

[Zheng *et al.*, 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[Zheng *et al.*, 2017] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[Zheng *et al.*, 2019] Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and Jie Zhou. Hardness-aware deep metric learning. In *CVPR*, 2019.

[Zhong *et al.*, 2018] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *ICCV*, 2017.