Joint Generative and Contrastive Learning for Unsupervised Person Re-identification

Hao Chen[1,2,3]* Yaohui Wang[1,2]* Benoit Lagadec[3] Antitza Dantcheva[1,2] Francois Bremond[1,2]
[1]Inria [2]Université Côte d'Azur [3]European Systems Integration
{hao.chen, yaohui.wang, antitza.dantcheva, francois.bremond}@inria.fr benoit.lagadec@esifrance.net

Abstract

Recent self-supervised contrastive learning provides an effective approach for unsupervised person re-identification (ReID) by learning invariance from different views (transformed versions) of an input. In this paper, we incorporate a Generative Adversarial Network (GAN) and a contrastive learning module into one joint training framework. While the GAN provides online data augmentation for contrastive learning, the contrastive module learns view-invariant features for generation. In this context, we propose a mesh-based view generator. Specifically, mesh projections serve as references towards generating novel views of a person. In addition, we propose a view-invariant loss to facilitate contrastive learning between original and generated views. Deviating from previous GAN-based unsupervised ReID methods involving domain adaptation, we do not rely on a labeled source dataset, which makes our method more flexible. Extensive experimental results show that our method significantly outperforms state-of-the-art methods under both, fully unsupervised and unsupervised domain adaptive settings on several large scale ReID datasets. Source code and models are available under https://github.com/chenhao2345/GCL.

1. Introduction

A person re-identification (ReID) system is targeted at identifying subjects across different camera views. In particular, given an image containing a person of interest (as query) and a large set of images (gallery set), a ReID system ranks gallery-images based on visual similarity with the query. Towards this, ReID systems are streamlined to bring to the fore discriminative representations, which allow for robust comparison of query and gallery images. In this context, *supervised* ReID methods [4, 33] learn representations guided by human-annotated labels, which is time-consuming and cumbersome. Towards omitting such human annotation, researchers increasingly place emphasis on *unsupervised* person ReID algorithms [35, 24, 27], which learn directly from unlabeled images and thus allow for scalability in real world deployments.

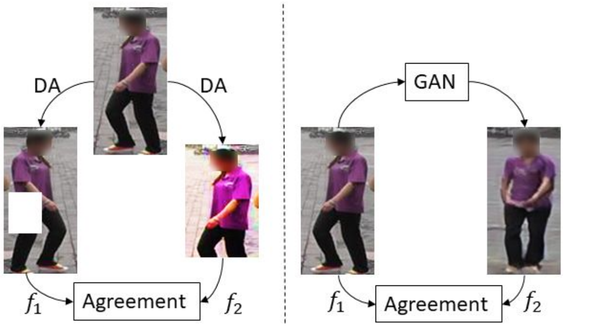

Figure 1: **Left**: Traditional self-supervised contrastive learning maximizes agreement between representations (f_1 and f_2) of augmented views from Data Augmentation (DA). **Right**: Joint generative and contrastive learning maximizes agreement between original and generated views.

Recently, self-supervised contrastive methods [16, 6] have provided an effective retrieval-based approach for unsupervised representation learning. Given an image, such methods maximize agreement between two augmented views of one instance (see Fig. 1). *Views* refer to transformed versions of the same input. As shown in very recent works [6, 7], data augmentation enables a network to explore view-invariant features by providing augmented views of a person, which are instrumental in building robust representations. Such and similar methods considered traditional data augmentation techniques, *e.g.*, 'random flipping', 'cropping', and 'color jittering'. Generative Adversarial Networks (GANs) [15] constitute a novel approach for data augmentation. As opposed to traditional data augmentation, GANs are able to modify id-unrelated features substantially, while preserving id-related features, which is highly beneficial in contrastive ReID.

Previous GAN-based methods [1, 9, 54, 25, 41, 49] considered unsupervised ReID as an unsupervised domain adaptation (UDA) problem. Under the UDA setting, researchers used both, a labeled source dataset, as well as an unlabeled target dataset to gradually adjust a model from a source domain into a target domain. GANs can be used in cross-domain style transfer, where labeled source domain images are generated in the style of a target domain. However, the UDA setting necessitates a large-scale labeled source dataset. Scale and quality of the source dataset

[*]Equal contribution.

strongly affect the performance of UDA methods. Recent research has considered fully unsupervised ReID [35, 24], where under the fully unsupervised setting, a model directly learns from unlabeled images without any identity labels. Self-supervised contrastive methods [16, 6] belong to this category. In this work, we use a GAN as a novel view generator for contrastive learning, which does not require a labeled source dataset.

Here, we aim at enhancing view diversity for contrastive learning via generation under the fully unsupervised setting. Towards this, we introduce a mesh-based novel view generator. We explore the possibility of disentangling a person image into identity features (color distribution and body shape) and structure features (pose and view-point) under the fully unsupervised ReID setting. We estimate 3D meshes from unlabeled training images, then rotate these 3D meshes to simulate new structures. Compared to skeleton-guided pose transfer [14, 25], which neglects body shape, mesh recovery [21] jointly estimates pose and body shape. Estimated meshes preserve body shape during the training, which facilitates the generation and provides more visual clues for fine-grained ReID. Novel views can be generated by combining identity features with new structures.

Once we obtain the novel views, we design a pseudo label based contrastive learning module. With the help of our proposed view-invariant loss, we maximize representation similarity between original and generated views of a same person, whereas representation similarity of other persons is minimized.

Our proposed method incorporates generative and contrastive modules into one framework, which are trained jointly. Both modules share the same identity feature encoder. The generative module disentangles identity and structure features, then generates diversified novel views. The novel views are then used in the contrastive module to improve the capacity of the shared identity feature encoder, which in turn improves the generation quality. Both modules work in a mutual promotion way, which significantly enhances the performance of the shared identity feature encoder in unsupervised ReID. Moreover, our method is compatible with both UDA and fully unsupervised settings. With a labeled source dataset, we obtain better performance by alleviating the pseudo label noise.

Our contributions can be summarized as follows.

1. We propose a joint generative and contrastive learning framework for unsupervised person ReID. Generative and contrastive modules mutually promote each other's performance.

2. In the generative module, we introduce a 3D mesh based novel view generator, which is more effective in body shape preservation than skeleton-guided generators.

3. In the contrastive module, a view-invariant loss is proposed to reduce intra-class variation between original and generated images, which is beneficial in building view-invariant representations under a fully unsupervised ReID setting.

4. We overcome the limitation of previous GAN-based unsupervised ReID methods that strongly rely on a labeled source dataset. Our method significantly surpasses the performance of state-of-the-art methods under both, fully unsupervised, as well as UDA settings.

2. Related Work

Unsupervised representation learning. Recent contrastive instance discrimination methods [44, 16, 6] have witnessed a significant progress in unsupervised representation learning. The basic idea of instance discrimination has to do with the assumption that each image is a single class. Contrastive predictive coding (CPC) [30] included an InfoNCE loss to measure the ability of a model to classify positive representation amongst a set of unrelated negative samples, which has been commonly used in following works on contrastive learning. Recent contrastive methods treated unsupervised representation learning as a retrieval task. Representations can be learnt by matching augmented views of a same instance from a memory bank [44, 16] or a large mini-batch [6]. MoCoV2 [7] constitutes the improved version of the MoCo [16] method, incorporating larger data augmentation. We note that data augmentation is pertinent in allowing a model to learn robust representations in contrastive learning. However, only traditional data augmentation was used in aforementioned methods.

Data augmentation. MoCoV2 [7] used 'random crop', 'random color jittering', 'random horizontal flip', 'random grayscale' and 'gaussian blur'. However, 'random color jittering' and 'grayscale' were not suitable for fine-grained person ReID, because such methods for data augmentation tend to change the color distribution of original images. In addition, 'Random Erasing' [48] has been a commonly used technique in person ReID, which randomly erases a small patch from an original image. Cross-domain Mixup [29] interpolated source and target domain images, which alleviated the domain gap in UDA ReID. Recently, Generative Adversarial Networks (GANs) [15] have shown great success in image [23, 22, 2] and video synthesis [34, 37, 3, 39, 38]. GAN-based methods can serve as a method for evolved data augmentation by conditionally modifying id-unrelated features (style and structure) for supervised ReID. CamStyle [52] used the CycleGAN-architecture [53] in order to transfer images from one camera into the style of another camera. FD-GAN [14] was targeted to generate images in a pre-defined pose, so that images could be compared in the same pose. IS-GAN [10] was streamlined to disentangle id-related and id-unrelated features by switching both local and global level identity features. DG-Net [47] recolored grayscale images with a color distribution of other images, targeting to disentangle identity features. Deviating from such supervised GAN-based methods, our method generates novel views by rotating 3D meshes in an *unsupervised* manner.

Unsupervised person ReID. Recent unsupervised person ReID methods were predominantly based on UDA. Among UDA-based methods, several works [36, 26] used semantic attributes to facilitate domain adaptation. Other works [43, 12, 5, 45, 13] assigned pseudo labels to unlabeled images and proceeded to learn representations with pseudo labels. Transferring source dataset images into the style of a target dataset represents another line of research. SPGAN [9] and PTGAN [41] used CycleGAN [53] as domain style transfer-backbone. HHL [49] aims at transferring cross-dataset camera styles. ECN [50, 51] exploited invariance from camera style transferred images for UDA ReID. CR-GAN [8] employed parsing-based masks to remove noisy backgrounds. PDA [25] included skeleton estimation to generate person images with different poses and cross-domain styles. DG-Net++ [54] jointly disentangled id-related/id-unrelated features and transferred domain styles. While the latter is related to our our method, we aim at training jointly a GAN-based online data augmentation, as well as a contrastive discrimination, which renders the labeled source dataset unnecessary, rather than transferring style.

Fully unsupervised methods do not require any identity labels. BUC [27] represented each image as a single class and gradually merged classes. In addition, TSSL [42] considered each tracklet as a single class to facilitate cluster merging. SoftSim [28] utilized similarity-based soft labels to alleviate label noise. MMCL [35] assigned multiple binary labels and trained a model in a multi-label classification way. JVTC and JVTC+ [24] added temporal information to refine visual similarity based pseudo labels. We note that all aforementioned fully unsupervised methods learn from pseudo labels. We show in this work that disentangling view-invariant identity features is possible in fully unsupervised ReID, which can be an add-on to boost the performance of previous pseudo label based methods.

3. Proposed Method

We refer to our proposed method as joint *Generative and Contrastive Learning* as GCL. The general architecture of GCL comprises of two modules, namely a View Generator, as well as a View Contrast Module, see Fig. 2. Firstly, the View Generator uses cycle-consistency on both, image and feature reconstructions in order to disentangle identity and structure features. It combines identity features and mesh-guided structure features to generate one person in new view-points. Then, original and generated views are exploited as positive pairs in the View Contrast Module, which enables our network to learn view-invariant identity features. We proceed to elaborate on both modules in the following.

3.1. View Generator (Generative Module)

As shown in Fig. 2, the proposed View Generator incorporates 4 networks: an identity encoder E_{id}, a structure encoder E_{str}, a decoder G and an image discriminator D. Given an unlabeled person ReID dataset $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, we generate corresponding 3D meshes with a popular 3D mesh generator Human Mesh Recovery (HMR) [21], which simultaneously estimates body shape and pose from a single RGB image. Here, we denote the 2D projection of a 3D mesh as original structure s_{ori}. Then, as depicted in Fig. 3, we rotate each 3D mesh by $45^{\circ}, 90^{\circ}, 135^{\circ}, 180^{\circ}, 225^{\circ}, 270^{\circ}$ and 315°, respectively and proceed to randomly pick one 2D projection of these rotated meshes as a new structure s_{new}. We use the 3D mesh rotation to mimic view-point variance from different cameras. Next, unlabeled images are encoded to identity features by the identity encoder $E_{id} : x \to f_{id}$, while both original and new structures are encoded to structure features by the structure encoder $E_{str} : s_{ori} \to f_{str(ori)}, s_{new} \to f_{str(new)}$. Combining both, identity and structure features, the decoder generates synthesized images $G : (f_{id}, f_{str(ori)}) \to x'_{ori}, (f_{id}, f_{str(new)}) \to x'_{new}$, where a prime is used to represent generated images.

Given the lack of real images corresponding to the new structures, we consider a cycle consistency [53] to reconstruct the original image by swapping the structure features in the View Generator. We encode and decode once again to get synthesized images in original structures $G(E_{id}(x'_{new}), s_{ori}) \to x''_{ori}$. We calculate an image reconstruction loss as follows.

$$\mathcal{L}_{img} = \mathbb{E}[\|x - x'_{ori}\|_1] + \mathbb{E}[\|x - x''_{ori}\|_1] \tag{1}$$

In addition, we compute a feature reconstruction loss

$$\begin{aligned}\mathcal{L}_{feat} = &\mathbb{E}[\|f_{id} - E_{id}(x'_{new})\|_1] + \\ &\mathbb{E}[\|f_{id} - E_{id}(x''_{ori})\|_1].\end{aligned} \tag{2}$$

The discriminator D attempts to distinguish between real and generated images with the adversarial loss

$$\begin{aligned}\mathcal{L}_{adv} = &\mathbb{E}[\log D(x) + \log(1 - D(x'_{ori}))] + \\ &\mathbb{E}[\log D(x) + \log(1 - D(x'_{new}))] + \\ &\mathbb{E}[\log D(x) + \log(1 - D(x''_{ori}))].\end{aligned} \tag{3}$$

Consequently, the overall GAN loss combines the above named losses with weighting coefficients λ_{img} and λ_{feat}

$$\mathcal{L}_{gan} = \lambda_{img}\mathcal{L}_{img} + \lambda_{feat}\mathcal{L}_{feat} + \mathcal{L}_{adv}. \tag{4}$$

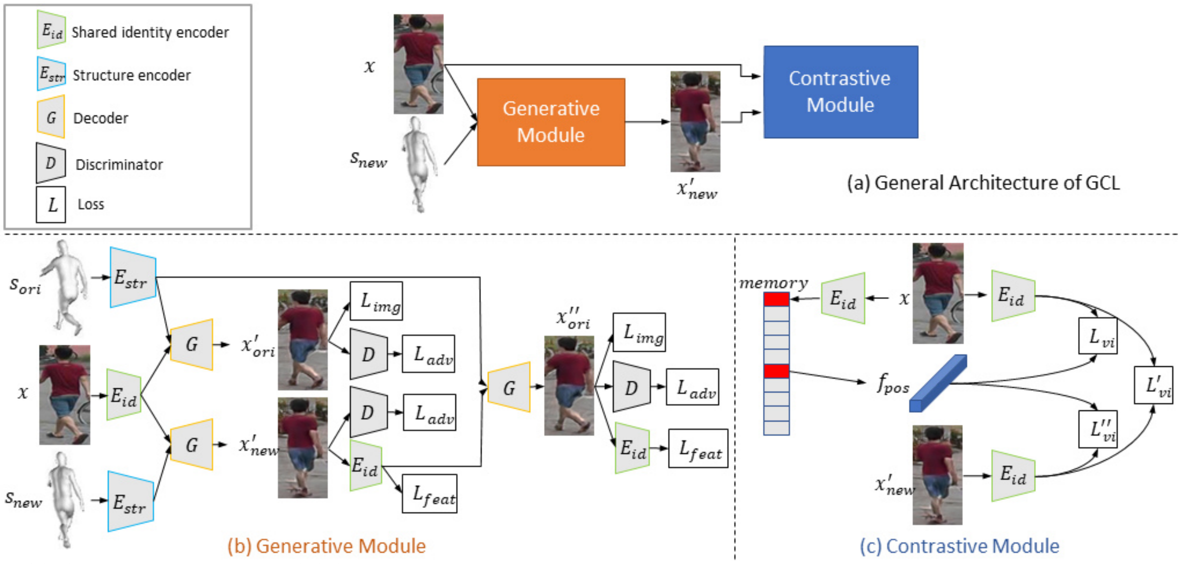

Figure 2: **(a) General architecture of GCL**: Generative and contrastive modules are coupled by the shared identity encoder E_{id}. **(b) Generative module**: The decoder G combines the identity features encoded by E_{id} and structure features E_{str} to generate a novel view x'_{new} with a cycle consistency. **(c) Contrastive module**: View-invariance is enhanced by maximizing the agreement between original $E_{id}(x)$, synthesized $E_{id}(x'_{new})$ and memory f_{pos} representations.

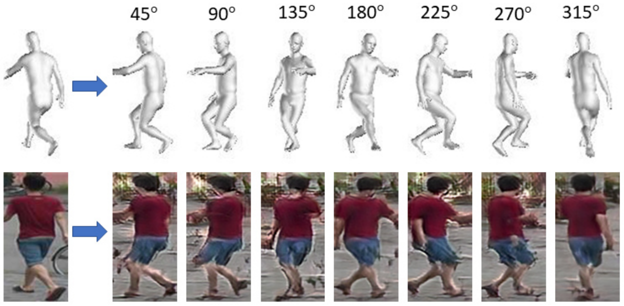

Figure 3: Example images as generated by the View Generator via 3D mesh rotation based on left input image.

3.2. View Contrast (Contrastive Module)

The previous reconstruction and adversarial losses work in an unconditional manner. They only explore identity features within the original view-point, which renders appearance representations view-variant. In rotating an original mesh to a different view-point, *e.g.*, from front to side view-point, the generation is prone to fail due to lack of information pertained to the side view. This issue can be alleviated by enhancing the view-invariance of representations.

Given an anchor image x, the first step is to find positive images that belong to the same identity and negative images that belong to different identities. Here, we store all instance representations in a memory bank [44], which stabilizes pseudo labels and enlarges the number of negatives during the training with mini-batches. The memory bank $\mathcal{M}$ is updated with a momentum coefficient α.

$$\mathcal{M}[i]^t = \alpha \cdot \mathcal{M}[i]^{t-1} + (1-\alpha) \cdot f^t \tag{5}$$

where $\mathcal{M}[i]^t$ and $\mathcal{M}[i]^{t-1}$ respectively refer to the identity feature vector in the t and $t-1$ epochs.

We use a clustering algorithm DBSCAN [11] on all memory bank feature vectors to generate pseudo identity labels $\mathcal{Y} = \{y_1, y_2, ..., y_J\}$, which are renewed at the beginning of every epoch. Given the obtained pseudo labels, we have N_{pos} positive and N_{neg} negative instances for each training instance. N_{pos} and N_{neg} vary for different instances. For simplicity in a mini-batch training, we fix common positive and negative numbers for every training instance. Given an image x, we randomly sample K instances that have different pseudo identities and one instance representation f_{pos} that has the same pseudo identity with x from the memory bank. Note that f_{pos} is from a random positive image that usually has a pose and camera style different from x and x'_{new}. x and x'_{new} are encoded by E_{id} into identity feature vectors f and f'_{new}. Next, f, f'_{new} and f_{pos} are used in turn to form three positive pairs. The f'_{new} and K different identity instances in the memory bank are used as K negative pairs. Towards learning robust view-invariant representations, we extend the InfoNCE loss [30] into a view-invariant loss between original and generated views. We use $sim(u, v) = \frac{u}{\|u\|_2} \cdot \frac{v}{\|v\|_2}$ to denote the cosine similarity. We define the view-invariant loss as a softmax log loss of $K + 1$ pairs as following.

$$\mathcal{L}_{vi} = \mathbb{E}[\log\left(1 + \frac{\sum_{i=1}^{K} \exp\left(sim(f'_{new}, k_i)/\tau\right)}{\exp\left(sim(f, f_{pos})/\tau\right)}\right)] \tag{6}$$

$$\mathcal{L}'_{vi} = \mathbb{E}[\log\left(1 + \frac{\sum_{i=1}^{K} \exp\left(sim(f'_{new}, k_i)/\tau\right)}{\exp\left(sim(f'_{new}, f)/\tau\right)}\right)] \tag{7}$$

$$\mathcal{L}''_{vi} = \mathbb{E}[\log(1 + \frac{\sum_{i=1}^{K} \exp(sim(f'_{new}, k_i)/\tau)}{\exp(sim(f'_{new}, f_{pos})/\tau)})], \quad (8)$$

where τ indicates a temperature coefficient that controls the scale of calculated similarities. $\mathcal{L}_{vi}$ maximizes the invariance between original and memory positive views. $\mathcal{L}'_{vi}$ maximizes the invariance between synthesized and original views. $\mathcal{L}''_{vi}$ maximizes the invariance between synthesized and memory positive views. Meanwhile, the synthesized view is pushed away from K negative views in the latent space. Replacing $sim(f'_{new}, k_i)$ in Eq. 6, Eq. 7 and Eq. 8 with $sim(f, k_i)$ is another possibility, which pushes away the original view from negative instances. After testing, $sim(f'_{new}, k_i)$ works better, because pushing away the synthesized view from negative instances aid the generation of more accurate synthesized views that look different from the K negative instances.

3.3. Joint Training

Our proposed GCL framework is trained in a joint training way. Both GAN and contrastive instance discrimination can be trained in a self-supervised manner. While the GAN learns a data distribution via adversarial learning on each instance, contrastive instance discrimination learns representations by retrieving each instance from candidates. In our designed joint training, the two modules work as two collaborators with the same objective: enhancing the quality of representations built by the shared identity encoder E_{id}. We formulate our GCL as an approach to augment contrast for unsupervised ReID. Firstly, the generative module generates online data augmentation, which enhances the positive view diversity for contrastive module. Secondly, the contrastive module, in turn, learns view-invariant representations by matching original and generated views, which refine the generation quality. The joint training boosts both modules simultaneously. Our joint training conducts forward propagation initially on the generative module and subsequently on the contrastive module. Back-propagation is then conducted with an overall loss that combines Eq. 4, Eq. 6, Eq. 7 and Eq. 8.

$$\mathcal{L}_{all} = \mathcal{L}_{gan} + \mathcal{L}_{vi} + \mathcal{L}'_{vi} + \mathcal{L}''_{vi} \quad (9)$$

To accelerate the training process and alleviate the noise from imperfect generation quality at beginning epochs, we need to warm up the four modules used in the View Generator E_{id}, E_{str}, G and D. We firstly use a state-of-the-art unsupervised ReID method to warm up E_{id}, which is then considered as a baseline in our ablation studies. Generally speaking, any unsupervised ReID method can be used to warm up E_{id}. Before conducting the View Contrast, we freeze E_{id} and warm up E_{str}, G, and D only with GAN loss in Eq. 4 for 40 epochs. In the following, we bring in the memory bank and the pseudo labels to jointly train the whole framework with $\mathcal{L}_{all}$ for another 20 epochs. During the joint training, pseudo labels are updated at the beginning of every epoch.

4. Experiments

4.1. Datasets and Evaluation Protocols

Three mainstream person ReID datasets are considered in our experiments, including Market-1501 [46], DukeMTMC-reID [31] and MSMT17 [41]. Market-1501 is composed of 12,936 images of 751 identities for training and 19,732 images of 750 identities for test captured from 6 cameras. DukeMTMC-reID contains 16,522 images of 702 persons for training, 2,228 query images and 17,661 gallery images of 702 persons for test from 8 cameras. MSMT17 is a larger dataset, which contains 32,621 training images of 1,041 identities and 93,820 testing images of 3,060 identities collected from 15 cameras.

Following state-of-the-art unsupervised ReID methods [35, 24], we evaluate our proposed method GCL under fully unsupervised setting on the three datasets and under four UDA benchmark protocols, including Market→Duke, Duke→Market, Market→MSMT and Duke→MSMT. We report both quantitative and qualitative results for unsupervised person ReID and view generation.

4.2. Implementation Details

We firstly present network design details of E_{id}, E_{str}, G and D. In the following descriptions, we write the size of feature maps in channel×height×width. Our model design is mainly inspired by [47, 54]. (1) E_{id} is a ImageNet [32] pre-trained ResNet50 [17] with slight modifications. The original fully connected layer is replaced by a fully connected embedding layer, which outputs identity representations f in 512×1×1 for the View Contrast. In parallel, we add a part average pooling that outputs identity features f_{id} in 2048×4×1 for the View Generator. (2) E_{str} is composed of four convolutional and four residual layers, which output structure features f_{str} in 128×64×32. (3) G contains four residual and four convolutional layers. Every residual layer contains two adaptive instance normalization layers [19] that transform f_{id} into scale and bias parameters. (4) D is a multi-scale PatchGAN [20] discriminator at 64×32, 128×64 and 256×128.

Then, we present the training and testing configuration details. Our framework is implemented in Pytorch and trained with one Nvidia Titan RTX GPU. (1) For the E_{id} warm-up, we consider JVTC [24], because it is a state-of-the-art ReID method that is compatible with both fully unsupervised and UDA settings. We also test other baselines, *e.g.*, MMCL [35] and ACT [45] to demonstrate the generalizability of our method. (2) For training, inputs are resized to 256×128. We empirically set a large weight $\lambda_{img} = \lambda_{feat} = 5$ for reconstruction in Eq. 4. With a batch size of 16, we use SGD to train E_{id} and Adam optimizer to

Method	Reference	Market1501					DukeMTMC-reID				
		Source	mAP	Rank1	Rank5	Rank10	Source	mAP	Rank1	Rank5	Rank10
BUC [27]	AAAI'19	None	29.6	61.9	73.5	78.2	None	22.1	40.4	52.5	58.2
SoftSim [28]	CVPR'20	None	37.8	71.7	83.8	87.4	None	28.6	52.5	63.5	68.9
TSSL [42]	AAAI'20	None	43.3	71.2	-	-	None	38.5	62.2	-	-
MMCL [35]	CVPR'20	None	45.5	80.3	89.4	92.3	None	40.2	65.2	75.9	80.0
JVTC [24]	ECCV'20	None	41.8	72.9	84.2	88.7	None	42.2	67.6	78.0	81.6
JVTC+ [24]	ECCV'20	None	47.5	79.5	89.2	91.9	None	50.7	74.6	82.9	85.3
MMCL*	This paper	None	45.1	79.5	89.0	91.9	None	40.9	64.8	75.2	79.8
JVTC*	This paper	None	47.2	75.4	86.7	90.5	None	43.9	66.8	77.6	81.0
JVTC+*	This paper	None	50.9	79.1	89.8	92.9	None	52.8	74.9	83.3	85.8
ours(MMCL*)	This paper	None	54.9	83.7	91.6	94.0	None	49.3	69.7	79.7	82.8
ours(JVTC*)	This paper	None	63.4	83.7	91.6	94.3	None	53.3	72.4	82.0	84.9
ours(JVTC+*)	This paper	None	**66.8**	**87.3**	**93.5**	**95.5**	None	**62.8**	**82.9**	**87.1**	**88.5**
ECN [50]	CVPR'19	Duke	43.0	75.1	87.6	91.6	Market	40.4	63.3	75.8	80.4
PDA [25]	ICCV'19	Duke	47.6	75.2	86.3	90.2	Market	45.1	63.2	77.0	82.5
CR-GAN [8]	ICCV'19	Duke	54.0	77.7	89.7	92.7	Market	48.6	68.9	80.2	84.7
SSG [12]	ICCV'19	Duke	58.3	80.0	90.0	92.4	Market	53.4	73.0	80.6	83.2
MMCL [35]	CVPR'20	Duke	60.4	84.4	92.8	95.0	Market	51.4	72.4	82.9	85.0
ACT [45]	AAAI'20	Duke	60.6	80.5	-	-	Market	54.5	72.4	-	-
DG-Net++ [54]	ECCV'20	Duke	61.7	82.1	90.2	92.7	Market	63.8	78.9	87.8	90.4
JVTC [24]	ECCV'20	Duke	61.1	83.8	93.0	95.2	Market	56.2	75.0	85.1	88.2
ECN+ [51]	PAMI'20	Duke	63.8	84.1	92.8	95.4	Market	54.4	74.0	83.7	87.4
JVTC+ [24]	ECCV'20	Duke	67.2	86.8	95.2	97.1	Market	66.5	80.4	**89.9**	92.2
MMT [13]	ICLR'20	Duke	71.2	87.7	94.9	96.9	Market	65.1	78.0	88.8	**92.5**
CAIL [29]	ECCV'20	Duke	71.5	88.1	94.4	96.2	Market	65.2	79.5	88.3	91.4
ACT*	This paper	Duke	59.1	78.8	88.9	91.7	Market	51.5	70.9	80.0	83.4
JVTC*	This paper	Duke	65.0	85.7	93.6	95.6	Market	56.5	73.9	84.5	87.7
JVTC+*	This paper	Duke	67.6	87.0	95.2	97.0	Market	66.7	81.0	**89.9**	91.5
ours(ACT*)	This paper	Duke	66.7	83.9	91.4	93.4	Market	55.4	71.9	81.6	84.6
ours(JVTC*)	This paper	Duke	73.4	89.1	95.0	96.6	Market	60.4	77.2	86.2	88.4
ours(JVTC+*)	This paper	Duke	**75.4**	**90.5**	**96.2**	**97.1**	Market	**67.6**	**81.9**	88.9	90.6

Table 1: Comparison of unsupervised ReID methods (%) with a ResNet50 backbone on Market and Duke datasets. We test our proposed method on several baselines, whose names are in brackets. * refers to our implementation based on authors' code.

train E_{str}, G and D. Learning rate is set to 1×10^{-4} during the warm-up. In the joint-training, learning rate in Adam is set to 1×10^{-4} and 3.5×10^{-4} in SGD and are multiplied by 0.1 after 10 epochs. (3) In the View Contrast module, we set the momentum coefficient $\alpha = 0.2$ in Eq. 5 and the temperature $\tau = 0.04$ in Eq. 6. The number of negatives K is 8192. DBSCAN density radius is set to 2×10^{-3}. (4) For testing, only E_{id} is conserved and outputs representations f of dimension 512.

Important parameters are set by a grid search on the fully unsupervised Market-1501 benchmark. The temperature τ is searched from $\{0.03, 0.04, 0.05, 0.06, 0.07\}$ and finally is set to 0.04. A smaller τ increases the scale of similarity scores in the Eq. 6, Eq. 7 and Eq. 8, which makes view-invariant losses more sensitive to inter-instance difference. However, when τ is set to 0.03, these losses become too sensitive and make the training unstable. The number of negatives K is searched from $\{2048, 4096, 8192\}$. A larger K pushes away more negatives in the view-invariant losses. Since the Market-1501 dataset has only 12936 training images, we set $K = 8192$.

4.3. Unsupervised ReID Evaluation

Comparison with state-of-the-art methods. Tab. 1 shows the quantitative results on the Market-1501 and DukeMTMC-reID datasets. Tab. 2 shows the quantitative results on the MSMT17 dataset. Our method is mainly designed for fully unsupervised ReID. Under this setting, we test the performance of GCL with three different baselines, including MMCL, JVTC and JVTC+. Our implementation of the three baselines provides results that are slightly different from those mentioned in the corresponding papers. Thus, we firstly report results of our implementations and then add our GCL on these baselines. Our method improves the performance of the baselines by large margins. These improvements show that GANs are not limited to cross-domain style transfer for unsupervised ReID.

Under the UDA setting, we also evaluate the performance of GCL with three different baselines, including ACT, JVTC and JVTC+. The labeled source dataset is only used to warm up our identity encoder E_{id}, but not used in our joint generative and contrastive training. Compared to fully unsupervised methods, the UDA warmed E_{id} is stronger and extracts improved identity features. Thus, the performance of UDA methods is generally higher than fully unsupervised methods. With a strong baseline JVTC+, our GCL achieves state-of-the-art performance.

Method	Reference	MSMT17				
		Source	mAP	R1	R5	R10
MMCL [35]	CVPR'20	None	11.2	35.4	44.8	49.8
JVTC [24]	ECCV'20	None	15.1	39.0	50.9	56.8
JVTC+ [24]	ECCV'20	None	17.3	43.1	53.8	59.4
JVTC*	This paper	None	13.4	36.0	48.8	54.9
JVTC+*	This paper	None	16.3	40.4	55.6	61.8
ours(JVTC*)	This paper	None	18.0	41.6	53.2	58.4
ours(JVTC+*)	This paper	None	**21.3**	**45.7**	**58.6**	**64.5**
ECN [50]	CVPR'19	Market	8.5	25.3	36.3	42.1
SSG [12]	ICCV'19	Market	13.2	31.6	49.6	-
MMCL [35]	CVPR'20	Market	15.1	40.8	51.8	56.7
ECN+ [51]	PAMI'20	Market	15.2	40.4	53.1	58.7
JVTC [24]	ECCV'20	Market	19.0	42.1	53.4	58.9
DG-Net++ [54]	ECCV'20	Market	22.1	48.4	60.9	66.1
CAIL [29]	ECCV'20	Market	20.4	43.7	56.1	61.9
MMT [13]	ICLR'20	Market	22.9	49.2	63.1	68.8
JVTC+ [24]	ECCV'20	Market	25.1	48.6	65.3	68.2
JVTC*	This paper	Market	17.1	39.6	53.3	59.3
JVTC+*	This paper	Market	20.5	44.0	59.5	71.1
ours(JVTC*)	This paper	Market	21.5	45.0	57.1	66.5
ours(JVTC+*)	This paper	Market	**27.0**	**51.1**	**63.9**	**69.9**
ECN [50]	CVPR'19	Duke	10.2	30.2	41.5	46.8
SSG [12]	ICCV'19	Duke	13.3	32.2	51.2	-
MMCL [35]	CVPR'20	Duke	16.2	43.6	54.3	58.9
ECN+ [51]	PAMI'20	Duke	16.0	42.5	55.9	61.5
JVTC [24]	ECCV'20	Duke	20.3	45.4	58.4	64.3
DG-Net++ [54]	ECCV'20	Duke	22.1	48.8	60.9	65.9
MMT [13]	ICLR'20	Duke	23.3	50.1	63.9	69.8
CAIL [29]	ECCV'20	Duke	24.3	51.7	64.0	68.9
JVTC+ [24]	ECCV'20	Duke	27.5	52.9	70.5	75.9
JVTC*	This paper	Duke	19.9	45.4	59.1	64.9
JVTC+*	This paper	Duke	23.6	49.4	65.2	71.1
ours(JVTC*)	This paper	Duke	24.9	50.8	63.4	68.9
ours(JVTC+*)	This paper	Duke	**29.7**	**54.4**	**68.2**	**74.2**

Table 2: Comparison of unsupervised Re-ID methods (%) with a ResNet50 backbone on MSMT17. * refers to our implementation based on authors' code.

Ablation Study. To better understand the contribution of generative and contrastive modules, we conduct ablation experiments on the two fully unsupervised benchmarks: Market-1501 and DukeMTMC-reID. Quantitative results with a JVTC baseline are reported in Tab. 3. By gradually adding loss functions on the baseline, our ablation experiments correspond to three scenarios. (1) Only Generation: with only $\mathcal{L}_{gan}$, our generation module disentangles identity and structure features. Since there is no inter-view constraint, E_{id} tends to extract view-specific identity features, which decreases the ReID performance. (2) Only Contrast: we use $\mathcal{L}_{vi}^{woGAN} = \mathbb{E}[\log(1 + \frac{\sum_{i=1}^{K} \exp(sim(f,k_i)/\tau)}{\exp(sim(f,f_{pos})/\tau)})]$ to train our contrastive module without generation. We also add a set of traditional data augmentation, including random flipping, cropping, jittering, erasing, to train our contrastive module like a traditional memory bank based contrastive method. (3) Joint Generation and Contrast: $\mathcal{L}_{vi}$, $\mathcal{L}'_{vi}$ and $\mathcal{L}''_{vi}$ enhance the view-invariance of identity representations between original, synthesized and memory-stored positive views, while negative views are pushed away. We provide how view-invariant representations learned from generated views affect pseudo labels in Appendix B.

Loss	Market-1501		DukeMTMC-reID	
	mAP	Rank1	mAP	Rank1
Baseline	47.2	75.4	43.9	66.8
$+\mathcal{L}_{gan}$	41.6	69.0	25.8	45.9
$+\mathcal{L}_{vi}^{woGAN}$	47.8	75.2	44.1	67.8
$+\mathcal{L}_{vi}^{woGAN} + TDA$	53.7	78.7	48.5	70.0
$+\mathcal{L}_{gan} + \mathcal{L}_{vi}$	54.1	79.4	47.4	68.4
$+\mathcal{L}_{gan} + \mathcal{L}_{vi} + \mathcal{L}'_{vi}$	59.2	82.2	50.5	71.0
$+\mathcal{L}_{gan} + \mathcal{L}_{vi} + \mathcal{L}'_{vi} + \mathcal{L}''_{vi}$	**63.4**	**83.7**	**53.3**	**72.4**

Table 3: Ablation study on loss functions used in two modules. (1). $\mathcal{L}_{gan}$ corresponds to generation w/o contrast. (2). $\mathcal{L}_{vi}^{woGAN}$ corresponds to contrast w/o generation. TDA denotes traditional data augmentation. (3). $\mathcal{L}_{gan} + \mathcal{L}_{vi}$ ($\mathcal{L}'_{vi}$ and $\mathcal{L}''_{vi}$) correspond to joint generative and contrastive learning.

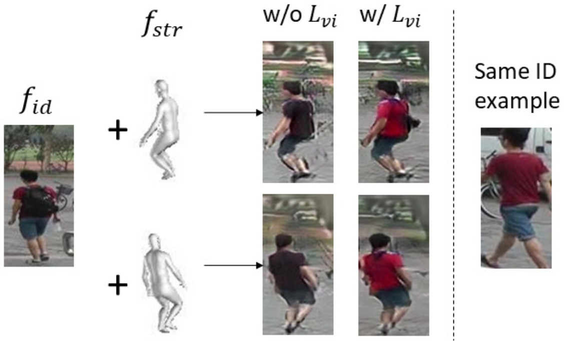

Figure 4: Qualitative ablation study on the view-invariant losses. For simplicity, $\mathcal{L}_{vi}$ denotes three view-invariant losses $\mathcal{L}_{vi} + \mathcal{L}'_{vi} + \mathcal{L}''_{vi}$, which helps E_{id} to extract view-invariant features (red shirt).

Method	FID(realism)	SSIM(diversity)
Real	**7.22**	0.350
FD-GAN [14]	216.88	0.271
IS-GAN [10]	281.63	0.165
DG-Net [47]	18.24	0.360
Ours(U)	59.86	0.367
Ours(UDA)	53.07	**0.369**

Table 4: Comparison of FID (lower is better) and SSIM (higher is better) on Market-1501 dataset. U denotes the fully unsupervised setting. UDA denotes Duke→Market setting.

We also conduct a qualitative ablation study, where synthesized novel views without and with view-invariant losses are illustrated in Fig. 4. Results confirm that E_{id} extracts view-specific identity features (black bag), in the case that view-invariant losses are not used. Given view-invariant losses, E_{id} is able to extract view-invariant identity features (red shirt). Another example is provided in Appendix C.

4.4. Generation Quality Evaluation

Comparison with state-of-the-art methods. We compare generated images between our proposed GCL under the JVTC [24] warmed fully unsupervised setting and state-of-the-art GAN-based ReID methods in Fig. 5. FD-GAN [14], IS-GAN [10] and DG-Net [47] are supervised Re-ID methods. Since the source code of these three methods is available, we compare generated images of same identities. We observe that there exists blur in images generated by

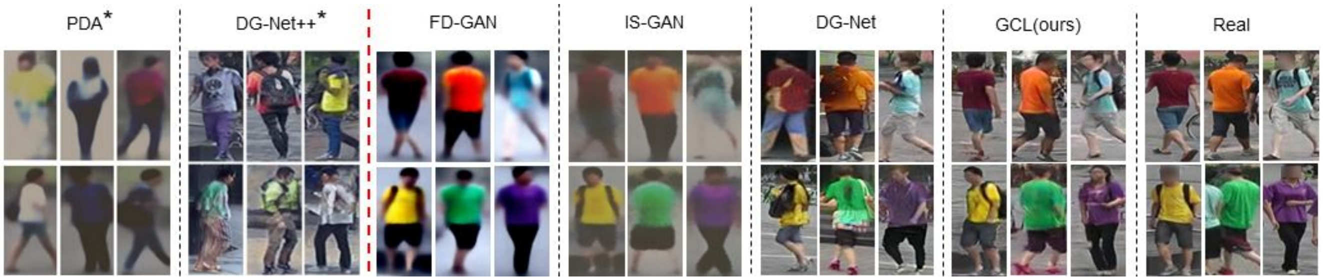

Figure 5: Comparison of the generated images on Market-1501 dataset. ⋆ refers to methods without sharing source code, whose examples are cropped from their papers. Examples of FD-GAN, IS-GAN, DG-Net and GCL are generated from six real images shown in the figure.

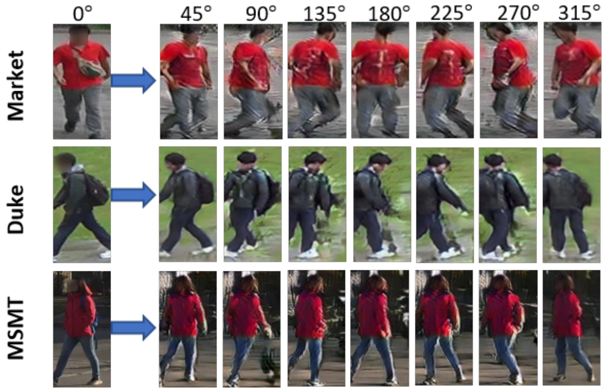

Figure 6: Generated novel views on the three datasets.

FD-GAN and IS-GAN. DG-Net generates sharper images, but different body shapes and some incoherent objects (bags and clothes) are observed. PDA [25] and DG-Net++ [54] are UDA methods, whose source code is not yet released. We can only compare several generated images with unknown identities as illustrated in their papers. PDA generates blurred cross-domain images, whose quality is similar to FD-GAN and IS-GAN. DG-Net++ extends DG-Net into cross-domain generation, which has same problems of body shape and incoherent objects. Our GCL preserves better body shape information and does not generate incoherent objects. Moreover, our GCL is a fully unsupervised method.

We use Fréchet Inception Distance (FID) [18] to measure visual quality, as well as Structural SIMilarity (SSIM) [40] to capture structure diversity of generated images. In Tab. 4, we compare our method with FD-GAN [14], IS-GAN [10] and DG-Net [47], whose source code is available. FID measures the distribution distance between generated and real images, where a lower FID represents the case, where generated images are similar to real ones. SSIM measures the intra-class structural similarity, where a larger SSIM represents a larger diversity. We note that DG-Net is outperforms our method w.r.t. FID, because the distribution is better maintained with ground truth identities in the supervised method DG-Net. However, our method is superior to DG-Net w.r.t. SSIM, as DG-Net swaps intra-dataset structures, whereas our rotated meshes build structures that do not exist in the original dataset.

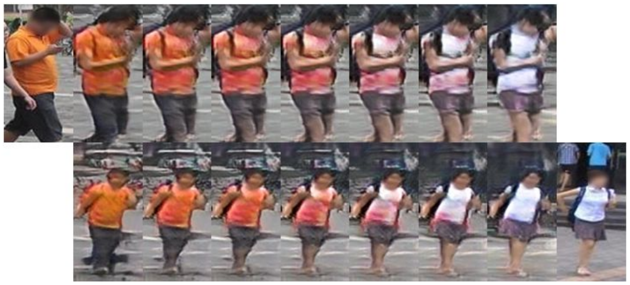

Figure 7: Linear interpolation on identity features. Identity features are swapped between left and right persons.

More discussion. To validate, whether identity and structure features can be really disentangled under a fully unsupervised ReID setting, two experiments are conducted by changing firstly only structure features and then only identity features. Results in Fig. 6 show that changing structure features only change structures and do not affect appearances. We also fix structure features and linearly interpolate two random identity feature vectors. Results in Fig. 7 show that identity features only change appearances and do not affect structures in generated images. More examples are provided in Appendix D.

5. Conclusions

In this paper, we propose a joint generative and contrastive learning framework for unsupervised person ReID. Deviating from previous contrastive methods with traditional data augmentation techniques, we generate diversified views with a 3D mesh guided GAN. These generated novel views are then combined with original images in memory based contrastive learning, in order to learn view-invariant representations, which in turn improve generation quality. Our generative and contrastive modules mutually promote each other's performance in unsupervised ReID. Moreover, our framework does not rely on a source dataset, which is mandatory in style transfer based methods. Extensive experiments on three datasets validate the effectiveness of our framework in both unsupervised person ReID and multi-view person image generation.

References

[1] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *ECCV*, 2018. 1

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 2

[3] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, 2019. 2

[4] Hao Chen, Benoit Lagadec, and Francois Bremond. Learning discriminative and generalizable representations by spatial-channel partition for person re-identification. In *WACV*, 2020. 1

[5] Hao Chen, Benoit Lagadec, and Francois Bremond. Enhancing diversity in teacher-student networks via asymmetric branches for unsupervised person re-identification. In *WACV*, 2021. 3

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 2

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2

[8] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *ICCV*, 2019. 3, 6

[9] 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. 1, 3

[10] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. In *NeurIPS*, 2019. 2, 7, 8

[11] 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. 4

[12] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV*, 2019. 3, 6, 7

[13] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 3, 6, 7

[14] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NeurIPS*, 2018. 2, 7, 8

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 2

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 8

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 5

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 5

[21] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 3

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 2

[24] Jianing Li and Shiliang Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7

[25] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *ICCV*, 2019. 1, 2, 3, 6, 8

[26] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *BMVC*, 2018. 3

[27] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019. 1, 3, 6

[28] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *CVPR*, 2020. 3, 6

[29] Chuanchen Luo, Chunfeng Song, and Zhaoxiang Zhang. Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup. In *ECCV*, 2020. 2, 6, 7

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 4

[31] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCVW*, 2016. 5

[32] Olga Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5

[33] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 1

[34] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *CVPR*, 2018. 2

[35] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *CVPR*, 2020. 1, 2, 3, 5, 6, 7

[36] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *CVPR*, 2018. 3

[37] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3AN: Disentangling appearance and motion for video generation. In *CVPR*, 2020. 2

[38] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *WACV*, 2020. 2

[39] Yaohui Wang, Francois Bremond, and Antitza Dantcheva. Inmodegan: Interpretable motion decomposition generative adversarial network for video generation. *arXiv preprint arXiv:2101.03049*, 2021. 2

[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 8

[41] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 1, 3, 5

[42] Guile Wu, Xiatian Zhu, and Shaogang Gong. Tracklet self-supervised learning for unsupervised person re-identification. In *AAAI*, 2020. 3, 6

[43] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wei Bian, and Yi Yang. Progressive learning for person re-identification with one example. *TIP*, 2019. 3

[44] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 2, 4

[45] Fengxiang Yang, Ke Li, Zhun Zhong, Zhiming Luo, Xing Sun, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Shaozi Li. Asymmetric co-teaching for unsupervised cross-domain person re-identification. In *AAAI*, 2020. 3, 5, 6

[46] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. *ICCV*, 2015. 5

[47] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019. 3, 5, 7, 8

[48] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 2

[49] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *ECCV*, 2018. 1, 3

[50] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019. 3, 6, 7

[51] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *PAMI*, 2020. 3, 6, 7

[52] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018. 2

[53] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 3

[54] Yang Zou, Xiaodong Yang, Zhiding Yu, B. V. K. Vijaya Kumar, and Jan Kautz. Joint disentangling and adaptation for cross-domain person re-identification. In *ECCV*, 2020. 1, 3, 5, 6, 7, 8