



**HAL**  
open science

# Learning Invariance from Generated Variance for Unsupervised Person Re-identification

Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, Francois  
Bremond

► **To cite this version:**

Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, Francois Bremond. Learning Invariance from Generated Variance for Unsupervised Person Re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, inPress, pp.1-15. 10.1109/TPAMI.2022.3226866 . hal-03931340

**HAL Id: hal-03931340**

**<https://hal.science/hal-03931340v1>**

Submitted on 9 Jan 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Learning Invariance from Generated Variance for Unsupervised Person Re-identification

Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, Francois Bremond

**Abstract**—This work focuses on unsupervised representation learning in person re-identification (ReID). Recent self-supervised contrastive learning methods learn invariance by maximizing the representation similarity between two augmented views of a same image. However, traditional data augmentation may bring to the fore undesirable distortions on identity features, which is not always favorable in id-sensitive ReID tasks. In this paper, we propose to replace traditional data augmentation with a generative adversarial network (GAN) that is targeted to generate augmented views for contrastive learning. A 3D mesh guided person image generator is proposed to disentangle a person image into id-related and id-unrelated features. Deviating from previous GAN-based ReID methods that only work in id-unrelated space (pose and camera style), we conduct GAN-based augmentation on both id-unrelated and id-related features. We further propose specific contrastive losses to help our network learn invariance from id-unrelated and id-related augmentations. By jointly training the generative and the contrastive modules, our method achieves new state-of-the-art unsupervised person ReID performance on mainstream large-scale benchmarks.

**Index Terms**—Person re-identification, image synthesis, representation disentanglement, data augmentation, contrastive learning

## 1 INTRODUCTION

GIVEN an image of a target person, a person re-identification (ReID) system [1], [2] aims at matching images of the same person across non-overlapping cameras. With the help of human-annotated labels, *supervised person ReID* methods [3], [4] have yielded impressive results. However, there usually exist strong domain gaps between different domains, such as illumination condition, camera property and scenario variation. As shown in previous methods [5], [6], a ReID model trained on a specific domain is hard to generalize to other domains. One straightforward solution is to annotate and re-train the ReID model in a new domain, which is cumbersome and time-consuming for real-world deployments. Towards an automatic adaptive system, *unsupervised person ReID* [7], [8], [9] has attracted increasing attention in the research community. Compared with supervised counterparts, unsupervised methods directly learn from unlabeled images and therefore entail better scalability in real-world deployments.

Recent *self-supervised contrastive learning* studies [10], [11] have shown promising performance in unsupervised representation learning. By maximizing the representation similarity between two different views (augmented versions) of a same image, contrastive methods learn representations that are invariant to different conditions. In this context, data augmentation plays a crucial role in mimicking real-world condition variance. Contrastive learning methods are able to build more robust representations, given they were provided with better augmented views. Previous methods generally consider traditional data augmentation techniques,

e.g., random flipping, cropping, color jittering, blurring and erasing [12]. However, these random augmentation techniques may cause undesirable distortion to crucial identity information. To overcome this issue, we propose to use a Generative Adversarial Network (GAN) [13] as an augmentation substitute, as it is able to disentangle a representation into id-related and id-unrelated features (see Table 1). More accurate augmented views can be obtained by modifying a certain factor while preserving other factors.

Previous GAN-based unsupervised ReID methods [14], [15], [16], [17] often treat unsupervised ReID as an unsupervised domain adaptation task, which attempts to adapt a model trained on a labeled source domain to an unlabeled target domain. Under this setting, it is intuitive to use GAN-based style transfer [18], [19] to generate source domain images in the style of a target domain. A model can be re-trained on the generated images in target domain style with source domain labels. However, unsupervised domain adaptation performance often strongly relies on quality and scale of the source domain. Differently, we treat unsupervised ReID as a *contrastive representation learning* task, where the source domain is not mandatory. To this end, we integrate a generative module and a contrastive module into a joint learning framework.

For the generative module, we propose a 3D mesh based generator. Conventional pose transfer methods [20], [21] use 2D pose [22] to guide generation, not preserving body shape information. 3D mesh recovery [23] jointly estimates body shape, as well as 3D pose, which conserves more identity information for unsupervised ReID. We use 3D meshes to guide the generation, where generated images in new poses are then used as augmented views in the contrastive module.

For the contrastive module, we use a clustering algorithm to generate pseudo labels, aimed at maximizing representation similarity between different views of a same

- H. Chen, Y. Wang, A. Dantcheva and F. Bremond are with Inria and Université Côte d’Azur, 2004 Route des Lucioles, 06902 Valbonne, France. E-mail: {hao.chen, yaohui.wang, antitza.dantcheva, francois.bremond}@inria.fr
- B. Lagadec is with European Systems Integration, 362 Avenue du Campan, 06110 Le Cannet, France. E-mail: benoit.lagadec@esifrance.net

TABLE 1  
Id-related and Id-unrelated factors in a person image.

Id-related	Id-unrelated
cloth color, hair color, texture, body shape	pose, view-point, illumination, camera style background













pseudo identity. Our model attracts a generated view to its original view, while repulsing the generated view from images of different identities. The contrastive module permits an identity encoder to extract view-invariant identity features, which, in turn, improves the generation quality.

In our previous work [9], GAN-based augmentation was only conducted on id-unrelated features, which has been common practice in previous GAN-based ReID methods [20], [24], [25]. Modifying id-unrelated features allows for learning identity features that are more invariant to id-unrelated variations. In this paper, we explore the possibility of conducting GAN-based augmentation on the id-related features to further improve the ReID performance. Inspired by Mixup [26] that interpolates two images to learn a smoother decision boundary between two classes, we propose to interpolate disentangled id-related features inside the generative module, namely **Disentangled Mixup (D-Mixup)**. As shown in Table 2, if two persons  $P_1$  and  $P_2$  respectively wear red and yellow clothes, an in-between identity in orange clothes should be marked as  $0.5P_1 + 0.5P_2$ . However, in a dataset, such a person in orange clothes is normally labeled as a totally different identity  $P_3$ , which hinders a network from learning the accurate relationship between different identities. Compared to traditional image-level Mixup [26] and feature-level Mixup [27], our proposed D-Mixup generates more accurate in-between identity images, which are more suitable for fine-grained person ReID. In our D-Mixup, we try to make our network understand the mixed identity  $0.5P_1 + 0.5P_2$  is not related to id-unrelated features (pose and view-point), but only related to id-related features (cloth color).

To summarize, our contributions include the following:

- We propose a 3D mesh guided generator to disentangle representations into id-related and id-unrelated features. Two novel data augmentation techniques are proposed respectively on id-unrelated and id-related features.
- We propose Rotation Contrast and Mixup Contrast modules to respectively learn invariance from id-unrelated and id-related augmented views.
- We propose an enhanced joint generative and contrastive learning framework. We comprehensively investigate how the generative and contrastive modules mutually promote each other and contribute to unsupervised ReID performance.
- Extensive experiments validate the superiority of proposed GAN-based augmentation over traditional augmentation for unsupervised person ReID. Our method achieves new state-of-the-art unsupervised person ReID performance on mainstream image-based datasets, including Market-1501, DukeMTMC-ReID and MSMT17.

TABLE 2  
Interpolation results between two random persons  $P_1$  and  $P_2$  with image-level Mixup [26], feature-level Mixup (F-Mixup) [27] and our proposed disentangled Mixup (D-Mixup). To visualize results from F-Mixup, we follow AMR [28] to train a VAE-GAN for mixed image reconstruction. Our D-Mixup only interpolates disentangled identity features in the generation, which alleviates noise from mixed structural features.

	Inputs		Mixup	F-Mixup	D-Mixup	
Image						
Image						
Label	$1.0P_1$ $0.0P_2$	$0.0P_1$ $1.0P_2$	$0.5P_1$ $0.5P_2$	$0.5P_1$ $0.5P_2$	$0.5P_1$ $0.5P_2$	$0.5P_1$ $0.5P_2$

- Our method can be also applied to video-based person ReID. Our method significantly outperforms previous unsupervised video person ReID methods on MARS and DukeMTMC-VideoReID datasets.

## 2 RELATED WORK

### 2.1 Contrastive learning

Contrastive learning [29] has shown impressive performance for un-/self-supervised representation learning [10], [11], [30], [31], [32], [33]. Such contrastive methods target at learning representations that are invariant to different distortions by attracting positive pairs, while repulsing negative pairs. For each image, a positive pair can be constituted by two augmented views, whereas all other images in a dataset are regarded as negative samples. Contrastive learning methods benefit from a set of well defined data augmentation techniques, which can mimic real-world image distortions. For example, MoCo [11] used random cropping, color jittering, horizontal flipping and grayscale conversion to obtain positive view pairs. As an extension, MoCo-v2 [34] included blurring and stronger color distortion, which enhanced the original method. However, most of data augmentation settings in contrastive learning methods were designed for general image classification datasets, e.g., ImageNet [35]. These traditional augmentation techniques are not always suitable for color-sensitive person ReID, especially those that introduce strong color distortion.

### 2.2 Data augmentation

As a technique to constitute positive pairs, data augmentation plays an important role in contrastive learning. Recently, GAN and Mixup have provided new approaches for data augmentation in person ReID.

#### 2.2.1 GAN-based augmentation

Zheng et al. [36] unconditionally generated a lot of unlabeled person images with DCGAN [37] to enlarge data

volume for supervised ReID. Following GAN-based methods were usually conditionally conducted on some factors from Table 1. **1) Pose:** With the guidance of 2D poses, FD-GAN [20] and PN-GAN [38] generated a target person in new poses to learn pose-irrelevant representations for single-domain supervised ReID. Similar pose transfer [21] was then proposed to address unsupervised domain adaptive (UDA) ReID. **2) Dataset style (illumination):** As a dataset is usually recorded in a uniform illumination condition, PTGAN [14] and SyRI [15] used CycleGAN [39] to minimize the domain gap between different datasets by generating person images in the style of a target domain. **3) Camera style:** Instead of the general dataset style, CamStyle [24] transferred images captured from one camera into the style of another camera, in order to reduce inter-camera style gaps. Similar method [16] was then applied to UDA ReID. **4) Background:** SBSGAN [40] and CR-GAN [41] respectively were targeted at removing and switching the background of a person image to mitigate background influence for UDA ReID. **5) General structure:** By switching global and local level identity-unrelated features, IS-GAN [42] disentangled a representation into identity-related and identity-unrelated features without any concrete guidance. As a concrete guidance, a gray-scaled image contains multiple id-unrelated factors of a person image, including pose, background and carrying structures. By recoloring gray-scaled person images with the color distribution of other images, DG-Net [25] and DG-Net++ [17] learned disentangled identity representations invariant to structure factors. Our proposed 3D mesh guided generator shares certain similarity with pose transfers and DG-Net++. However, both pose transfers and DG-Net++ lose body shape information, which can be conserved by 3D meshes. Moreover, as opposed to DG-Net++, we do not transfer style in a cross-domain manner, which allows our method to operate without a source domain.

### 2.2.2 Mixup

Mixup [26] is a simple yet effective data augmentation technique that interpolates two samples and labels into one new in-between sample, which encourages a smoother decision boundary between two classes. The interpolation can be conducted between two images [26], [43], two feature representations [27] and two portions of different images [44]. Initially proposed for supervised image classification [26], [43], Mixup has been successfully extended to semi-supervised learning [45], [46], unsupervised domain adaptation [47], as well as novel class discovery [48]. Aug-Mix [49] combines multiple augmented versions of an image into a mixed image and proves that such technique can enhance robustness on corrupted data. CAIL [50] applies image-level Mixup between a source domain image and a target domain image to create a between-domain person image, which facilitates cross-domain knowledge transfer in unsupervised domain adaptive ReID. The above methods usually interpolate whole images or whole representations, resulting in noise from overlapping person structures. To reduce noise from mixed person structures, we propose to interpolate only disentangled identity features, which is compatible with our proposed 3D mesh guided GAN.

## 2.3 Unsupervised person ReID

Depending on the necessity of a large-scale labeled source dataset, unsupervised person ReID methods can be roughly categorized into unsupervised domain adaptive (UDA) and fully unsupervised ReID. We note that above mentioned GAN-based unsupervised ReID methods [14], [15], [16], [17], [21], [41] fall into the setting of UDA ReID. Several works [51], [52] leveraged semantic attributes to facilitate the domain adaptation. Another prominent approach has to do with assigning pseudo labels to unlabeled images and conducting pseudo label learning [7], [8], [50], [53], [54], [55], [56]. Pseudo labels can be obtained by existing clustering algorithms, e.g., K-means [8] and DBSCAN [17], [55], or newly designed pseudo labelling algorithms [53], [56]. Since the performance of UDA ReID is highly correlated to the scale and quality of a source domain, recent fully unsupervised ReID methods have attracted more attention. Most of previous fully unsupervised methods [57], [58], [59], [60], [61] were based on pure pseudo label learning. Our previous method GCL [9] has entailed a hybrid GAN and pseudo label learning method, which is compatible with both UDA and fully unsupervised settings. We here propose a new id-related augmentation D-Mixup, which enhances our framework to achieve new state-of-the-art performance under both UDA and fully unsupervised settings.

## 3 METHOD

In this paper, we propose an enhanced joint Generative and Contrastive Learning (GCL+) for unsupervised person ReID. We define unsupervised ReID as a problem of learning invariance from self-augmented variance. As illustrated in Fig. 1. (a), the proposed GCL+ constitutes of two modules: a generative module that provides GAN-based augmented views, as well as a contrastive module that learns invariance from augmented views. These two modules are coupled by a shared identity encoder. After the joint training, only the shared identity encoder is conserved for inference. In the following sections, we proceed to provide details related to both modules. To facilitate the reading, we include a list of abbreviations in Supplementary Materials Section C.

### 3.1 Generative Module

Our generative module is composed of 4 networks, including an identity encoder  $E_{id}$ , a structure encoder  $E_{str}$ , a decoder  $G$  and a discriminator  $D$ . Given an unlabeled person ReID dataset  $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$ , we use the prominent algorithm HMR [23] to generate corresponding 3D meshes, which are then used as structure guidance in the generative module. By recoloring a specific 3D mesh to reconstruct a real image, a person representation can be disentangled into identity and structure features. We conduct data augmentation in two pathways: one on id-unrelated structure features with rotated meshes, the other one on identity features with D-Mixup.

#### 3.1.1 Mesh-guided Rotation (id-unrelated augmentation)

As shown in Fig. 1. (b), given a person image and an estimated 3D mesh, we denote the 2D projection of the mesh as original structure  $s_{ori}$ . To mimic real-world camera

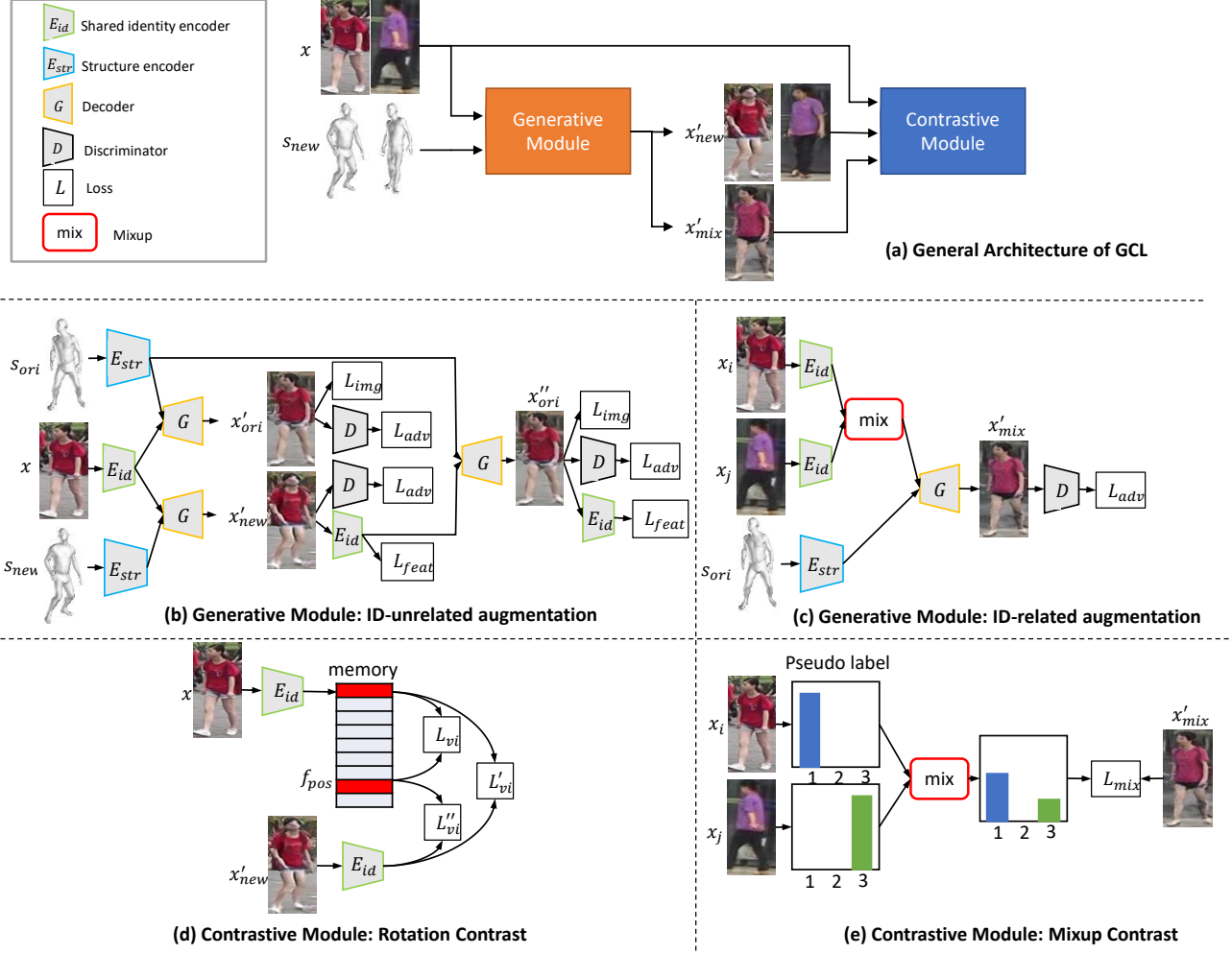


Fig. 1. **(a) General architecture of GCL+:** The framework is composed of a generative module **(b, c)** and a contrastive module **(d, e)**, which are coupled by the shared identity encoder  $E_{id}$ . **(b) Mesh rotation (id-unrelated augmentation):** The decoder  $G$  combines the identity features encoded by  $E_{id}$  and structure features  $E_{str}$  to generate an augmented view  $x'_{new}$  with a cycle consistency. **(c) D-mixup (id-related augmentation):** The decoder  $G$  generates an identity-mixed augmented view  $x'_{mix}$  with the mixed identity features. **(d) Rotation Contrast:** Viewpoint-invariance is enhanced by maximizing the agreement between original  $E_{id}(x)$ , synthesized  $E_{id}(x'_{new})$  and memory  $f_{pos}$  representations. **(e) Mixup Contrast:** A smoother decision boundary can be learnt with  $x'_{mix}$  and the interpolated pseudo label.

view-point, as shown in Table 3, we rotate the 3D mesh by  $45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ$  and  $315^\circ$  and randomly take one 2D projection from these rotated meshes as a new structure  $s_{new}$ . The unlabeled image is encoded to identity features by the identity encoder  $E_{id} : x \rightarrow f_{id}$ , while both original and new structures are encoded to structure features by the structure encoder  $E_{str} : s_{ori} \rightarrow f_{str(ori)}, s_{new} \rightarrow f_{str(new)}$ . Combining both identity and structure features, the decoder generates synthesized images  $G : (f_{id}, f_{str(ori)}) \rightarrow x'_{ori}, (f_{id}, f_{str(new)}) \rightarrow x'_{new}$ , where a prime is used to represent generated images.

As we do not have real images in new structures (paired data), a cycle consistency reconstruction [39] becomes indispensable for the generative module. We encode the generated image in the new structure  $x'_{new}$  and decode once again to get synthesized images in original structures  $G(E_{id}(x'_{new}), s_{ori}) \rightarrow x''_{ori}$ , where double primes denote cycle-generated images. We calculate a  $\ell_1$  image reconstruction loss between the original image  $x$ , the generated image

$x'_{ori}$  and the cycle-generated image:

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

To enhance the disentanglement in the cycle consistency reconstruction, we also calculate a  $\ell_1$  feature reconstruction loss:

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

The discriminator  $D$  attempts to distinguish between real and generated images with adversarial losses:

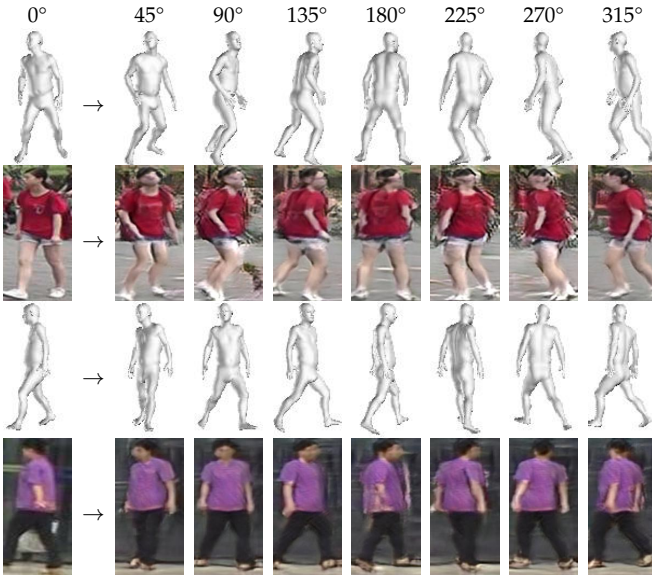
$$\mathcal{L}_{adv} = \mathbb{E}[\log D(x) + \log(1 - D(x'_{ori}))] + \mathbb{E}[\log D(x) + \log(1 - D(x'_{new}))] + \mathbb{E}[\log D(x) + \log(1 - D(x''_{ori}))]. \quad (3)$$

**Remark.** As shown in Fig. 2, we can switch 2D gray images [17], [25], switch meshes between random persons or rotate one's own mesh to introduce new structures for generation guidance. Although stronger pose and view-point variances can be introduced into generation, random



TABLE 3

Examples of 3D mesh guided generation on Market-1501 dataset. Each mesh is rotated by  $45^\circ$ ,  $90^\circ$ ,  $135^\circ$ ,  $180^\circ$ ,  $225^\circ$ ,  $270^\circ$  and  $315^\circ$ .



switching hinders conservation of body shape information. After testing, we find that the most appropriate way to preserve body shape and generate accurate images is Mesh rotation, which yields higher performance in Table 4.

### 3.1.2 D-mixup (id-related augmentation)

As shown in Fig. 1. (c), given two random person images  $x_i$  and  $x_j$  in a mini-batch, we encode the images into identity features  $E_{id}(x_i) \rightarrow f_{id(i)}$  and  $E_{id}(x_j) \rightarrow f_{id(j)}$ . We follow the original Mixup [26] in using a Beta distribution with a hyper-parameter  $\alpha$  to randomly sample a mixing coefficient  $\lambda$ :

$$\begin{aligned} \lambda &= \text{Beta}(\alpha, \alpha), \lambda^* = \max(\lambda, 1 - \lambda) \\ f_{id(mix)} &= \lambda^* \cdot f_{id(i)} + (1 - \lambda^*) \cdot f_{id(j)}, \end{aligned} \quad (4)$$

where  $\lambda^*$  renders the mixed identity more similar to  $x_i$ . To conserve corresponding body shape information, we use the original structure of  $x_i$ , rather than  $x_j$  as the generation guidance. A mixed person image (see more interpolated examples in Fig. 3) can be generated by combining mixed identity features and original structure features  $G(f_{id(mix)}, s_{ori(i)}) \rightarrow x'_{mix}$ . The discriminator  $D$  attempts to distinguish between real and mixed images with the adversarial loss:

$$\mathcal{L}_{adv\_mix} = \mathbb{E}[\log D(x) + \log(1 - D(x'_{mix}))]. \quad (5)$$

More discussion about feature regularization losses is provided in Supplementary Materials Section A.

### 3.1.3 Overall generative loss

The overall GAN loss combines the above losses (1), (2), (3) and (5) with a weighting coefficient  $\lambda_{recon}$ :

$$\mathcal{L}_{gan} = \lambda_{recon}(\mathcal{L}_{img} + \mathcal{L}_{feat}) + \mathcal{L}_{adv} + \mathcal{L}_{adv\_mix}. \quad (6)$$

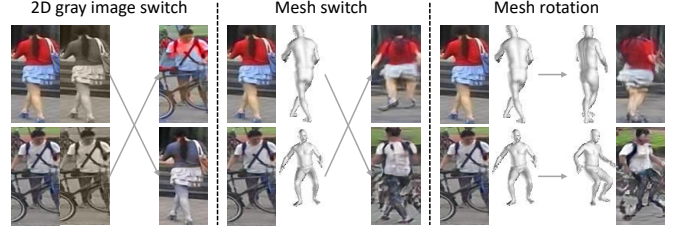


Fig. 2. Different ways of introducing structural variance (2D gray image switch [25], Mesh switch and Mesh rotation) into generation.

TABLE 4

Performance comparison of rotating one mesh and switching two random meshes in the generation.

Method	Duke→Market		Market→Duke	
	mAP	Rank1	mAP	Rank1
2D gray image switch [25]	60.1	78.8	59.5	76.2
Mesh switch	74.2	88.5	60.6	76.9
Mesh rotation	<b>74.4</b>	<b>89.7</b>	<b>61.3</b>	<b>78.0</b>

## 3.2 Contrastive Module

The described generative module generates augmented views of a person image, which can form positive view pairs for the contrastive module. By maximizing similarity between positive pairs, the shared identity encoder is aimed at building robust representations that are invariant to distortions. For one identity, there are commonly several positive images in the dataset, which are recorded in different poses, camera styles and backgrounds. Only maximizing similarity between an image and its self-augmented views leads to sub-optimal performance. Moreover, previous methods [10], [11] have demonstrated the effectiveness of mining a large number of negative samples in contrastive learning.

In order to mine more positives and a large number of negatives, we generate pseudo labels on a memory bank [30] that stores all representations  $\mathcal{M}$  corresponding to dataset images  $\mathcal{X}$ . Given a representation  $f^t$  in the current epoch, the corresponding memory bank representation  $\mathcal{M}[i]$  is updated with a momentum hyper-parameter  $\beta$ :

$$\mathcal{M}[i]^t = \beta \cdot \mathcal{M}[i]^{t-1} + (1 - \beta) \cdot f^t, \quad (7)$$

where  $\mathcal{M}[i]^t$  and  $\mathcal{M}[i]^{t-1}$  respectively refer to the memory bank representations in the  $t$  and  $t - 1$  epochs. The memory bank stores moving averaged representations, which stabilize the pseudo label generation. To further enhance the pseudo label quality, we compute k-reciprocal re-ranked Jaccard distance [62] between memory bank representations, which are then fed into a clustering algorithm DBSCAN [63] to generate pseudo labels  $\mathcal{Y} = \{y_1, y_2, \dots, y_N\}$ . During the training, the pseudo labels are renewed at the beginning of each epoch. We design a Rotation Contrast and a Mixup Contrast respectively for the two types of generated views.

### 3.2.1 Rotation Contrast (for id-unrelated augmentation)

As shown in Fig. 1. (d), the original image  $x$  and the generated image  $x'_{new}$  are encoded by the shared identity encoder into two identity feature vectors  $E_{id}(x) \rightarrow f$  and  $E_{id}(x'_{new}) \rightarrow f'_{new}$ . For a representation  $f$  with a pseudo label  $y_i$ , we randomly sample a positive representation  $f_{pos}$

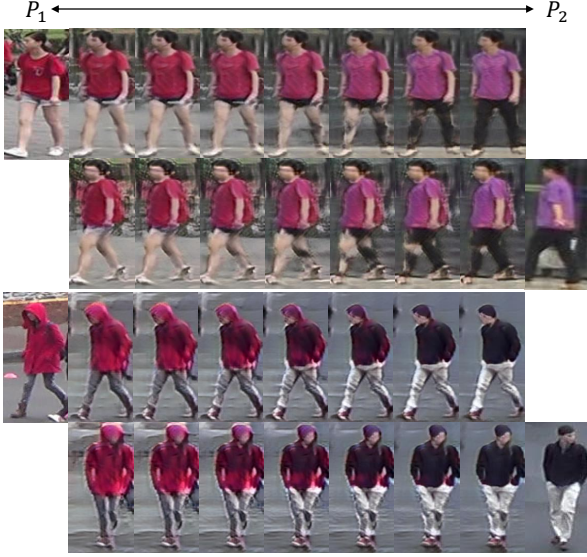


Fig. 3. Linear interpolation of disentangled identity features between two persons respectively from Market-1501 and DukeMTMC-reID.

of the same pseudo label  $y_i$  and  $K$  negative representations of pseudo labels different to  $y_i$  from the memory bank. Three positive pairs can be formed, i.e.,  $(f, f_{pos})$ ,  $(f, f'_{new})$  and  $(f_{pos}, f'_{new})$ . The  $f'_{new}$  and sampled  $K$  negative representations from the memory bank are used to form  $K$  negative pairs. We define three view-invariant losses to attract three positive pairs while repulsing  $K$  negative pairs:

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

$$\mathcal{L}'_{vi} = \mathbb{E}[\log(1 + \frac{\sum_{i=1}^K \exp(\langle f'_{new} \cdot k_i \rangle / \tau)}{\exp(\langle f'_{new} \cdot f \rangle / \tau)})], \quad (9)$$

$$\mathcal{L}''_{vi} = \mathbb{E}[\log(1 + \frac{\sum_{i=1}^K \exp(\langle f'_{new} \cdot k_i \rangle / \tau)}{\exp(\langle f'_{new} \cdot f_{pos} \rangle / \tau)})], \quad (10)$$

where  $\langle \cdot \rangle$  denotes the cosine similarity between two feature vectors.  $\tau$  is a temperature hyper-parameter to sharpen the cosine similarity.  $k_i$  denotes negative representations sampled from the memory bank. Presented three loss functions enable the contrastive module to maximize the similarity between original view  $f$ , generated view  $f'_{new}$  and positive memory view  $f_{pos}$ . At the same time, the similarity between generated view  $f'_{new}$  and  $K$  negative memory views is minimized, which encourages the generative module to refine the generated view  $f'_{new}$  that should be different from a large number of negative samples.

### 3.2.2 Mixup Contrast (for id-related augmentation)

The mixed image  $x'_{mix}$  is encoded by the shared identity encoder into a mixed identity feature vector  $E_{id}(x'_{mix}) \rightarrow f'_{mix}$ , see Fig. 1. (e). Towards learning a smoother decision boundary between two clusters, as illustrated in Fig. 4, we design a Mixup Contrast for  $f'_{mix}$ . As certain instances in a cluster are close to the decision boundary between two

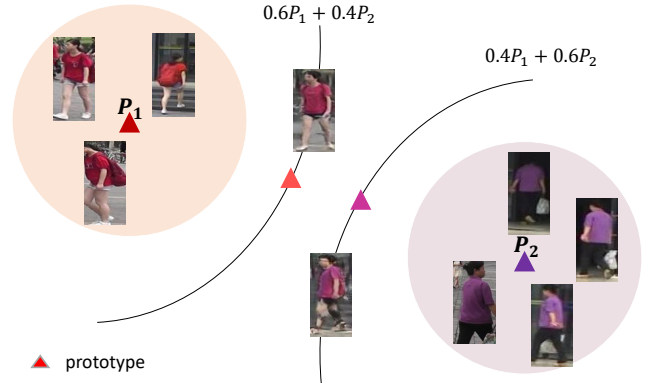


Fig. 4. Mixup Contrast targets at learning a smoother decision boundary between two persons  $P_1$  and  $P_2$  by contrasting in-between samples with in-between prototypes.

clusters, whereas the others are far away, we define an averaged prototype for a cluster:

$$p_a = \frac{1}{N_a} \sum_{\mathcal{M}[i] \in y_a} \mathcal{M}[i], \quad (11)$$

where  $N_a$  is the number of instances belonging to the cluster  $a$ .

Given a random image representation  $f$ , we use a softmax cross-entropy loss  $\mathcal{L}_{proto}$  to make  $f$  converge to the cluster prototype, which encourages the compactness of a cluster:

$$\mathcal{L}_{proto} = \mathbb{E}[\log(1 + \frac{\sum_{i=1}^{|\mathcal{Y}|-1} \exp(f \cdot p_i)}{\exp(f \cdot p_+)})], \quad (12)$$

where  $p_+$  is the corresponding prototype of  $f$  and  $p_i$  denotes other cluster prototypes.  $|\mathcal{Y}|$  is the number of clusters. Given that certain clusters may contain more instances that are close to decision boundaries with other clusters, compact clusters provide stable mixed prototypes.

Based on the pseudo labels, we define a mixed prototype vector between two clusters  $i$  and  $j$ :

$$p_{mix} = \lambda^* \cdot p_i + (1 - \lambda^*) \cdot p_j, \quad (13)$$

where  $\lambda^*$  is the same mixing coefficient as in Eq. (4).

For the mixed representation  $f'_{mix}$ , we use another softmax cross-entropy loss to maximize its similarity with the mixed prototype  $p_{mix}$  and minimize its similarity with  $|\mathcal{Y}| - 2$  negative prototypes that do not belong to the two clusters  $i$  and  $j$ :

$$\mathcal{L}_{mix} = \mathbb{E}[\log(1 + \frac{\sum_{i=1}^{|\mathcal{Y}|-2} \exp(f'_{mix} \cdot p_i)}{\exp(f'_{mix} \cdot p_{mix})})]. \quad (14)$$

As opposed to cosine similarity in Eq. (8), (9) and (10), we do not compute normalized similarity, as the average operation for computing prototype vectors performs as normalization.

### 3.2.3 Overall contrastive loss

The overall contrastive loss combines the above losses (8), (9), (10), (12) and (14):

$$\mathcal{L}_{contrast} = \lambda_{vi}(\mathcal{L}_{vi} + \mathcal{L}'_{vi} + \mathcal{L}''_{vi}) + \lambda_{mix}(\mathcal{L}_{proto} + \mathcal{L}_{mix}). \quad (15)$$

### 3.3 Joint Training

Our proposed framework incorporates a generative module and a contrastive module. The generative module disentangles a person image representation into identity and structure features, which allows for learning purified identity features for person ReID. The contrastive module learns invariance via contrasting augmented images. If we replace the GAN-based augmentation with traditional data augmentation techniques, both modules can be trained separately. However, a separate training leads to sub-optimal performance for both of them. To address this issue, we couple the two modules with a shared identity encoder in a joint training framework. In the setting of joint training, both modules work collaboratively to achieve one objective: enhancing the discriminability of identity representations. Inside GCL+, the generative module provides both, id-unrelated and id-related augmentations for the contrastive module. On the other hand, the contrastive module maximizes the similarity between positive views, while repulsing negative views, which, in turn, refines the identity representations for a better generation quality. Both modules mutually promote each other’s performance in the joint training, leading to an optimal ReID performance. In our proposed framework, a forward propagation is firstly conducted on the generative module and subsequently on the contrastive module. A backward propagation is then conducted with an overall loss that combines Eq. (6) and Eq. (15):

$$\mathcal{L}_{overall} = \mathcal{L}_{gan} + \mathcal{L}_{contrast}. \quad (16)$$

## 4 EXPERIMENT

### 4.1 Datasets and Evaluation Protocols

We evaluate our proposed method GCL+ on five mainstream person ReID benchmarks, including three image-based datasets: Market-1501 [64], DukeMTMC-reID [65], MSMT17 [14] and two video-based datasets: MARS [66] and DukeMTMC-VideoReID [67]. **Market-1501** dataset is collected in front of a supermarket in Tsinghua University from 6 cameras. It is composed of 12,936 images of 751 identities for training and 19,732 images of 750 identities for testing. **DukeMTMC-reID** is collected from 8 cameras installed in the campus of Duke University. It contains 16,522 images of 702 persons for training, 2,228 query images and 17,661 gallery images of 702 persons for testing. **MSMT17** is a large-scale Re-ID dataset, which includes 32,621 training images of 1,041 identities and 93,820 testing images of 3,060 identities collected from 15 cameras deployed in both indoor and outdoor scenes. **MARS** is a large-scale video-based person ReID dataset. The dataset contains 17,503 tracklets of 1,261 identities collected from 6 cameras, where 625 identities are used for training and the other 636 identities are used for testing. **DukeMTMC-VideoReID** is a video-based person ReID dataset derived from DukeMTMC [65] dataset. DukeMTMC-VideoReID contains 2,196 training tracklets of 702 identities and 2,636 testing tracklets of other 702 identities.

As our method includes a GAN and a contrastive module, we report results for both unsupervised person ReID and generation quality evaluations. For unsupervised person ReID evaluation, we provide results under both,

unsupervised domain adaptation and fully unsupervised settings. We report both, Cumulative Matching Characteristics (CMC) at Rank1, Rank5, Rank10 accuracies, as well as mean Average Precision (mAP) on the testing set. For the generation quality evaluation, we conduct a qualitative comparison between our method and state-of-the-art methods on generated images.

### 4.2 Implementation details

We introduce implementation details pertained to network design and general training configurations, as well as three-step optimization.

**Network design.** Our network design related to the identity encoder  $E_{id}$ , the structure encoder  $E_{str}$ , the decoder  $G$  and the discriminator  $D$  has been mainly inspired by [17], [25]. In the following descriptions, we denote the size of feature maps in channel×height×width. **1)**  $E_{id}$  is an ImageNet [35] pre-trained ResNet50 [68] with slight modifications. The original fully connected layer is replaced by a batch normalization layer and a fully connected embedding layer, which outputs identity representations  $f$  in  $512 \times 1 \times 1$  for the contrastive module. In parallel, we add a part average pooling that outputs identity features  $f_{id}$  in  $2048 \times 4 \times 1$  for the generative module. **2)**  $E_{str}$  is composed of four convolutional and four residual layers, which output structure features  $f_{str}$  in  $128 \times 64 \times 32$ . **3)**  $G$  contains four residual and four convolutional layers. Every residual layer contains two adaptive instance normalization layers [18] that transform  $f_{id}$  into scale and bias parameters. **4)**  $D$  is a multi-scale PatchGAN [19] discriminator at  $64 \times 32$ ,  $128 \times 64$  and  $256 \times 128$ .

**General training configurations.** Our framework is implemented under Pytorch [69] and trained with one Nvidia V100 GPU. The inputs are resized to  $256 \times 128$ . We empirically set a large weight  $\lambda_{recon} = 5$  for reconstruction in Eq. (6). With a batch size of 16, we use SGD to train  $E_{id}$  and Adam optimizer to train  $E_{str}$ ,  $G$  and  $D$ . Learning rate in Adam is set to  $1 \times 10^{-4}$  and  $3.5 \times 10^{-4}$  in SGD and are multiplied by 0.1 after 10 epochs. DBSCAN maximal neighborhood distance is set to 0.5 and minimal sample number is set to 4. The number of negatives  $K$  is 8192. For testing,  $E_{id}$  outputs representations  $f$  of dimension 512. For video-based person ReID, due to our GPU memory constraint, we randomly sample 2 frames per tracklet on MARS and 8 frames per tracklet on DukeMTMC-VideoReID for training. For testing, all the frames from each tracklet are used to calculate a unified tracklet representation for similarity ranking. Other settings are kept the same as image-based person ReID settings.

**Three-stage optimization.** To reduce the noise from imperfect generated images at early epochs, we train the four modules  $E_{id}$ ,  $E_{str}$ ,  $G$  and  $D$  in a three-stage optimization. **Stage 1**  $E_{id}$  warm-up: we use a state-of-the-art unsupervised ReID method to warm up  $E_{id}$ , e.g., ACT [55], MMCL [59] and JVTc [60]. **Stage 2**  $E_{str}$ ,  $G$  and  $D$  warm-up: we freeze  $E_{id}$  and warm up  $E_{str}$ ,  $G$ , and  $D$  only with the overall GAN loss in Eq. (6) for 40 epochs. **Stage 3** joint training: we bring in the memory bank and the pseudo labels to jointly train the whole framework with the overall loss in Eq. (16) for another 20 epochs.



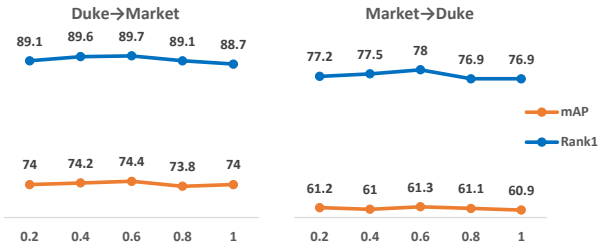


Fig. 5. Hyper-parameter analysis on  $\alpha$  for mixup coefficient on Duke  $\rightarrow$  Market and Market  $\rightarrow$  Duke tasks.

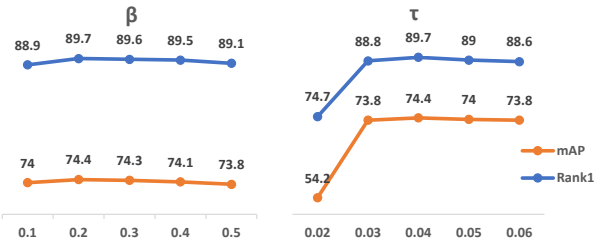


Fig. 6. Hyper-parameter analysis on  $\beta$  for memory momentum and  $\tau$  for contrastive temperature on Duke  $\rightarrow$  Market task.

### 4.3 Unsupervised ReID Evaluation

To validate the effectiveness of each component, we conduct parameter analysis and ablation experiments with a JVTC [60] baseline. As JVTC+ is the enhanced version of JVTC with a camera temporal distribution post-processing, the performance boost from the post-processing is almost fixed. Thus, the ablation experiments show similar variance with JVTC and JVTC+ baselines. We further compare our method with state-of-the-art unsupervised person ReID with three different baselines to show the generalizability of our method.

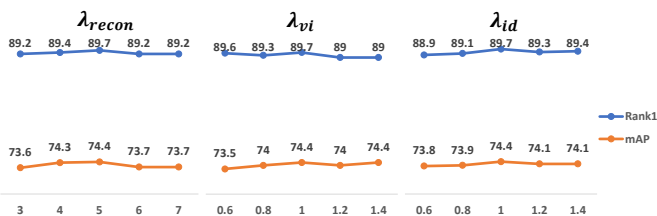


Fig. 7. Hyper-parameter analysis on balancing coefficients  $\lambda_{recon}$  for reconstruction weight,  $\lambda_{vi}$  for rotation contrast weight and  $\lambda_{mix}$  for mixup contrast weight on Duke  $\rightarrow$  Market task.

TABLE 5

Performance under different clustering neighborhood distance threshold. 'N' is the approximate number of pseudo-identities.

Threshold	Duke $\rightarrow$ Market			Market $\rightarrow$ Duke		
	N	mAP	Rank1	N	mAP	Rank1
0.4	~642	74.5	89.4	~840	60.9	77.1
0.45	~605	74.4	89.4	~810	61.2	77.4
0.5	~584	74.4	89.7	~786	61.3	78.0
0.55	~540	73.6	88.4	~744	61.1	76.8
0.6	~500	72.4	87.6	~697	60.7	77.7

#### 4.3.1 Parameter analysis

Hyper-parameters, such as mixing coefficient  $\alpha$ , memory momentum  $\beta$  and view-invariant contrastive loss temperature  $\tau$ , play important roles inside our proposed GCL+ framework for better unsupervised person ReID performance. We vary their values to analyze the sensitivity of each hyper-parameter inside our proposed framework GCL+.

For Beta distribution, a larger  $\alpha$  results in a higher possibility that  $\lambda$  gets closer to 0.5. ReID performance on both Duke  $\rightarrow$  Market and Market  $\rightarrow$  Duke tasks with reference to  $\alpha$  is reported in Fig. 5. On both tasks, the optimal performance is achieved, in case of  $\alpha$  is around 0.6. As a consequence,  $\alpha$  is set to 0.6 in our framework.

The value of  $\beta$  controls the memory updating speed. The value of  $\tau$  amplifies the cosine similarity between contrastive views. An overlarge or undersized value, generally speaking, introduces more noise for contrastive learning. We report the performance variation with reference to  $\beta$  and  $\tau$  on Duke  $\rightarrow$  Market task in Fig. 6. We find that the performance is more sensitive to the similarity temperature  $\tau$ . Based on the results, we set  $\beta$  to 0.2 and  $\tau$  to 0.04.

The number of possible pseudo-identities  $N$  is related to clustering hyper-parameters, such as maximal neighborhood distance threshold and minimal cluster sample number. The distance threshold of DBSCAN is the maximal distance between two samples for one to be considered as in the neighborhood of the other. A larger distance threshold enlarges the radius of a cluster, making more samples be considered into a same cluster ( $N$  becomes smaller). As shown in Table 5, the threshold value only slightly affects ReID performance.

As our framework jointly optimize the generative and contrastive modules, we set weight coefficients to balance different loss functions in the two modules. We vary the balancing coefficients  $\lambda_{recon}$ ,  $\lambda_{vi}$  and  $\lambda_{mix}$  in Equation (6) and (15). The corresponding results are reported in Fig. 7. Overall, the different values in the tested range only slightly influence the final results. Based on the results, we set  $\lambda_{recon} = 5$ ,  $\lambda_{vi} = 1$  and  $\lambda_{mix} = 1$ .

#### 4.3.2 Ablation study

Contrastive learning methods strongly rely on data augmentation to create different augmented views for contrasting. Our proposed GCL+ outperforms traditional contrastive learning methods by replacing traditional data augmentation techniques with GAN-based augmentation techniques. To validate the effectiveness of our proposed GAN-based augmentation techniques and contrastive losses, we conduct ablation experiments on both Market-1501 and DukeMTMC-reID datasets.

**Data augmentation.** Data augmentation techniques can be categorized into id-unrelated and id-related augmentation. Id-unrelated augmentation creates intra-image visual distortions. In contrast, id-related augmentation creates inter-image visual distortions, which affects image identities. We compare results of traditional and generative data augmentation under fully unsupervised setting and domain adaptation setting in Table 6. For traditional data augmentation, we use multiple popular person ReID

TABLE 6

Ablation study under fully unsupervised and UDA settings on traditional (w/o GAN) and generative (w/ GAN) data augmentation for the contrastive module. ‘Multi’ refers to multiple commonly used data augmentation techniques for person ReID, including random flipping, padding, cropping and erasing. ‘Rotation’ refers to our proposed mesh-guided rotation. ‘Mixup’ is conducted on image level, while ‘F-Mixup’ is conducted on feature level.

Fully unsupervised	ID-unrelated		ID-related			Market				Duke			
	Multi	Rotation	Mixup	F-Mixup	D-Mixup	mAP	R1	R5	R10	mAP	R1	R5	R10
w/o GAN	Baseline					47.2	75.4	86.7	90.5	43.9	66.8	77.6	81.0
	✓					58.2	81.1	91.0	93.5	50.8	70.8	80.9	83.8
	✓		✓			60.0	82.5	91.6	94.0	51.0	71.1	80.8	84.1
w/ GAN		✓				63.8	83.4	91.8	94.3	53.1	72.8	81.2	83.7
		✓	✓			65.9	84.8	92.5	94.3	54.3	73.6	82.5	84.9
		✓		✓		66.1	84.3	92.4	94.6	54.2	73.7	82.4	85.5
		✓			✓	<b>66.3</b>	<b>85.3</b>	<b>92.9</b>	<b>94.6</b>	<b>54.6</b>	<b>74.2</b>	<b>82.8</b>	<b>85.6</b>
UDA	ID-unrelated		ID-related			Duke→Market				Market→Duke			
	Multi	Rotation	Mixup	F-Mixup	D-Mixup	mAP	R1	R5	R10	mAP	R1	R5	R10
w/o GAN	Baseline					65.0	85.7	93.4	95.9	56.5	73.9	84.4	87.8
	✓					70.4	86.9	94.3	95.8	57.0	74.2	84.2	87.2
	✓		✓			70.7	87.8	94.1	96.3	57.7	74.5	85.0	88.0
w/ GAN		✓				72.5	88.7	94.8	96.3	59.9	75.9	86.2	88.5
		✓	✓			73.0	88.9	94.8	96.4	60.4	76.5	85.9	88.3
		✓		✓		72.7	88.8	95.1	96.3	60.2	76.7	86.1	88.1
		✓			✓	<b>74.4</b>	<b>89.7</b>	<b>95.5</b>	<b>96.7</b>	<b>61.3</b>	<b>78.0</b>	<b>86.8</b>	<b>89.1</b>

TABLE 7

Ablation study on three view-invariant losses in Rotation Contrast and two prototype losses in Mixup Contrast.

$\mathcal{L}_{vi}$	$\mathcal{L}'_{vi}$	$\mathcal{L}''_{vi}$	$\mathcal{L}_{proto}$	$\mathcal{L}_{mix}$	Duke→Market		Market→Duke	
					mAP	R1	mAP	R1
✓					61.6	82.4	51.7	70.6
✓	✓				69.1	85.6	58.3	74.8
✓	✓	✓			72.5	88.7	59.9	75.9
✓	✓	✓	✓		72.8	88.8	60.6	76.9
✓	✓	✓	✓	✓	<b>74.4</b>	<b>89.7</b>	<b>61.3</b>	<b>78.0</b>

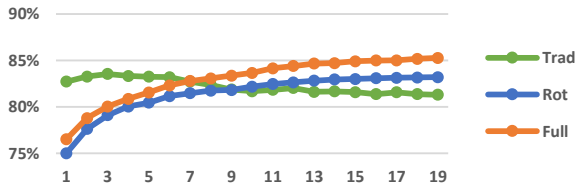


Fig. 8. Normalized Mutual Information (NMI) during 20 joint training epochs on Market-1501. ‘Trad’ refers to traditional data augmentation techniques. ‘Rot’ refers to id-unrelated mesh-guided rotation. ‘Full’ refers to combining id-unrelated mesh-guided rotation and id-related D-Mixup.

data augmentation techniques, including random flipping, padding, cropping and erasing [12], as id-unrelated augmentation and Mixup [26] as id-related augmentation. Even with these traditional data augmentation, our contrastive module significantly outperforms the baseline. When we replace traditional data augmentation with generative data augmentation, the unsupervised person ReID performance can be further improved. Our proposed mesh-guided rotation (Rotation) works better than the multiple commonly used data augmentation techniques (Multi) for id-unrelated augmentation. Meanwhile, our proposed D-Mixup achieves better performance than the image-level Mixup and feature-level Mixup (F-Mixup) for id-related augmentation.

**Effects on pseudo labels.** Robust identity representations should have a better intra-class compactness and inter-class separability, which leads to better pseudo label quality. We evaluate our pseudo label quality by measuring the Normalized Mutual Information (NMI) [71] between our pseudo labels and ground truth labels. As illustrated in Fig. 8, traditional data augmentation (Trad) works well at

the beginning, but ends up in a worse quality. We argue that traditional data augmentation brings to the fore undesirable distortions on identity features, which easily leads to over-fitting for id-sensitive tasks. Deviating from that, GAN-based augmentation introduces more noise at the beginning, however avoids over-fitting in the final training epochs. In addition, our full GCL+ (Full) conducts both GAN-based id-unrelated and id-related augmentation, which achieves better pseudo label quality than only id-unrelated mesh-guided rotation (Rot).

**Contrastive loss.** To learn maximal invariance from generated image and memory stored image, we have formed three positive pairs for Rotation Contrast, namely  $(f, f_{pos})$ ,  $(f, f'_{new})$  and  $(f_{pos}, f'_{new})$ . By maximizing the similarity between these three positive pairs in Equation (8), (9) and (10), our objective is to build identity representations, which are invariant to instance-level pose, view-point and background variance. Meanwhile, we use identity prototypes and mixed prototypes in Mixup Contrast to learn a smoother class-level decision boundary with Equation (12) and (14). To confirm the contribution from these contrastive losses, we gradually add each into our framework and report the corresponding results in Table 7. The results indicate that our proposed contrastive losses effectively contribute to learning robust representations for unsupervised person ReID.

#### 4.3.3 Comparison with state-of-the-art methods

**Image-based person ReID.** We compare our proposed GCL+ with state-of-the-art unsupervised ReID methods under three purely unsupervised and four unsupervised domain adaptation evaluation protocols. We evaluate the performance of GCL+ with different baselines, including MMCL [59], JVTC [60] and ACT [55], to demonstrate the generalizability of our proposed method.

Under the fully unsupervised setting, we report associated results on Market-1501, DukeMTMC-reID and MSMT17 dataset in Table 8. We firstly provide results of state-of-the-art methods, including BUC [57], SoftSim [58], TSSL [61], MMCL [59], JVTC [60], JVTC+ [60], Meta-Cam [70], as well as our previous work GCL [9], on the three datasets. Our proposed method GCL+ significantly improves the unsupervised person ReID performance from

TABLE 8

Comparison of fully unsupervised ReID methods (%) on Market1501, DukeMTMC-reID and MSMT17 datasets. We test our proposed method on several baselines, see names in parentheses.

Method	Reference	Market1501				DukeMTMC-reID				MSMT17			
		mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
BUC [57]	AAAI'19	29.6	61.9	73.5	78.2	22.1	40.4	52.5	58.2	-	-	-	-
SoftSim [58]	CVPR'20	37.8	71.7	83.8	87.4	28.6	52.5	63.5	68.9	-	-	-	-
TSSL [61]	AAAI'20	43.3	71.2	-	-	38.5	62.2	-	-	-	-	-	-
MMCL [59]	CVPR'20	45.5	80.3	89.4	92.3	40.2	65.2	75.9	80.0	11.2	35.4	44.8	49.8
JVTC [60]	ECCV'20	41.8	72.9	84.2	88.7	42.2	67.6	78.0	81.6	15.1	39.0	50.9	56.8
JVTC+ [60]	ECCV'20	47.5	79.5	89.2	91.9	50.7	74.6	82.9	85.3	17.3	43.1	53.8	59.4
MetaCam [70]	CVPR'21	61.7	83.9	92.3	-	53.8	73.8	84.2	-	15.5	35.2	48.3	-
GCL(MMCL) [9]	CVPR'21	54.9	83.7	91.6	94.0	49.3	69.7	79.7	82.8	-	-	-	-
GCL(JVTC) [9]	CVPR'21	63.4	83.7	91.6	94.3	53.3	72.4	82.0	84.9	18.0	41.6	53.2	58.4
GCL(JVTC+) [9]	CVPR'21	66.8	87.3	93.5	95.5	62.8	82.9	87.1	88.5	21.3	45.7	58.6	64.5
GCL+(MMCL)	This paper	56.0	84.0	91.4	93.7	49.5	70.2	80.2	83.3	-	-	-	-
GCL+(JVTC)	This paper	66.3	85.3	92.9	94.6	54.6	74.2	82.8	85.6	19.2	44.7	56.4	61.4
GCL+(JVTC+)	This paper	<b>69.3</b>	<b>89.0</b>	<b>94.6</b>	<b>96.0</b>	<b>63.5</b>	<b>83.1</b>	<b>87.4</b>	<b>88.8</b>	<b>22.0</b>	<b>47.9</b>	<b>61.3</b>	<b>67.1</b>

TABLE 9

Comparison of unsupervised domain adaptive ReID methods (%) between Market1501, DukeMTMC-reID and MSMT17 datasets. We test our proposed method on several baselines, see names in parentheses.

Method	Reference	Duke→Market				Market→Duke				Market→MSMT17				Duke→MSMT17			
		mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
ECN [7]	CVPR'19	43.0	75.1	87.6	91.6	40.4	63.3	75.8	80.4	8.5	25.3	36.3	42.1	10.2	30.2	41.5	46.8
PDA [21]	ICCV'19	47.6	75.2	86.3	90.2	45.1	63.2	77.0	82.5	-	-	-	-	-	-	-	-
CR-GAN [41]	ICCV'19	54.0	77.7	89.7	92.7	48.6	68.9	80.2	84.7	-	-	-	-	-	-	-	-
SSG [54]	ICCV'19	58.3	80.0	90.0	92.4	53.4	73.0	80.6	83.2	13.2	31.6	49.6	-	13.3	32.2	51.2	-
MMCL [59]	CVPR'20	60.4	84.4	92.8	95.0	51.4	72.4	82.9	85.0	15.1	40.8	51.8	56.7	16.2	43.6	54.3	58.9
ACT [55]	AAAI'20	60.6	80.5	-	-	54.5	72.4	-	-	-	-	-	-	-	-	-	-
DG-Net++ [17]	ECCV'20	61.7	82.1	90.2	92.7	63.8	78.9	87.8	90.4	22.1	48.4	60.9	66.1	22.1	48.8	60.9	65.9
JVTC [60]	ECCV'20	61.1	83.8	93.0	95.2	56.2	75.0	85.1	88.2	19.0	42.1	53.4	58.9	20.3	45.4	58.4	64.3
ECN+ [56]	TPAMI'20	63.8	84.1	92.8	95.4	54.4	74.0	83.7	87.4	15.2	40.4	53.1	58.7	16.0	42.5	55.9	61.5
JVTC+ [60]	ECCV'20	67.2	86.8	95.2	97.1	66.5	80.4	89.9	92.2	25.1	48.6	65.3	68.2	27.5	52.9	70.5	75.9
MMT [8]	ICLR'20	71.2	87.7	94.9	96.9	65.1	78.0	88.8	92.5	22.9	49.2	63.1	68.8	23.3	50.1	63.9	69.8
CAIL [50]	ECCV'20	71.5	88.1	94.4	96.2	65.2	79.5	88.3	91.4	20.4	43.7	56.1	61.9	24.3	51.7	64.0	68.9
MetaCam [70]	CVPR'21	76.5	90.1	-	-	65.0	79.5	-	-	-	-	-	-	-	-	-	-
GCL(ACT) [9]	CVPR'21	66.7	83.9	91.4	93.4	55.4	71.9	81.6	84.6	-	-	-	-	-	-	-	-
GCL(JVTC) [9]	CVPR'21	73.4	89.1	95.0	96.6	60.4	77.2	86.2	88.4	21.5	45.0	57.1	66.5	24.9	50.8	63.4	68.9
GCL(JVTC+) [9]	CVPR'21	75.4	90.5	96.2	97.1	67.6	81.9	88.9	90.6	27.0	51.1	63.9	69.9	29.7	54.4	68.2	74.2
GCL+(ACT)	This paper	67.5	84.3	92.6	94.2	56.8	73.5	82.8	85.1	-	-	-	-	-	-	-	-
GCL+(JVTC)	This paper	74.4	89.7	95.5	96.7	61.3	78.0	86.8	89.1	23.0	48.3	60.6	65.8	25.5	52.7	65.2	70.2
GCL+(JVTC+)	This paper	<b>76.5</b>	<b>91.6</b>	<b>96.3</b>	<b>97.6</b>	<b>68.3</b>	<b>82.6</b>	<b>89.4</b>	<b>91.2</b>	<b>27.8</b>	<b>53.8</b>	<b>66.9</b>	<b>72.5</b>	<b>31.5</b>	<b>57.9</b>	<b>70.3</b>	<b>76.1</b>

the three baselines MMCL, JVTC and JVTC+. The proposed new D-Mixup and Mixup Contrast in our framework GCL+ consistently surpasses the performance of our previous work GCL with the three different baselines. With the strong baseline JVTC+, our method achieves state-of-the-art performance on the three datasets.

Under the unsupervised domain adaptation setting, we report related results on four mainstream benchmarks, including Duke→Market, Market→Duke, Market→MSMT17 and Duke→MSMT17 in Table 9. Our proposed method GCL+ additionally achieves better performance than state-of-the-art methods, including ECN [7], PDA [21], CR-GAN [41], SSG [54], MMCL [59], ACT [55], DG-Net++ [17], JVTC [60], ECN+ [56], JVTC+ [60], MMT [8], CAIL [50], MetaCam [70], as well as our previous work GCL [9]. Among these methods, PDA, CR-GAN and DG-Net++ share certain similarity with our proposed method GCL+, in that they are based on GAN. However, PDA and DG-Net++ used either 2D skeleton or random gray-scaled images as guidance, which could not preserve body shape information. Further, PDA, CR-GAN and DG-Net++ did not manipulate identity features to generate in-between identity images. CAIL [50] has considered cross-domain Mixup, where interpolated structures may introduce more noise on identity

features. Our proposed D-Mixup does not suffer from such interpolated structures. In addition, cross-domain Mixup interpolates images from two domains, while our proposed D-Mixup interpolates intra-domain images, which is more flexible for fully unsupervised ReID.

**Video-based person ReID.** We compare our proposed GCL+ with state-of-the-art unsupervised video person ReID methods on MARS and DukeMTMC-VideoReID datasets. RACE [72] and EUG [67] leverage a labeled video tracklet per identity to initialize their models. These one-example video-based ReID methods can not actually be considered as unsupervised. DAL [73], TAUDL [74] and UTAL [75] utilize camera labels of each tracklet and try to associate tracklets of a same person across different cameras. OIM [76], BUC [57] and TSSL [61] are fully unsupervised video person ReID methods. We use the fully unsupervised method BUC as our baseline. As shown in Table 10, our proposed methods GCL (view-point augmentation) and GCL+ (view-point and in-between identity augmentation) significantly outperform previous unsupervised video-based person ReID methods.

TABLE 10

Comparison with the state-of-the-art methods on two video-based re-ID datasets, MARS and DukeMTMC-VideoReID. The “Labels” column indicates the labels used in each method. “OneEx” denotes the one-example annotation per identity. “Camera” refers to camera annotation. “Baseline (BUC)” refers to our reproduced results.

Method	Labels	MARS				DukeMTMC-VideoReID			
		mAP	R1	R5	R10	mAP	R1	R5	R10
RACE [72]	OneEx	24.5	43.2	57.1	62.1	-	-	-	-
EUG [67]	OneEx	42.4	62.6	74.9	-	63.2	72.7	84.1	-
DAL [73]	Camera	23.0	49.3	65.9	72.2	-	-	-	-
TAUDL [74]	Camera	29.1	43.8	59.9	72.8	-	-	-	-
UTAL [75]	Camera	35.2	49.9	66.4	77.8	-	-	-	-
OIM [76]	None	13.5	33.7	48.1	54.8	43.8	51.1	70.5	76.2
BUC [57]	None	29.4	55.1	68.3	72.8	66.7	74.8	86.8	89.7
TSSL [61]	None	30.5	56.3	-	-	64.6	73.9	-	-
Baseline (BUC [57])	None	32.0	51.1	66.5	71.6	67.1	72.9	86.2	90.0
GCL	None	48.6	64.8	77.5	82.0	75.9	80.1	90.5	93.7
GCL+	None	50.1	66.5	78.7	82.2	76.3	80.9	91.5	94.2

## 4.4 Generation Quality Evaluation

### 4.4.1 Ablation study

We conduct a qualitative ablation study, represented in Fig. 9 to demonstrate that our proposed contrastive module can improve generative quality for person image generation. Unconditional GANs learn a data distribution via reconstruction and adversarial training of each image, which then generate new images that fit the learned distribution. However, unconditional GANs generate from features of a single image and neglect the shared features of different images of one person (or class). Conditional GANs generally use human-annotated identity labels to learn shared class-level features, which are more view-invariant. Our proposed GCL+ introduces an unsupervised way to learn view-invariant class-level features for person image generation by contrasting pseudo positive views.

We illustrate two examples respectively from the Market-1501 and DukeMTMC-reID datasets in Fig. 9 to validate the effectiveness of our proposed contrastive module for person image generation. Given a target person, a robust identity representation should contain salient features shared by the majority of observations in different view-points and poses. In the case that GCL+ is trained without  $\mathcal{L}_{contrast}$ , our generative module tends to focus only on salient features of original image (black backpack for the first example and blue jacket for the second example), while neglecting salient features of other images of the same person (yellow t-shirt for the first example and red backpack for the second example). The contrastive module ensures the consistency of identity features for generation in different poses and view-points.

### 4.4.2 Comparison with state-of-the-art methods

We conduct a qualitative comparison between our proposed method GCL+ and state-of-the-art GAN-based person ReID methods, including FD-GAN [20], IS-GAN [42], DG-NET [25] and DG-NET++ [17]. We re-implement these GAN-based person ReID methods based on their published source code and generate six images per real image of the Market-1501 dataset, as shown in Fig. 10. FD-GAN, IS-GAN and DG-Net are supervised methods, which rely on human-annotated labels to learn robust identity-level features. We observe that images generated by FD-GAN and IS-GAN suffer from evident visual blur, which may lose detailed identity information after generation. Compared

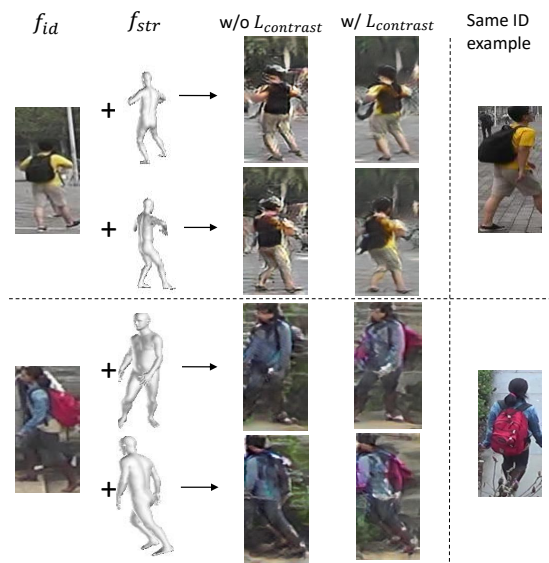


Fig. 9. Qualitative ablation study on the effectiveness of contrastive loss in Eq. (15) for generation quality.  $\mathcal{L}_{contrast}$  allows for preserving salient features from other views (yellow t-shirt for the first example and red backpack for the second example) in identity representations for generation in different poses and view-points.

TABLE 11

Examples of 3D mesh guided generation on DukeMTMC-reID dataset.







Fig. 10. Comparison of generated images on Market-1501 dataset. Examples of FD-GAN, IS-GAN, DG-Net, DG-Net++ and GCL+ are generated from same real images shown in the figure. We note that DG-Net++ and GCL+ are unsupervised methods.

TABLE 12  
Examples of 3D mesh guided generation on MSMT17 dataset.



to FD-GAN and IS-GAN, DG-Net can generate sharper images. However, using randomly switched gray-scaled images as guidance is prone to result in incoherent body shape and carrying. More comparison on the generative quality between FD-GAN, IS-GAN, DG-Net and our method is provided in Supplementary Materials Section B. As an UDA method, DG-Net++ uses cross-domain gray-scaled images as guidance, which, however, shares same problems in generation as DG-Net. Different from DG-Net++, our proposed GCL+ is a fully unsupervised ReID method, which directly augments data diversity in the target domain without the need for a labeled source domain. Moreover, an image in GCL+ is generated from its own rotated mesh, which helps to conserve body shape information and does not add extra carrying structures. The generated images from GCL+ have higher quality and similarity to real images than other methods. To validate the generative quality on DukeMTMC-reID and MSMT17 datasets, we provide more examples in Table 11 and Table 12. Consistency in the id-related space and variance in the id-unrelated space validate the purity (disentanglement quality) of identity representations in our framework GCL+. We further provide tracklet examples before and after our view-point rotation for video-based person ReID in Fig. 11. The results show that our method also works well for video-based person ReID.

#### 4.4.3 Failure case analysis

We show some failure cases from the rotation generative model in Fig. 12. Actually, when there exists inconsistent front-side and back-side patterns, the rotation-based generation can hardly generate accurate images after large rotation.

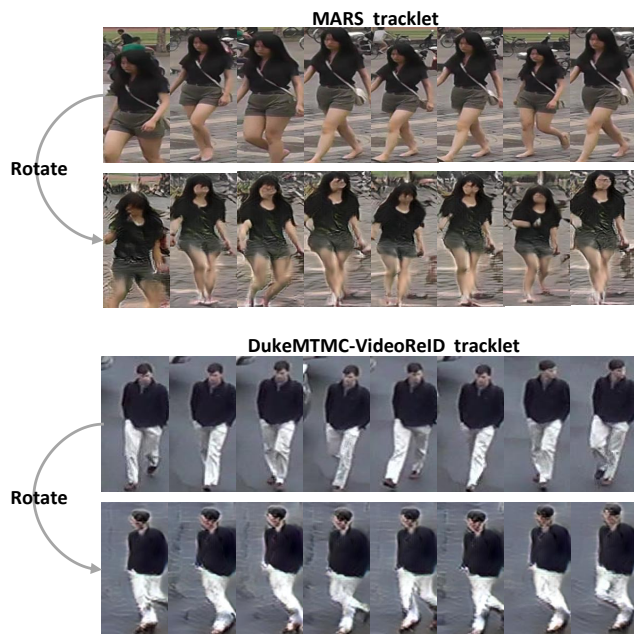


Fig. 11. Examples of tracklet frames before and after our view-point rotation. Tracklets are respectively sampled from MARS and DukeMTMC-VideoReID datasets.

For example, the model may consider visual patterns only in the back side (backpack in the first row) and patterns only in the front side (carrying objects in the second row) as whole-body appearance features for generation. One possible solution is to use a 3D human-object arrangement mesh generator [77] to help the generative model distinguish humans and objects.

## 5 CONCLUSION

In this paper, we propose an enhanced joint generative and contrastive learning (GCL+) framework for unsupervised person ReID. The framework is composed of a *generative module* for data augmentation, as well as a *contrastive module* aimed at learning invariance from generated variance. For the generative module, we propose a *3D mesh guided GAN* to realize id-unrelated and id-related augmentation by respectively rotating 3D meshes as generation guidance and interpolating two identity representations. For the contrastive





Fig. 12. Failure cases of rotation-based generation. First row: the backpack can be generated onto the front side. Second row: the carrying object can be generated onto the back side.

module, we design *Rotation Contrast* and *Mixup Contrast*, respectively for the two data augmentation techniques to learn robust identity representations. Extensive experiments are conducted to validate the superiority of the proposed GAN-based augmentation over traditional augmentation techniques for contrastive representation learning. The generative module benefits from learned robust identity representations that preserve fine-grained identity information for better generation quality. GCL+ outperforms state-of-the-art methods under both, fully unsupervised and unsupervised domain adaptation settings. Moreover, our contrastive module can be regarded as a contrastive discriminator in a GAN, which provides a new unsupervised approach for identity-preserving person image generation.

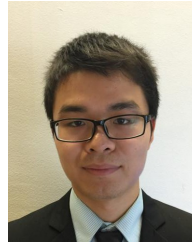
## ACKNOWLEDGMENTS

This work has been supported by the French government, through the 3IA Côte d’Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. The authors are grateful to the OPAL infrastructure from Université Côte d’Azur for providing resources and support.

## REFERENCES

- [1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, “Deep learning for person re-identification: A survey and outlook,” *IEEE TPAMI*, 2021.
- [2] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. Radke, “A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets,” *IEEE TPAMI*, 2019.
- [3] 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 *ECCV*, 2018.
- [4] H. Chen, B. Lagadec, and F. Bremond, “Learning discriminative and generalizable representations by spatial-channel partition for person re-identification,” in *WACV*, 2020.
- [5] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, “Generalizable person re-identification by domain-invariant mapping network,” in *CVPR*, June 2019.
- [6] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, “Style normalization and restitution for generalizable person re-identification,” in *CVPR*, June 2020.
- [7] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, “Invariance matters: Exemplar memory for domain adaptive person re-identification,” in *CVPR*, 2019.
- [8] Y. Ge, D. Chen, and H. Li, “Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification,” in *ICLR*, 2020.
- [9] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, “Joint generative and contrastive learning for unsupervised person re-identification,” in *CVPR*, 2021.
- [10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *ICML*, 2020.
- [11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *CVPR*, 2020.
- [12] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” in *AAAI*, 2020.
- [13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *NeurIPS*, 2014.
- [14] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer gan to bridge domain gap for person re-identification,” in *CVPR*, 2018.
- [15] S. Bak, P. Carr, and J.-F. Lalonde, “Domain adaptation through synthesis for unsupervised person re-identification,” in *ECCV*, 2018.
- [16] Z. Zhong, L. Zheng, S. Li, and Y. Yang, “Generalizing a person retrieval model hetero- and homogeneously,” in *ECCV*, 2018.
- [17] Y. Zou, X. Yang, Z. Yu, B. V. K. V. Kumar, and J. Kautz, “Joint disentangling and adaptation for cross-domain person re-identification,” in *ECCV*, 2020.
- [18] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *ICCV*, 2017.
- [19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *CVPR*, 2017.
- [20] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and H. Li, “Fd-gan: Pose-guided feature distilling gan for robust person re-identification,” in *NeurIPS*, 2018.
- [21] Y.-J. Li, C.-S. Lin, Y.-B. Lin, and Y.-C. F. Wang, “Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation,” in *ICCV*, 2019.
- [22] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *CVPR*, 2017.
- [23] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in *CVPR*, 2018.
- [24] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, “Camera style adaptation for person re-identification,” in *CVPR*, 2018.
- [25] Z. Zhong, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, “Joint discriminative and generative learning for person re-identification,” in *CVPR*, 2019.
- [26] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” in *ICLR*, 2018.
- [27] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, “Manifold mixup: Better representations by interpolating hidden states,” in *ICML*, 2019.
- [28] C. Beckham, S. Honari, V. Verma, A. M. Lamb, F. Ghadiri, R. D. Hjelm, Y. Bengio, and C. Pal, “On adversarial mixup resynthesis,” *NeurIPS*, 2019.
- [29] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *CVPR*, 2006.
- [30] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in *CVPR*, 2018.
- [31] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” in *NeurIPS*, 2020.
- [32] J.-B. Grill, F. Strub, F. Alché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, “Bootstrap your own latent: A new approach to self-supervised learning,” in *NeurIPS*, 2020.
- [33] X. Chen and K. He, “Exploring simple siamese representation learning,” in *CVPR*, 2021.
- [34] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” *arXiv preprint arXiv:2003.04297*, 2020.
- [35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” *IJCV*, 2015.
- [36] Z. Zhong, L. Zheng, and Y. Yang, “Unlabeled samples generated by gan improve the person re-identification baseline in vitro,” in *ICCV*, 2017.
- [37] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in *ICLR*, 2016.
- [38] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, “Pose-normalized image generation for person re-identification,” in *ECCV*, 2018.

- [39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *CVPR*, 2017.
- [40] Y. Huang, Q. Wu, J. Xu, and Y. Zhong, "Sbsgan: Suppression of inter-domain background shift for person re-identification," in *ICCV*, 2019.
- [41] Y. Chen, X. Zhu, and S. Gong, "Instance-guided context rendering for cross-domain person re-identification," in *ICCV*, 2019.
- [42] C. Eom and B. Ham, "Learning disentangled representation for robust person re-identification," in *NeurIPS*, 2019.
- [43] Y. Tokozume, Y. Ushiku, and T. Harada, "Between-class learning for image classification," in *CVPR*, 2018.
- [44] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *ICCV*, 2019.
- [45] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *NeurIPS*, 2019.
- [46] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *ICLR*, 2020.
- [47] M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang, "Adversarial domain adaptation with domain mixup," in *AAAI*, 2020.
- [48] Z. Zhong, L. Zhu, Z. Luo, S. Li, Y. Yang, and N. Sebe, "Openmix: Reviving known knowledge for discovering novel visual categories in an open world," in *CVPR*, 2021.
- [49] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," in *ICLR*, 2020.
- [50] C. Luo, C. Song, and Z. Zhong, "Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup," in *ECCV*, 2020.
- [51] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," *CVPR*, 2018.
- [52] S. Lin, H. Li, C.-T. Li, and A. C. Kot, "Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification," in *BMVC*, 2018.
- [53] H.-X. Yu, W. Zheng, A. Wu, X. Guo, S. Gong, and J. Lai, "Unsupervised person re-identification by soft multilabel learning," in *CVPR*, 2019.
- [54] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *ICCV*, 2019.
- [55] F. Yang, K. Li, Z. Zhong, Z. Luo, X. Sun, H. Cheng, X. Guo, F. Huang, R. Ji, and S. Li, "Asymmetric co-teaching for unsupervised cross-domain person re-identification," in *AAAI*, 2020.
- [56] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Learning to adapt invariance in memory for person re-identification," *IEEE TPAMI*, 2020.
- [57] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *AAAI*, 2019.
- [58] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, "Unsupervised person re-identification via softened similarity learning," in *CVPR*, 2020.
- [59] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," in *CVPR*, 2020.
- [60] J. Li and S. Zhang, "Joint visual and temporal consistency for unsupervised domain adaptive person re-identification," in *ECCV*, 2020.
- [61] G. Wu, X. Zhu, and S. Gong, "Tracklet self-supervised learning for unsupervised person re-identification," in *AAAI*, 2020.
- [62] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017.
- [63] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD*, 1996.
- [64] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," *ICCV*, 2015.
- [65] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCVW*, 2016.
- [66] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *ECCV*, 2016.
- [67] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," *CVPR*, 2018.
- [68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [69] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019.
- [70] F. Yang, Z. Zhong, Z. Luo, Y. Cai, S. Li, and S. Nicu, "Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification," in *CVPR*, 2021.
- [71] A. Strehl and J. Ghosh, "Cluster ensembles — a knowledge reuse framework for combining multiple partitions," *JMLR*, 2002.
- [72] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *ECCV*, 2018.
- [73] Y. Chen, X. Zhu, and S. Gong, "Deep association learning for unsupervised video person re-identification," in *BMVC*, 2018.
- [74] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in *ECCV*, 2018.
- [75] —, "Unsupervised tracklet person re-identification," *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [76] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," *CVPR*, 2017.
- [77] J. Y. Zhang, S. PePOSE, H. Joo, D. Ramanan, J. Malik, and A. Kanazawa, "Perceiving 3d human-object spatial arrangements from a single image in the wild," in *ECCV*, 2020.



**Hao Chen** received the B.S. degree from Wuhan University in 2014, and the M.S. degree from CentraleSupélec and Université Paris Saclay in 2017. He is currently working towards his Ph.D. at Inria Sophia Antipolis and Université Côte d'Azur. His research interests include person re-identification and unsupervised learning. Homepage: <https://chenhao2345.github.io/>.



**Yaohui Wang** received the B.S. degree from Xidian University in 2015, and the M.S. degree from ENSIE and Université Paris Saclay in 2017. He is currently working towards his Ph.D. at Inria Sophia Antipolis, STARS Team and Université Côte d'Azur. His current research focuses on image and video synthesis, activity recognition and representation learning.



**Benoit Lagadec** is a Research Engineer at European Systems Integration. He currently works on developing video analysis solutions based on abnormal human behavior. Previously, he worked in public research at Ifremer, where he was able to develop image processing algorithms adapted to the difficulty of underwater imaging : denoising, segmentation.



**Antitza Dantcheva** is a Research Scientist (CRCN) with the STARS team of INRIA Sophia Antipolis, France. Previously, she was a Marie Curie fellow at Inria and a Postdoctoral Fellow at the Michigan State University and the West Virginia University, USA. She received her Ph.D. degree from Télécom ParisTech/Eurecom in image processing and biometrics in 2011. Her research is in computer vision and specifically in designing algorithms that seek to learn suitable representations of the human face in interpreta-

tion and generation.



**Francois Bremond** received the PhD degree from INRIA in video understanding in 1997, and he pursued his research work as a post doctorate at the University of Southern California (USC) on the interpretation of videos taken from Unmanned Airborne Vehicle (UAV). In 2007, he received the HDR degree (Habilitation a Diriger des Recherches) from Nice University on Scene Understanding. He created the STARS team on the 1st of January 2012. He is the research director at INRIA Sophia Antipolis, France. He

has conducted research work in video understanding since 1993 at Sophia- Antipolis. He is author or co-author of more than 140 scientific papers published in international journals or conferences in video understanding. He is a handling editor for MVA and a reviewer for several international journals (CVIU, IJPRAI, IJHCS, PAMI, AIJ, Eurasip, JASP) and conferences (CVPR, ICCV, AVSS, VS, ICVS). He has (co-)supervised 26 PhD theses. He is an EC INFOS and French ANR Expert for reviewing projects.