Adversarial Feature Disentanglement for Long-Term Person Re-identification

Wanlu Xu[1], **Hong Liu**[1*], **Wei Shi**[1], **Ziling Miao**[1], **Zhisheng Lu**[1] and **Feihu Chen**[2]

[1]Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, China
[2]Department of Precision Instrument, Tsinghua University, China
{xuwanlu, hongliu, pkusw, zilingmiao, zhisheng_lu}@pku.edu.cn, cfh19@mails.tsinghua.edu.cn

Abstract

Most existing person re-identification methods are effective in short-term scenarios because of their appearance dependencies. However, these methods may fail in long-term scenarios where people might change their clothes. To this end, we propose an adversarial feature disentanglement network (AFD-Net) which contains intra-class reconstruction and inter-class adversary to disentangle the identity-related and identity-unrelated (clothing) features. For intra-class reconstruction, the person images with the same identity are represented and disentangled into identity and clothing features by two separate encoders, and further reconstructed into original images to reduce intra-class feature variations. For inter-class adversary, the disentangled features across different identities are exchanged and recombined to generate adversarial clothes-changing images for training, which makes the identity and clothing features more independent. Especially, to supervise these new generated clothes-changing images, a re-feeding strategy is designed to re-disentangle and reconstruct these new images for image-level self-supervision in the original image space and feature-level soft-supervision in the disentangled feature space. Moreover, we collect a challenging Market-Clothes dataset and a real-world PKU-Market-Reid dataset for evaluation. The results on one large-scale short-term dataset (Market-1501) and five long-term datasets (three public and two we proposed) confirm the superiority of our method against other state-of-the-art methods.

1 Introduction

Person re-identification (Re-ID), the process of matching person images across different camera views, is widely used in person search [Shi *et al.*, 2020] and multi-object tracking [Ke *et al.*, 2019]. In recent years, many person Re-ID methods have made great success in short-term scenarios. These methods often rely heavily on the appearance of the people, which makes them cannot be easily applied to address the clothes-changing problem in long-term scenarios. From Fig. 1, we find that the clothing features that are effective in short-term task become ineffective in long-term task, while identity features are discriminative to tackle the clothes-changing problem. Based on this observation, we aim to find how to split the identity and clothing information of person images and make them both independent and representative, so that our method can be adapted to both short and long-term scenarios.

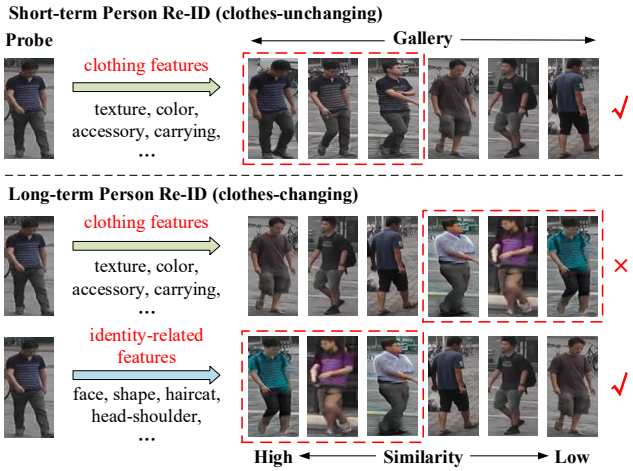

Figure 1: Examples of probe image and gallery candidates in the short-term and long-term scenarios from our generated dataset. The red dashed boxes represent the correct matching samples. Note that the identity-related features (face, shape, haircut, head-shoulder features, etc.) are more effective than clothing features (texture, color, accessory, carrying, etc.) to deal with clothes-changing problem.

Recently, great efforts have been devoted to extracting the robust identity representations of the human body for long-term person Re-ID task. Some works [Yang *et al.*, 2020; Wan *et al.*, 2020] proposed to use contour sketch or facial information as identity-related features for person re-identification. These methods only extract a certain type of explicit representation, but cannot get the complete identity-related information of the human body. Other works [Li *et al.*, 2020] proposed to use generated adversarial learning strategy to learn implicit shape-based features. These methods only pay attention to the adversarial learning between intra-class samples but ignore the important inter-class information.

*Contact Author

Considering the above problems, we propose an end-to-end adversarial feature disentanglement network (AFD-Net) to disentangle identity-related and identity-unrelated (clothing) features. The AFD-Net combines intra-class reconstruction and inter-class adversary to make the two disentangled features independent and representative to adapt to different scenarios. At the same time, the inter-class network can recombine the identity and clothing information across different identities to obtain new adversarial clothes-changing images. Since there is no ground truth for generated images, a re-feeding strategy is designed to supervise the adversarial image generation. To be specific, the generated images are fed back to the AFD-Net to re-disentangle and reconstruct for image-level self-supervision in the original image space and feature-level soft-supervision in the disentangled feature space. Moreover, we collect two long-term datasets for evaluation. The Market-Clothes dataset is a generated challenging dataset in which each identity has 2 to 71 sets of clothes. The PKU-Market-Reid dataset is a collected real-world dataset, which has complex background and occlusion in a supermarket. It is worth mentioning that our method does not need to be trained on the clothes-changing dataset like many long-term methods. It can be used for long or short-term tasks only by training on short-term datasets such as Market-1501.

The main contributions can be summarized as follows:

- We propose an adversarial feature disentanglement network (AFD-Net) which contains intra-class reconstruction and inter-class adversary to disentangle the identity-related and identity-unrelated (clothing) features. The inter-class network can recombine two features across identities to obtain adversarial clothes-changing images.
- To supervise the adversarial image generation, a re-feeding strategy is designed to re-disentangle and reconstruct new generated images for image-level self-supervision in the original image space and feature-level soft-supervision in the disentangled feature space.
- Two challenging datasets (Market-Clothes and PKU-Market-Reid) are collected for evaluation. The performance on one short-term dataset (Market-1501) and five long-term datasets (PRCC, Celeb-reID, Celeb-reID-light, Market-Clothes, PKU-Market-Reid) confirm the effectiveness and generalizability of our method.

2 Related Work

Short-term person Re-ID. Short-term person Re-ID task is carried out based on the assumption that people do not change clothes. Most short-term methods focus on tackling the challenges of matching images under posture, viewpoint, background, illumination, or occlusion variations. We classify these methods into six categories: representation learning based methods [Geng *et al.*, 2016], metric learning based methods [Varior *et al.*, 2016], global features based methods [Zheng *et al.*, 2017], local features based methods [Sun *et al.*, 2018], video-based methods [Liu *et al.*, 2018], and GAN-based methods [Zheng *et al.*, 2019]. These methods can achieve high performance on many short-term datasets, but due to their appearance dependencies, they cannot be easily applied to address the clothes-changing problem.

Dataset	Subjects	Images	Cameras	Clothes	Type
Celeb-reID-light [Huang *et al.*, 2020]	590	10,842	-	-	real
Celeb-reID [Huang *et al.*, 2020]	1,052	34,186	-	-	real
PRCC [Yang *et al.*, 2020]	221	33,698	3	-	real
LTCC [Qian *et al.*, 2020]	152	17,138	12	2-14	real
COCAS [Yu *et al.*, 2020]	5,266	62,382	30	2-3	real
Real28 [Wan *et al.*, 2020]	28	4,324	4	3	real
PKU-Market-Reid (Ours)	**25**	**6,028**	**4**	**3-10**	**real**
VC-Clothes [Wan *et al.*, 2020]	512	19,060	4	1-3	rendered
Div-Market [Li *et al.*, 2020]	200	24,732	6	-	generated
Market-Clothes (Ours)	**751**	**16,483**	**6**	**2-71**	**generated**

Table 1: Comparison of various long-term person Re-ID datasets.

Long-term person Re-ID. People may change their clothes in long-term person Re-ID task, which makes long-term task face a huge challenge of appearance variations. Most of the current long-term person Re-ID methods aim at finding human representations that are independent of appearance information. These representations include global features fusing facial information [Wan *et al.*, 2020], clothing features fusing biometric information [Yu *et al.*, 2020], local features combining with attention mechanisms [Huang *et al.*, 2020], and adversarial learning features [Li *et al.*, 2020]. The first three representations belong to explicit feature learning, and the last one belongs to implicit feature learning. We believe that explicit learning can obtain exact body representations, while implicit learning is more likely to obtain discriminative features autonomously. Since the identity information in the long-term person Re-ID is critical and ambiguous, we aim to use the implicit learning method. [Zheng *et al.*, 2019] proposed a joint learning method to couple Re-ID learning and data generation end-to-end. They think that the appearance information indicates the identity of the person, which is effective for short-term task. However, in our long-term task, it is more reasonable to regard identity as information independent of appearance. [Li *et al.*, 2020] proposed a novel representation learning method that can generate a shape-based feature representation that is invariant to clothing. They only pay attention to the adversarial learning between intra-class samples but ignore the inter-class case. In our method, we simultaneously obtain intra-class and inter-class information to enhance the independence of identity and clothing features.

Long-term datasets. Due to the difficulty of collection, there are fewer long-term datasets. Huang *et al*. [Huang *et al.*, 2020] built two large-scale long-term datasets called "Celeb-reID-light" and "Celeb-reID", which were acquired from the Internet using street snap-shots of celebrities. Yu *et al*. [Yu *et al.*, 2020] constructed a novel large-scale Re-ID benchmark named clothes changing person set (COCAS), which combined a new Re-ID setting with clothes template. Yang *et al*. [Yang *et al.*, 2020] and Qian *et al*. [Qian *et al.*, 2020] proposed two real-world long-term datasets PRCC and LTCC respectively. Wan *et al*. [Wan *et al.*, 2020] and Li *et al*. [Li *et al.*, 2020] used generation or rendering methods to get two "fake" clothes-changing datasets. We choose the three largest datasets mentioned above to evaluate our method. And taking into account the high efficiency of the generation method and the practicability of the real scene, we also collect two long-term datasets: Market-Clothes and PKU-Market-Reid. The comparison of various long-term datasets is shown in Table 1.

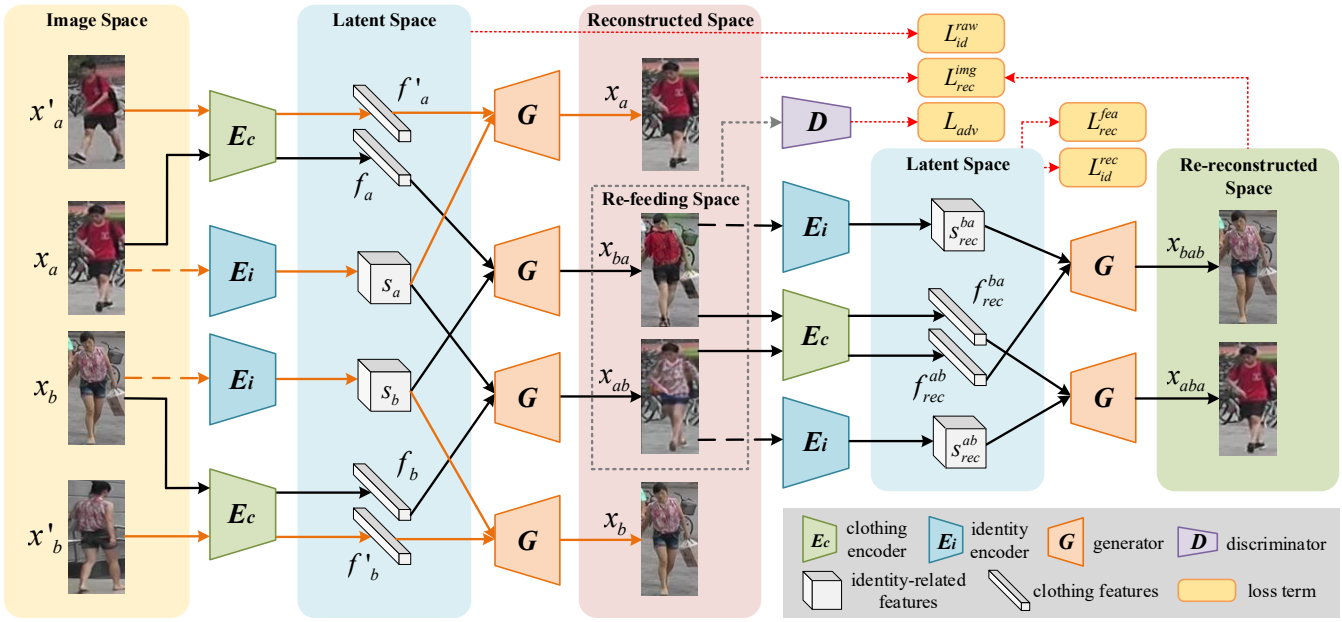

Figure 2: Overview of the proposed adversarial feature disentanglement network (AFD-Net). In disentanglement part, take x_a as the target: (a) for intra-class, the identity and clothing encoders E_i, E_c separately encode input images x_a, x'_a to produce identity-related and clothing features s_a, f'_a. The generator G jointly takes the s_a and f'_a to reconstruct the image x_a. (b) For inter-class, s_a and f_b are obtained through the same encoders E_i, E_c from x_a, x_b. The generator G jointly takes the s_a and f_b to generate a new clothes-changing image x_{ab}. The discriminator D is developed to determine whether the x_{ab} and x_a are from the same distribution. All operations are the same for x_b. In re-feeding part: (c) the generated images x_{ab}, x_{ba} are fed back to E_i, E_c to reconstruct the original images x_a, x_b for supervision of adversarial image generation. For better viewing, the dashed inputs of E_i denote the gray images, the orange lines represent the intra-class operations.

3 Proposed Method

In this section, the proposed adversarial feature disentanglement network (AFD-Net) is illustrated in Fig. 2. The inputs of the AFD-Net are four images including two pairs from two identities. The intra-class image pair is disentangled and reconstructed into original images to reduce intra-class feature variations. The inter-class image pair is disentangled and recombined to generate new adversarial clothes-changing images. Moreover, a re-feeding strategy is designed to supervise the adversarial image generation by re-disentangling and reconstructing these new images, which contains image-level self-supervision in the original image space and feature-level soft-supervision in the disentangled feature space.

3.1 Intra-Class Reconstruction

The goal of this part is to encourage the identity and clothing encoders to pull their codes of the same identity together so that intra-class feature variations are reduced. Specifically, the encoders E_i, E_c are used to separately extract identity and clothing features s, f of two different images x, x' from the same person. Then the generator G combines these two features to reconstruct the original image x that provides identity features before. The image-level reconstruction loss L_{recon}^{img} for x_a, x_b is defined as:

$$L_{recon}^{img} = ||x_a - G(s_a, f'_a)||_1 + ||x_b - G(s_b, f'_b)||_1, \quad (1)$$

where $||\cdot||_1$ denotes the L1 norm. L1 norm is widely adopted in reconstruction loss since it can preserve image sharpness.

In the meantime, to utilize labeled information of training data for person re-ID, we employ classification loss on the identity and clothing features of person images. That can force these two features of different identities to stay apart. The classification loss for identity and clothing features are:

$$L_{id}^{s} = L_{id}^{s_a} + L_{id}^{s_b} = -log(p(\hat{y}_a|x_{s_a})) - log(p(\hat{y}_b|x_{s_b})), \quad (2)$$

$$L_{id}^{f'} = L_{id}^{f'_a} + L_{id}^{f'_b} = -log(p(\hat{y}_a|x_{f'_a})) - log(p(\hat{y}_b|x_{f'_b})), \quad (3)$$

where $p(\hat{y}_a|x_{s_a})$ and $p(\hat{y}_a|x_{f'_a})$ are the predicted probabilities that x_a and x'_a belong to the ground-truth class $\hat{y}_a$ separately based on their identity and clothing codes. The probabilities $p(\hat{y}_b|x_{s_b})$ and $p(\hat{y}_b|x_{f'_b})$ have the same definition.

3.2 Inter-Class Adversary

After the intra-class feature variations are reduced, the inter-class adversary is carried out across different identities, which can force the identity and clothing encoders to learn their respective relevant information. To be more precise, first, two encoders E_i, E_c are used to separately extract identity and clothing features of two different identities. Second, the identity and clothing features are recombined to generate new adversarial clothes-changing images, like $x_{ab} = G(s_a, f_b)$ and $x_{ba} = G(s_b, f_a)$. Since there are no clothes-changing images in the same pose as the original images in the dataset, we cannot get the ground truth of the generated images. It means that the adversarial image generation has no pixel-level ground-truth supervision, so we introduce a re-feeding strategy for

supervision later. To make features representational as intra-class, the classification loss is also employed on clothing features of x_a and x_b:

$$L_{id}^{f} = L_{id}^{f_a} + L_{id}^{f_b} = -log(p(\hat{y}_a|x_{f_a})) - log(p(\hat{y}_b|x_{f_b})), \quad (4)$$

where $p(\hat{y}_a|x_{f_a})$, $p(\hat{y}_b|x_{f_b})$ are the predicted probabilities that x_a, x_b belong to the class $\hat{y}_a$, $\hat{y}_b$ based on clothing code.

The total classification loss of raw features L_{id}^{raw} is:

$$L_{id}^{raw} = L_{id}^{s} + L_{id}^{f} + L_{id}^{f'}. \quad (5)$$

To further enforce G to perform content recovery, we produce perceptually realistic outputs by using the discriminator D to discriminate the real images x_a, x_b with the generated images x_{ab}, x_{ba}. The adversarial loss L_{adv}^{ab}, L_{adv}^{ba} are:

$$L_{adv}^{ab} = log(D(x_a)) + log(1 - D(x_{ab})), \quad (6)$$

$$L_{adv}^{ba} = log(D(x_b)) + log(1 - D(x_{ba})). \quad (7)$$

The total adversarial loss L_{adv} is:

$$L_{adv} = L_{adv}^{ab} + L_{adv}^{ba}. \quad (8)$$

3.3 Image Re-feeding

Since there is no ground truth for clothes-changing images, a re-feeding strategy is designed to supervise the adversarial image generation. The re-feeding strategy contains image-level self-supervision in the original image space and feature-level soft-supervision in the disentangled feature space. For image-level self-supervision, the generated images are fed back to AFD-Net to re-disentangle and reconstruct into the original images. Like x_{ab}, x_{ba} corresponding to x_a, x_b, we can get x_{aba}, x_{bab} from x_{ab}, x_{ba}, which is formulated as:

$$x_{aba} = G(s_{rec}^{ab}, f_{rec}^{ba}), \quad (9)$$

$$x_{bab} = G(s_{rec}^{ba}, f_{rec}^{ab}), \quad (10)$$

where the features s_{rec}^{ab}, f_{rec}^{ba} derive from x_a, and the features s_{rec}^{ba}, f_{rec}^{ab} derive from x_b. We find that the above-mentioned reconstructed images x_{aba}, x_{bab} are exactly the original images x_a, x_b. Thus, we can achieve self-supervision by supervising the original images. The image-level reconstruction loss $L_{re-recon}^{img}$ in re-feeding part can be expressed as:

$$L_{re-recon}^{img} = ||x_a - x_{aba}||_1 + ||x_b - x_{bab}||_1. \quad (11)$$

We add $L_{re-recon}^{img}$ and L_{recon}^{img} calculated before to obtain the total image-level reconstruction loss L_{rec}^{img}:

$$L_{rec}^{img} = L_{recon}^{img} + L_{re-recon}^{img}. \quad (12)$$

For feature-level soft-supervision, each generated image is softly supervised by two corresponding original images in the disentangled feature space. For example, the identity and clothing features of x_{ab} are supervised by x_a and x_b respectively. The feature-level reconstruction loss L_{rec}^{s}, L_{rec}^{f} are:

$$L_{rec}^{s} = ||s_a - s_{rec}^{ab}||_1 + ||s_b - s_{rec}^{ba}||_1, \quad (13)$$

$$L_{rec}^{f} = ||f_b - f_{rec}^{ab}||_1 + ||f_a - f_{rec}^{ba}||_1, \quad (14)$$

The total feature-level reconstruction loss L_{rec}^{fea} is:

$$L_{rec}^{fea} = L_{rec}^{s} + L_{rec}^{f}. \quad (15)$$

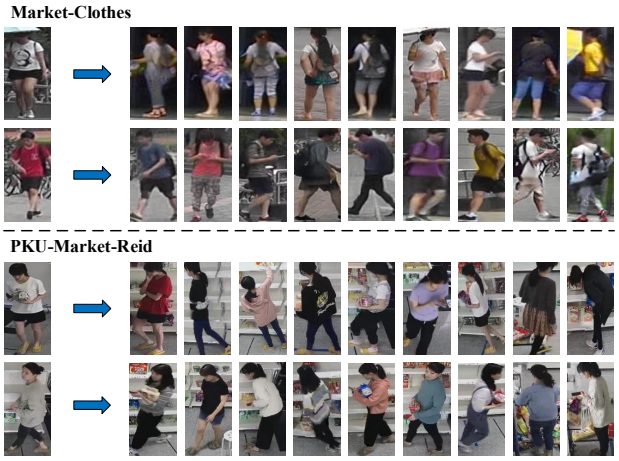

Figure 3: Examples of generated Market-Clothes and collected PKU-Market-Reid datasets. Each identity is shown in each row.

The classification loss is also employed on s_{rec} and f_{rec}. Since the two reconstructed images x_{ab} and x_{ba} are generated from two images with different ID, how to determine their labels is an important problem. We introduce a soft label strategy to tackle this problem, and prove its effectiveness in ablation experiments. For example, because the image x_{ab} is generated by features s_a and f_b, we set the ground truth of s_{rec}^{ab} as y_a, of f_{rec}^{ab} as y_b to be consistent with s_a and f_b. This can further force the two encoders to output mutually independent features. The classification loss L_{id}^{res}, L_{id}^{ref} are:

$$L_{id}^{res} = L_{id}^{s_{ab}} + L_{id}^{s_{ba}} = -log(p(\hat{y}_a|x_{s_{ab}})) - log(p(\hat{y}_b|x_{s_{ba}})), \quad (16)$$

$$L_{id}^{ref} = L_{id}^{f_{ab}} + L_{id}^{f_{ba}} = -log(p(\hat{y}_b|x_{f_{ab}})) - log(p(\hat{y}_a|x_{f_{ba}})), \quad (17)$$

where $p(\hat{y}_a|x_{s_{ab}})$ and $p(\hat{y}_b|x_{f_{ab}})$ are the predicted probabilities that x_{ab} belongs to the ground-truth class $\hat{y}_a$ and $\hat{y}_b$ separately based on its identity and clothing codes. The probabilities $p(\hat{y}_b|x_{s_{ba}})$ and $p(\hat{y}_a|x_{f_{ba}})$ have the same definition. The total classification loss of reconstructed features L_{id}^{rec} is:

$$L_{id}^{rec} = L_{id}^{res} + L_{id}^{ref}. \quad (18)$$

In summary, through the image reconstruction of x_{aba} and x_{bab} in image-level self-supervision, the feature reconstruction and classification of s_{rec} and f_{rec} in feature-level soft-supervision, the adversarial image generation is supervised and new reliable clothes-changing images are generated.

3.4 Optimization

We train the AFD-Net end-to-end, and the sum of the above-mentioned losses in the entire network is shown as:

$$\begin{aligned} L_{total} = {} & \lambda_{rec}^{img} \cdot L_{rec}^{img} + \lambda_{rec}^{fea} \cdot L_{rec}^{fea} + \lambda_{adv} \cdot L_{adv} \\ & + \lambda_{id}^{raw} \cdot L_{id}^{raw} + \lambda_{id}^{rec} \cdot L_{id}^{rec} \end{aligned} \quad (19)$$

Following [Huang *et al.*, 2018], we use a large weight $\lambda_{rec}^{img} = 5$ for the image reconstruction loss. The feature reconstruction loss is inaccurate in the initial stage, so we gradually increase λ_{rec}^{fea} from 0 to 1 during the training process. Similarly, the identification loss L_{id}^{rec} may make the training unstable due to the low quality of adversarial generated images at the beginning, so we set a small weight $\lambda_{id}^{rec} = 0.5$. The remaining weights λ_{id}^{raw} and λ_{adv} are set to 1.

4 Experiments

4.1 Datasets

We evaluate our method on six person Re-ID datasets, including three large-scale long-term datasets: Celeb-reID, Celeb-reID-light, PRCC, one benchmark short-term dataset: Market-1501, and two collected long-term datasets: Market-Clothes, PKU-Market-Reid which are shown in Fig. 3.

Celeb-reID / Celeb-reID-light. Celeb-reID [Huang *et al.*, 2020] is composed of 34,186 images of 1,052 identities that crawled from websites. It has 20,208 images from 632 identities for training and 13,978 images from 420 identities for testing. Celeb-reID-light is a light version of Celeb-reID, which has 490 and 100 identities for training and testing.

PRCC. PRCC [Yang *et al.*, 2020] consists of 33,698 images from 221 identities. Each person in Cameras A and B is wearing the same clothes, and in A and C is wearing different clothes. The dataset has 150 and 71 identities for training and testing, and also provides contour sketch images for use.

Market-1501. Market-1501 [Zheng *et al.*, 2015] is composed of 32,668 images of 1,501 identities collected from 6 camera views. The dataset is split into two non-overlapping fixed parts: 12,936 images from 751 identities for training and 19,732 images from 750 identities for testing.

Market-Clothes. Market-Clothes is our generated long-term dataset which is shown in Fig. 3. We change the clothes of person in Market-1501. It consists of 16,483 images from 751 identities, and each identity has 2 to 71 sets of clothes. It is a challenging dataset only used for the test phase.

PKU-Market-Reid. PKU-Market-Reid is our collected dataset which is shown in Fig. 3. It contains 6,028 images of 25 identities collected from 4 camera views, and each identity has 3 to 10 sets of clothes. This dataset contains different clothes, pose, illumination, shielding, and other complex conditions in supermarket. It is only used for the test phase too.

4.2 Implementation Details

We implement the AFD-Net using PyTorch with only one NVIDIA GTX 1080Ti GPU. All images are resized to 256×128 for input. The SGD optimizer is used to train E_i, E_c with $lr = 0.002$, $momentum = 0.9$. Encoders of the same type in AFD-Net share parameters. The Adam optimizer is applied to optimize G, D with $lr = 0.0001$, $(\beta 1, \beta 2) = (0, 0.999)$. We use ResNet50 [He *et al.*, 2016] pre-trained on ImageNet as our backbone of E_i and E_c, and convert the input of E_i to grayscale for color filtering. In the test phase, E_i outputs the identity code in $batch \times 2048$, E_c outputs the clothing code in $batch \times 1024$. We use the concatenation of two features for long-term task, and clothing features for short-term task.

4.3 Comparison with State-of-the-Arts

In this part, the common methods and long-term methods are split into two big rows, and the symbol R is short for $Rank$.

Celeb-reID / Celeb-reID-light. Experimental results on two high-resolution long-term datasets are shown in Table 2. Our AFD-Net outperforms all the compared methods on Rank-1 and Rank-5 but underperforms the mAP of method MGN. The reason is that MGN utilizes multi-granularity features for Re-ID, which not only increases the mAP but also increases the complexity. However, our method can achieve comparable mAP by only using two disentangled features.

Method	Celeb-reID			Celeb-reID-light		
	R1	R5	mAP	R1	R5	mAP
MLFN [Chang *et al.*, 2018]	41.4	54.7	6.0	10.6	31.0	6.3
IDE+DesNet121 [Zheng *et al.*, 2017]	42.9	56.4	5.9	10.5	24.8	5.3
ResNet-Mid [Yu *et al.*, 2017]	43.3	54.6	5.8	10.3	28.0	6.0
HACNN [Li *et al.*, 2018]	47.6	63.3	9.5	16.2	42.8	11.5
MGN [Wang *et al.*, 2018]	49.0	64.9	**10.8**	21.5	47.4	**13.9**
ReIDCaps [Huang *et al.*, 2020]	51.2	65.4	9.8	20.3	48.2	11.2
AFD-Net (Ours)	**52.1**	**66.1**	10.6	**22.2**	**51.0**	11.3

Table 2: Comparison on Cele-reID and Cele-reID-light datasets.

Method	Camera C to A (Change clothes)			Camera B to A (Same clothes)		
	R1	R10	R20	R1	R10	R20
Alexnet [Krizhevsky *et al.*, 2012]	16.3	48.0	65.9	63.3	91.7	94.7
VGG16 [Simonyan and Zisserman, 2015]	18.2	46.1	60.8	71.4	95.9	98.7
Res-Net50 [He *et al.*, 2016]	19.4	52.4	66.4	74.8	97.3	98.9
HACNN [Li *et al.*, 2018]	21.8	59.5	67.5	82.5	98.1	99.0
PCB [Sun *et al.*, 2018]	22.9	61.2	78.3	86.9	98.8	99.6
Deformable Conv [Dai *et al.*, 2017]	26.0	71.7	85.3	61.9	92.1	97.7
STN [Jaderberg *et al.*, 2015]	27.5	69.5	83.2	59.2	91.4	96.1
SPT+ASE [Yang *et al.*, 2020]	34.4	77.3	88.1	64.2	92.6	96.7
AFD-Net (Ours)	**42.8**	**79.6**	**89.0**	**95.7**	**99.0**	**99.8**

Table 3: Comparison on PRCC dataset.

Method	R1	R5	R10	mAP
MLFN [Chang *et al.*, 2018]	90.0	-	-	74.3
FD-GAN [Ge *et al.*, 2018]	90.5	96.0	97.7	77.9
HACNN [Li *et al.*, 2018]	91.2	-	-	75.7
PCB [Sun *et al.*, 2018]	93.2	97.3	98.2	81.7
DG-Net [Zheng *et al.*, 2019]	94.4	98.4	98.9	85.2
MGN [Wang *et al.*, 2018]	**95.7**	-	-	**86.9**
ReIDCaps [Huang *et al.*, 2020]	89.0	-	-	72.7
ReIDCaps+ [Huang *et al.*, 2020]	92.8	-	-	78.0
AFD-Net (Ours)	94.2	98.0	98.7	84.3

Table 4: Comparison on Market-1501 dataset.

Method	Market-Clothes				PKU-Market-Reid			
	R1	R5	R10	mAP	R1	R5	R10	mAP
FD-GAN [Ge *et al.*, 2018]	1.25	4.33	7.63	0.42	42.19	46.88	51.56	16.41
Part-aligned [Suh *et al.*, 2018]	1.31	3.56	6.53	0.39	39.84	44.53	46.88	15.84
PCB [Sun *et al.*, 2018]	2.82	8.14	12.90	0.71	45.66	48.78	55.03	16.65
P2Net [Guo *et al.*, 2019]	8.60	15.62	23.44	5.70	43.16	47.66	52.34	14.68
DG-Net [Zheng *et al.*, 2019]	20.04	33.09	39.64	6.32	44.53	49.78	53.91	17.30
ReIDCaps [Huang *et al.*, 2020]	15.52	22.59	27.64	6.60	43.91	46.69	50.81	15.66
AFD-Net (Ours)	**48.31**	**65.88**	**72.77**	**8.47**	**46.88**	**50.02**	**56.25**	**18.05**

Table 5: Comparison on Market-Clothes and PKU-Market-Reid.

Figure 4: Visualization of identity features on Market-Clothes via t-SNE. We visualize 30 different identities by using ReIDCaps (left) and our AFD-Net (right), each of which is shown in a unique color.

PRCC. As shown in Table 3, the Rank-1 / mAP of AFD-Net achieves 42.8% / 89.0% in clothes-changing task, and

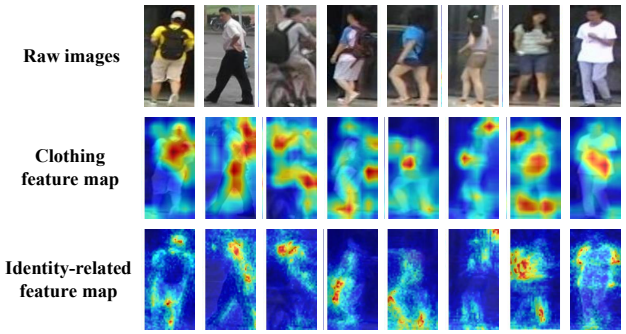

Figure 5: Visualization of clothing and identity feature maps on Market-1501 via AFD-Net. Each identity is shown in each column.

95.7% / 99.8% in clothes-unchanging task, which outperforms all the other methods by a large margin. This is because our approach can disentangle identity and clothing features, make them more independent to adapt to different scenarios.

Market-1501. To further prove that our method can also guarantee high performance in short-term task, we also evaluate our method on a large-scale short-term person Re-ID dataset: Market-1501. As shown in Table 4, our approach achieves comparable performance with short-term methods and far exceeds other long-term methods on Market-1501, which indicates the good generalization of our method.

Market-Clothes / PKU-Market-Reid. We use the model trained on the Market-1501 dataset to test on our two collected datasets. The experimental results are shown in the Table 5. On the challenging Market-Clothes dataset, other methods suffer significant performance declines, while ours remains stable and superior. On the real PKU-Market-Reid dataset, our method can still adapt to the complexity of real-world scenarios and achieve the best performance. These results can prove the domain adaptability of our method. The identity features clustering scatter plots are shown in Fig. 4.

4.4 Ablation Study

To investigate the effect of each component in our AFD-Net, we perform ablation studies on the Market-Clothes dataset.

Identity vs. clothing features. The best performance appears when identity and clothing features are concatenated as shown in Table 6. To further show the independence of the two features, the feature visualization is performed on the Market-1501 dataset in Fig. 5. It shows that our AFD-Net can effectively disentangle the local clothing information and the global context-aware identity information of human body.

Intra-class vs. inter-class networks. We also compare the intra-class reconstruction and inter-class adversary networks as shown in Table 7. It shows that the "intra+inter" surpasses "intra" and "inter" by a large margin, which indicates that the inter-class information is more important than the intra-class, and training the two networks end-to-end is effective.

Effect of re-feeding strategy. To evaluate the impact of the re-feeding strategy, we conduct an ablation study as shown in Table 8. Compared to "Non-Refeed", "Refeed" surpasses it by +7.34% Rank-1 accuracy and +2.51% mAP. It proves that the re-feeding part can effectively supervise the adversarial image generation, and further disentangle the identity and clothing features to make them more independent.

Features	s	f	R1	R5	R10	mAP
Identity	✓		33.52	47.51	54.33	4.45
Clothing		✓	38.24	55.34	62.74	6.32
Identity+Clothing	✓	✓	**48.31**	**65.88**	**72.77**	**8.47**

Table 6: Ablation studies on different features.

Type	intra-class	inter-class	R1	R5	R10	mAP
Intra	✓		39.58	54.51	61.85	5.75
Inter		✓	42.25	58.31	65.86	6.75
Intra+Inter	✓	✓	**48.31**	**65.88**	**72.77**	**8.47**

Table 7: Ablation studies on different networks.

Method	Re-feeding	R1	R5	R10	mAP
Non-Refeed	×	40.97	55.43	62.35	5.96
Refeed	✓	**48.31**	**65.88**	**72.77**	**8.47**

Table 8: Effectiveness evaluation of the re-feeding strategy.

Label	y_a	y_b	R1	R5	R10	mAP
L_aL_a	s, f	-	40.83	57.04	65.29	6.60
L_bL_b	-	s, f	40.35	57.51	65.05	6.61
L_bL_a	f	s	38.72	53.15	60.15	5.57
L_aL_b	s	f	**48.31**	**65.88**	**72.77**	**8.47**

Table 9: Effectiveness evaluation of the soft label strategy. As x_{ab} for example, we show the different label selections for s^{ab}_{rec} and f^{ab}_{rec}.

Effect of soft label strategy. We investigate how to set the supervised labels for disentangled features of generated images. As x_{ab} for example, the comparisons are shown in Table 9. "L_aL_a" and "L_bL_b" denote that both features are set to the same label, whereas "L_bL_a" and "L_aL_b" denote different labels. Our soft label strategy "L_aL_b" gets the excellent performance. This is because s^{ab}_{rec} and f^{ab}_{rec} are derive from x_a and x_b, so setting their labels to y_a and y_b is effective.

5 Conclusion

In this paper, we propose an adversarial feature disentanglement network (AFD-Net), intending to address the clothes-changing problem in long-term person Re-ID. First, the AFD-Net combines the intra-class reconstruction and inter-class adversary to disentangle the identity and clothing features, which can make these features more independent and discriminative. Second, AFD-Net can recombine the identity and clothing features across different identities to generate new reliable clothes-changing images for further disentangling. Third, a re-feeding strategy can supervise the adversarial image generation by image-level self-supervision in the original image space and feature-level soft-supervision in the disentangled feature space. Finally, to facilitate the research about the clothes-changing problem, we collect two long-term Re-ID datasets: Market-Clothes and PKU-Market-Reid.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No.62073004), Science and Technology Plan of Shenzhen (No.JCYJ20190808182209321).

References

[Chang *et al.*, 2018] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *ICCV*, pages 2109–2118, 2018.

[Dai *et al.*, 2017] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017.

[Ge *et al.*, 2018] 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*, pages 1230–1241, 2018.

[Geng *et al.*, 2016] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. Deep transfer learning for person re-identification. *CoRR*, abs/1611.05244, 2016.

[Guo *et al.*, 2019] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *ICCV*, pages 3641–3650, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Huang *et al.*, 2018] Xun Huang, MingYu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 179–196, 2018.

[Huang *et al.*, 2020] Yan Huang, Jingsong Xu, Qiang Wu, Yi Zhong, Peng Zhang, and Zhaoxiang Zhang. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *TCSVT*, 30(10):3459–3471, 2020.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015.

[Ke *et al.*, 2019] Bo Ke, Huicheng Zheng, Lvran Chen, Zhiwei Yan, and Ye Li. Multi-object tracking by joint detection and identification learning. *Neural Process. Lett.*, 50(1):283–296, 2019.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1106–1114, 2012.

[Li *et al.*, 2018] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *ICCV*, pages 2285–2294, 2018.

[Li *et al.*, 2020] Yu-Jhe Li, Zhengyi Luo, Xinshuo Weng, and Kris M. Kitani. Learning shape representations for clothing variations in person re-identification. *CoRR*, abs/2003.07340, 2020.

[Liu *et al.*, 2018] Hao Liu, Zequn Jie, Jayashree Karlekar, Meibin Qi, Jianguo Jiang, Shuicheng Yan, and Jiashi Feng. Video-based person re-identification with accumulative motion context. *TCSVT*, 28(10):2788–2802, 2018.

[Qian *et al.*, 2020] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. *CoRR*, abs/2005.12633, 2020.

[Shi *et al.*, 2020] Wei Shi, Hong Liu, and Mengyuan Liu. Identity-sensitive loss guided and instance feature boosted deep embedding for person search. *Neurocomputing*, 415:1–14, 2020.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[Suh *et al.*, 2018] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, pages 418–437, 2018.

[Sun *et al.*, 2018] 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*, pages 501–518, 2018.

[Varior *et al.*, 2016] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, pages 791–808, 2016.

[Wan *et al.*, 2020] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPR*, pages 830–831, 2020.

[Wang *et al.*, 2018] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACMM*, pages 274–282, 2018.

[Yang *et al.*, 2020] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *CoRR*, abs/2002.02295, 2020.

[Yu *et al.*, 2017] Qian Yu, Xiaobin Chang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. *CoRR*, abs/1711.08106, 2017.

[Yu *et al.*, 2020] Shijie Yu, Shihua Li, Dapeng Chen, Rui Zhao, Junjie Yan, and Yu Qiao. Cocas: A large-scale clothes changing person dataset for re-identification. In *CVPR*, pages 3400–3409, 2020.

[Zheng *et al.*, 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[Zheng *et al.*, 2017] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *CVPR*, pages 3346–3355, 2017.

[Zheng *et al.*, 2019] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 2138–2147, 2019.