Weakening the Influence of Clothing: Universal Clothing Attribute Disentanglement for Person Re-Identification

Yuming Yan[1,2], **Huimin Yu**[1,2,3,4*], **Shuzhao Li**[1,2], **Zhaohui Lu**[1,2,5], **Jianfeng He**[1]
Haozhuo Zhang[1,2] and **Runfa Wang**[1,2]
[1]College of Information Science and Electronic Engineering, Zhejiang University
[2]ZJU-League Research & Development Center
[3]State Key Lab of CAD&CG, Zhejiang University
[4]Zhejiang Provincial Key Laboratory of Information Processing, Communication and Networking
[5]Zhejiang Lijia Electronic Technology Company Ltd.
{12231016, yhm2005, 11631045, luzhaohui, 21931092, haozhuozhang, 21960189}@zju.edu.cn,

Abstract

Most existing Re-ID studies focus on the short-term cloth-consistent setting and thus dominate by the visual appearance of clothing. However, the same person would wear different clothes and different people would wear the same clothes in reality, which invalidates these methods. To tackle the challenge of clothes change, we propose a Universal Clothing Attribute Disentanglement network (UCAD) which can effectively weaken the influence of clothing (identity-unrelated) and force the model to learn identity-related features that are unrelated to the worn clothing. For further study of Re-ID in cloth-changing scenarios, we construct a large-scale dataset called CSCC with the following unique features: **(1) Severe**: A large number of people have cloth-changing over four seasons. **(2) High definition**: The resolution of the cameras ranges from 1920×1080 to 3840×2160, which ensures that the recorded people are clear. Furthermore, we provide two variants of CSCC considering different degrees of cloth-changing, namely moderate and severe, so that researchers can effectively evaluate their models from various aspects. Experiments on several cloth-changing datasets including our CSCC and short-term dataset Market-1501 prove the superiority of UCAD. The dataset is available at https://github.com/yomin-y/UCAD.

1 Introduction

Person Re-identification (Person Re-ID) aims at matching a specific person across non-overlapping cameras, which has large promising applications such as video surveillance and smart city. Extensive methods assuming a short-term cloth-consistent setting made great progress recently. However, these methods heavily rely on clothing features (texture, color, accessory, *etc.*) which makes them ineffective in cloth-changing scenarios. Fig.1 gives intuitive illustrations that the same person could look different on different clothes and different people could look similar on the same clothes, clothes become a disturbance to retrieval under cloth-changing setting. Therefore, we seek to alleviate the influence of clothing and learn identity-related features (face, shape, haircut, *etc.*) so that our method can apply to cloth-changing scenarios.

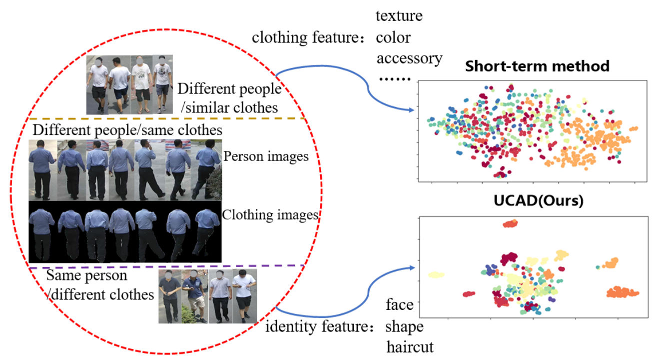

Figure 1: Different people wearing the same/similar clothes and the same person wearing different clothes would confuse short-term methods dramatically. Identity features (face, shape, haircut, *etc.*) are more robust than clothing features (texture, color, accessory, *etc.*) under the cloth-changing setting. It can be observed that the same clothes look different on different people which means the worn clothing may reveal some kind of information (*e.g.* body contour) of an identity. Feature distributions are visualized by t-SNE.

To tackle the challenge of clothes change, a few works have shifted their attention to the extraction of cloth-agnostic representations like shape [Qian *et al.*, 2020b], contour sketch [Yang *et al.*, 2021]. However, they have two main drawbacks: (1) Off-the-shelf networks from other tasks (*e.g.* the Human Keypoints Estimator) are employed both in the training and testing stage, which are of high time and resource cost thus limiting their scalability. (2) They paid too much attention to extracting certain kinds of features while failing to consider all identifiable identity-related features. Other works [Xu *et al.*, 2021; Li *et al.*, 2020b] proposed to use generated adversarial learning to disentangle the identity and clothing features. However, we notice that the same clothes would look different on different people as shown in Fig.1 which means the worn clothing may reveal some kind of information (*e.g.*

*Contact Author

Figure 2: Image samples of some identities in CSCC. Each row represents one identity. Winter, autumn, spring, and summer clothes are shown in each row from left to right.

body contour) of an identity. Therefore, the clothing features they extract contain part of the identity feature so they can not fully disentangle clothing features and identity features.

Considering the above problems, we propose a Universal Clothing Attribute Disentanglement (UCAD) network which can effectively weaken the influence of clothing and force the model to learn identity-related features that are unrelated to the worn clothing. Our proposed model includes two branches: the Clothing Attribute Extraction (CAE) branch and the Clothing Attribute Filtering (CAF) branch. First, with the help of the Human Parsing Estimator (HPE), the CAE branch is designed to obtain the clothing feature called the clothing attribute feature which truly contains no identity information (*e.g.* body contour). Next, the CAF branch distills the knowledge from the CAE branch aiming to decompose and disentangle clothing feature and identity feature to alleviate the influence of the worn clothing. Specifically, a consistent loss is used to distill from the clothing attribute feature to explicitly ensure the clothing feature contains no identity information. Besides, an orthogonal loss is utilized to encourage the clothing and identity features to be decoupled so that the identity feature could contain all identifiable information (shape, face, haircut, body contour *etc.*) except clothing thus weakening the influence of the worn clothing. The CAE branch is discarded at test stage to improve efficiency, which means the HPE is only used in the training stage.

To further study Person Re-ID in cloth-changing scenarios, large-scale datasets are acquired. Although some cloth-changing datasets have been collected, they are under moderate cloth-changing setting [Yang *et al.*, 2021] which means they do not include clothes change over four seasons and the maximum resolution of cameras they use is 1920×1080. However, changing clothes over four seasons is a common reality for human beings. In addition, it has been proven that higher resolution is good for Re-ID [Wang *et al.*, 2020]. Therefore, we contribute a large-scale cloth-changing dataset called CSCC with the following characteristics: **(1) Severe**: A large number of people have cloth-changing over all four seasons. **(2) High definition**: The resolution of the used 13 cameras ranges from 1920×1080 to 3840×2160, which ensures that the recorded people are clear. Examples are shown in Fig.2. Furthermore, we provide two variants of CSCC considering different degrees of cloth-changing, namely moderate and severe, for researchers to evaluate their models from various aspects.

The main contributions can be summarized as follows:

- A novel Universal Clothing Attribute Disentanglement network (UCAD) is proposed. It contains two branches. The Clothing Attribute Extraction branch aims to extract clothing feature which contains no identity information (*e.g.* body contour). In addition, the Clothing Attribute Filtering branch distills the knowledge from the CAE branch aiming to decompose and disentangle the clothing feature and the identity feature to alleviate the influence of the worn clothing.
- A severe cloth-changing person Re-ID dataset, called CSCC, is introduced to advance the study of more challenging yet realistic cloth-changing Person Re-ID, and a new evaluation criterion considering moderate and severe cloth-changing settings is proposed.
- Extensive experiments on several cloth-changing benchmarks including CSCC and short-term dataset Market-1501 demonstrate the effectiveness of our approach.

2 Related Works

2.1 Cloth-Changing Person Re-ID Datasets

A few datasets were collected specifically for cloth-changing setting. Celeb-reID [2019a] are acquired from the Internet which consists of street snapshots of celebrities. Yu *et al.* [2020] collected a cloth-changing Re-ID dataset called COCAS with cloth templetes. Qian *et al.* [2020b], Yang *et al.* [2021] and Wang *et al.* [2020] constructed three real-world cloth-changing Re-ID datasets LTCC, PRCC and NKUP respectively. Wan *et al.* [2020], Li *et al.* [2020b] and Xu *et al.* [2021] used generation or rendering methods to construct three virtual datasets VC-Clothes, Div-Market and Market-Clothes. However, all the existing datasets do not include clothes change over all the four seasons. Therefore, we construct a large-scale real-world cloth-changing dataset CSCC with a lot of people have cloth-changing over four seasons.

2.2 Cloth-Changing Person Re-ID Methods

The dramatic change in appearances when one person wears different clothes and the confusability when different people wear similar/same clothes are the main causes of failing short-term methods in cloth-changing scenarios. Some cloth-changing person Re-ID methods aim at finding cloth-agnostic representations. For example, Fan *et al.*[2020] proposed RF-ReID to use reflected wireless signals to identify persons. Yang *et al.*[2021] proposed a network to use human contour sketch information to identify persons. Chen *et al.*[2021] proposed to use 3D reconstruction to obtain the 3D shape feature to identify persons. We think certain kinds of feature is not discriminative enough and all identifiable features should be considered. Other works aim at separating clothing and identity features. Huang *et al.*[2019b] proposed a network using vector neurons instead of scalar neurons. The direction of the vector represents cloth-changing and the length of the vector stands for the identity. Qian *et al.*[2020b] proposed to distill shape information and disentangle cloth information using

Dataset	IDs	Imgs	Cams	Scene	Time span	Degree
Real28 [Wan *et al.*, 2020]	28	4324	4	ind,out	3 days	Moderate
LTCC [Qian *et al.*, 2020b]	152	17,138	12	ind	2 months	Moderate
PRCC [Yang *et al.*, 2021]	221	33,698	3	ind	-	Moderate
NKUP [Wang *et al.*, 2020]	107	9,738	15	ind,out	1 month	Moderate
CSCC(Ours)	**267**	**36,700**	13	out	a year	Severe

Table 1: Comparison with the existing real-world cloth-changing datasets. ('ind':indoor, 'out':outdoor, '-':unknown)

key points information. Yu *et al.*[2020] introduced a method using template clothes to achieve the purpose of retrieving a specific pedestrian wearing template clothes. Xu *et al.*[2021] proposed to use generated adversarial learning to disentangle the identity and clothing features. However, they ignore that the same clothes would look different on different people thus the clothing feature still contains some identity information (*e.g.* body contour). Therefore, in this paper, we aim to extract clothing feature which contains no identity information and learn the identity-related features to retrieve persons.

3 Severe Cloth-Changing Dataset

3.1 Data Collection and Labeling

A total of 13 cameras are used to record the raw videos. Five of the cameras have a resolution of 3840×2160, and the rest of them have a resolution of 1920×1080. Mask-RCNN [He *et al.*, 2017] detector is applied to obtain the person images. Each image is annotated with person ID, camera ID, cloth ID, and universal clothing attribute ID.

3.2 Statistics and Comparison

The proposed CSCC dataset has 36,700 person images of 267 identities. Each identity is captured by at least 2 cameras. The number of outfit changes ranges from 2 to 21 for each identity. In addition, a large number of identities (34 identities) have cloth-changing during all four seasons and almost all the identities have clothes variation in different seasons.

We randomly split the identities for training and testing. The training set contains 21,392 images of 134 identities and the rest 15,308 images of 133 identities are partitioned into the test set. For gallery and query splitting, we randomly picked one image per suit for each identity in each camera view to form the query set, and the remaining test set is left as the gallery set. The query set contains 1,614 images and the gallery set contains 13,694 images. Comparison with existing cloth-changing datasets is summarized in Tab.1.

4 Methodology

In this section, a novel Universal Clothing Attribute Disentanglement (UCAD) model is proposed to tackle the challenge of clothing changes. We aim to weaken the influence of the clothing and force the model to learn identity features that are unrelated to the worn clothing. The UCAD includes two branches: Clothing Attribute Extraction (CAE) Branch and Clothing Attribute Filtering (CAF) Branch. More details are listed below. The overall architecture is shown in Fig.3.

4.1 Clothing Attribute Extraction Branch

The goal of the CAE branch is to extract the clothing feature which contains no identity feature. Specifically, an off-the-shelf Human Parsing Estimation model is used to parse the clothing part from a person image to form a clothing image so that the extracted features are not affected by the background and head (clothing images are provided in our dataset). Next, we send the clothing image to the CAE branch which consists of a backbone network and an attention module to extract the clothing feature called the clothing attribute feature f^{CA}.

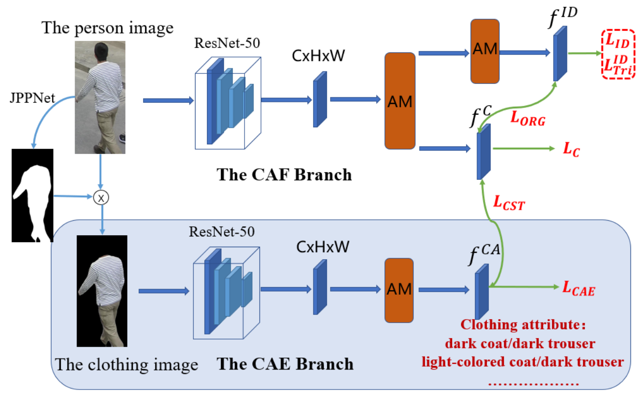

Figure 3: Overview of the proposed UCAD. It contains the CAE branch and the CAF branch. The former takes clothing images as input to extract the clothing attribute feature which contains no identity information. The latter distilling the knowledge from the former takes person images as input to decompose and disentangle the clothing feature and identity feature to alleviate the influence of the worn clothing. The CAE branch is discarded in the inference stage for computing efficiency. AM represents the attention module.

How to decouple with identity. We notice that the same clothes would look different on different people (as shown in Fig.1) which means the worn clothing may reveal some kind of information (*e.g.* body contour) of an identity. To extract clothing features that contain no identity information, we must find the generality of the same/similar clothing worn by different people. Therefore, we annotate each image with a universal clothing attribute label indicating the general attribute of clothing such as dark coat/dark trouser and light-colored coat/light-colored trouser *etc.*. In the training stage, a fully-connected classifier is used to predict the universal clothing attribute labels. In addition, a triplet loss is added to pull in the features of the similar clothes worn by different people and push away the features of different clothes worn by the same person. The features of similar clothes worn by different people will be clustered together thus containing no identity information.

$$L_{CA} = \sum_{i=1}^{N}(-y_c log(\hat{y_c})) \tag{1}$$

$$L_{Tri}^{CA} = \sum_{c=1}^{C}\sum_{a=1}^{K}[m + \max_{p=1..K}||f_c^a - f_c^p||_2 - \min_{n=1..C}||f_c^a - f_c^n||_2]_+ \tag{2}$$

$$L_{CAE} = L_{CA} + L_{Tri}^{CA} \tag{3}$$

where N indicates the number of training images, $y_c/\hat{y_c}$ represents the ground truth and predicted clothing attribute label respectively. f_c^a represents the clothing attribute feature of an anchor sample. f_c^p/f_c^n represents the clothing attribute feature of a positive/negative sample that belongs to the same/different class as the anchor sample. C, K stands for the number of clothing attributes and images of each clothing attribute in a triplet batch. $[\cdot]_+$ represents Hinge loss.

4.2 Clothing Attribute Filtering Branch

After getting the clothing attribute feature f^{CA}, the CAF branch distills knowledge from it aiming to decompose and disentangle the clothing feature and the identity feature so that the model would learn identity-related features that are unrelated to the worn clothing. Specifically, the original person image is sent into the backbone network to extract feature which contains both clothing and identity information. Then, the attention module is applied to decompose these two kinds of information and a clothing feature f^C and an identity feature f^{ID} are obtained. Another attention module is applied to get the final identity feature f^{ID} and then a fully-connected classifier is used to predict the identity. A triplet loss is added to improve the discrimination of the identity feature.

$$L_{ID} = \sum_{i=1}^{N}(-y_i log(\hat{y_i})) \tag{4}$$

$$L_{Tri}^{ID} = \sum_{i=1}^{P}\sum_{a=1}^{K}[m + \max_{p=1..K}||f_i^a - f_i^p||_2 - \min_{n=1..P}||f_i^a - f_i^n||_2]_+ \tag{5}$$

where y_i and $\hat{y_i}$ represent the ground truth and the predicted identity label respectively.

The clothing feature is passed through a fully-connected classifier to predict the universal clothing attribute labels. Besides, a consistent loss is added to distill the knowledge from the clothing attribute feature to explicitly ensure the clothing feature contains no identity information. In addition, an orthogonal loss is utilized to encourage the clothing and identity features to be decoupled so that the identity feature could include all identifiable information except clothing.

$$L_C = \sum_{i=1}^{N}(-y_c log(\hat{y_c})) \tag{6}$$

$$L_{CST} = \sum_{i=1}^{N} CF(f_i^{CA}, f_i^C) \tag{7}$$

$$L_{ORG} = \sum_{i=1}^{N} \frac{|f_i^C \cdot f_i^{ID}|}{||f_i^C||_2||f_i^{ID}||_2} \tag{8}$$

where *CF(·)* stands for consistency alignment function (Mean Square Error (MSE) function in this paper). f_i^{CA}, f_i^C, and f_i^{ID}represent the clothing attribute feature, the clothing feature, and the identity feature of the i'th image respectively. A·B represents dot product between vector A and vector B.

In the training stage, the overall loss function is,

$$L_{FD} = \alpha L_{ID} + \beta L_{Tri}^{ID} + \gamma L_C + \delta L_{CST} + \zeta L_{ORG} \tag{9}$$

where $\alpha,\ \beta,\ \gamma,\ \delta\ and\ \zeta$ are all weighting factors which control the contribution of each term.

4.3 Optimization

First of all, the clothing images are acquired in advance in an offline way and the universal clothing attribute labels are manually annotated. Then, we use Eq.3 to train the CAE branch for N_A epochs individually. Then, the clothing attribute feature f^{CA} is obtained for the CAF branch to distill knowledge. The Eq.9 is used to supervise the CAF branch training for N_F epochs with the CAE branch fixed.

In the testing stage, only the CAF branch network is remained to extract the identity feature for retrieving.

5 Experiments

In this section, we perform experiments on existing cloth-changing datasets including LTCC [2020b], PRCC [2021], VC-clothes [2020], NKUP [2020], our CSCC and short-term dataset Market-1501 [2015] to evaluate our approach.

5.1 Datasets

LTCC [2020b] contains 17,138 images of 152 identities captured by 12 cameras. Training set contains 9,576 images of 77 persons and the testing set contains 493 query images and 7,050 gallery images of 75 identities.

PRCC [2021] contains 33,698 images from 221 identities captured by 3 cameras. The whole dataset is divided into a training set containing 22,898 images of 150 identities and a test set containing 10,800 images of 71 identities.

VC-clothes [2020] contains 19,060 images of 512 identities with 4 cameras. 9,449 images of 256 identities are randomly selected as the training set, and the remaining 256 identities are divided into the testing set where contains 1,020 query images and 8,591 gallery images.

NKUP [2020] contains 9,738 images of 107 persons captured by 15 cameras. This dataset was split into 5,336 training images of 40 identities, 332 query images and 4,070 gallery images of 67 identities.

CSCC (Ours) contains 36,700 images of 267 identities captured by 13 cameras. This dataset was split into 21,392 training images of 134 identities, 1,614 query images, and 13,694 gallery images of 133 identities.

Market-1501 [2015] contains 32,668 images of 1,501 identities captured by 6 cameras. 12,936 images from 751 identities are selected for training and 19,732 images from 750 identities form the testing set.

5.2 Protocols and Metrics

Following [Qian *et al.*, 2020b], we use two settings in evaluation on our dataset CSCC as follows: (1)*Standard Setting*: For each query, the images in the test set with the same identity and same camera are discarded when evaluation. The test set contains both cloth-consistent and cloth-changing examples in this setting. (2)*Cloth-changing setting*: For each query, the images in the test set with the same identity and same camera (clothing) are discarded when evaluation. The test set contains only cloth-changing examples in this setting. The Cumulated Matching Characteristics (CMC) and mean average precision (mAP) are adopted as evaluation metrics.

As for the existing datasets, we follow the protocol as the authors proposed.

5.3 Implementation Details

Our model is implemented based on the Pytorch framework. The backbone network in our model is ResNet-50 [2016] pre-trained on ImageNet, other parameters are all initialized randomly. We employ JPPNet [2019] to get the clothing images. All images are resized to 384×192. The stochastic gradient descent (SGD) with a momentum of 0.9 is adopted. The hyper-parameters α, β, γ, δ, ζ in Eq.9 and m in Eq.5 are set to 1.0, 0.8, 0.3, 0.8, 0.3, 0.8 respectively. The learning rate

Methods	LTCC				PRCC		VC-clothes				NKUP(wo Face)	
	Standard		Cloth-changing		Same clothes	Cross-clothes	Seen suits		Unseen suits		Cloth-changing	
	Rank-1	mAP	Rank-1	mAP	Rank-1	Rank-1	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
ResNet-50 [He *et al.*, 2016]	58.8	25.9	20.0	9.0	74.8	19.4	27.5	28.8	18.4	19.5	9.0	6.7
HACNN [Li *et al.*, 2018]	60.2	26.7	21.5	9.2	82.4	21.8	68.6#	69.7#	49.6#	50.1#	13.5#	10.6#
MuDeep [Qian *et al.*, 2020a]	61.8	27.5	23.5	10.2	83.7#	22.6#	70.3#	48.5#	52.4#	54.2#	15.6#	11.1#
PCB [Sun *et al.*, 2018]	65.1	30.6	23.5	10.0	86.9	22.9	72.3	73.9	53.9	55.6	16.9	12.4
MGN [Wang *et al.*, 2018]	68.4#	32.4#	25.3#	11.5#	89.8#	25.9#	74.3#	75.2#	55.0#	57.3#	18.8	15.0
SPT+ASE [Yang *et al.*, 2021]	-	-	-	-	64.20	34.38	-	-	-	-	-	-
SE+CESD [Qian *et al.*, 2020b]	71.4	34.3	26.2	12.4	91.8*	37.6*	85.2*	79.1*	69.5*	65.5*	19.0*	16.0*
ReIDCaps [Huang *et al.*, 2019b]	70.4#	32.9#	26.1#	12.0#	92.3#	38.0#	86.5#	79.0#	70.3#	67.5#	20.7#	15.3#
AFD-Net[Xu *et al.*, 2021]	-	-	-	-	95.7	42.8	-	-	-	-	-	-
LSD[Yaghoubi *et al.*, 2021]	72.2	31.0	31.4	13.6	93.6	37.2	-	-	-	-	16.4	12.3
3DSL[Chen *et al.*, 2021]	-	-	31.2	14.8	-	**51.3**	-	-	79.9	**81.2**	-	-
UCAD(Ours)	**74.4**	**34.8**	**32.5**	**15.1**	**96.5**	45.3	**92.6**	**81.1**	**82.4**	73.8	**25.0**	**16.9**

Table 2: Comparative results on existing datasets. The best results are in bold. # denotes we use the source code the author provided and test on the dataset. * denotes that we reproduce the work and test on the dataset.

Methods	Standard		Cloth-changing	
	Rank-1	mAP	Rank-1	mAP
ResNet-50 [He *et al.*, 2016]	72.9	28.7	32.0	13.1
HACNN [Li *et al.*, 2018]	77.2	29.4	34.5	14.3
MuDeep [Qian *et al.*, 2020a]	78.9	30.4	34.3	14.2
PCB [Sun *et al.*, 2018]	81.5	32.1	37.6	15.5
SE+CESD [Qian *et al.*, 2020b]	87.5	35.8	47.6	21.9
ReIDCaps [Huang *et al.*, 2019b]	88.7	36.9	47.3	23.0
UCAL(Ours)	**92.0**	**40.9**	**53.8**	**25.9**

Table 3: Comparative results on CSCC.

Methods	Rank-1	mAP
HACNN [Li *et al.*, 2018]	91.2	75.7
PCB[Sun *et al.*, 2018]	93.2	81.7
MuDeep[Qian *et al.*, 2020a]	95.3	84.7
MGN[Wang *et al.*, 2018]	**95.7**	**86.9**
ReIDCaps[Huang *et al.*, 2019b]	89.0	72.7
AFD-Net[Xu *et al.*, 2021]	94.2	84.3
UCAD(Ours)(Identity)	88.7	71.8
UCAD(Ours)(Identity+Cloth)	92.6	79.5

Table 4: Comparative results on Market-1501.

of the CAE branch is set to 10^{-4} and trained for 200 epochs. The learning rate of the CAF branch is set to 10^{-3} at the beginning and decays to 10^{-4} and 10^{-5} after 70 and 170 epochs respectively. The whole training process lasts for 200 epochs.

5.4 Comparison with State-of-the-Arts

Some cloth-changing Re-ID models and short-term methods are our competitors. Comparative results on four cloth-changing datasets, CSCC and short-term dataset Market-1501 are shown in Tab.2, Tab.3 and Tab.4 respectively.

Long-term datasets. As is shown in Tab.2 and Tab.3, we have the following key findings: (1) Our method outperforms the short-term methods by a large margin as our method is specially designed for cloth-changing person Re-ID. (2) Our method still achieves improvements on five cloth-changing datasets compared with cloth-changing methods but underperforms 3DSL[Chen *et al.*, 2021] on dataset PRCC [Yang *et al.*, 2021]. The reason is that PRCC is a dataset that all people share the same clothes in one camera view. The 3D shape is unrelated to clothes thus more effective in this situation.

Methods	Standard		Cloth-changing	
	Rank-1	mAP	Rank-1	mAP
L_{Base}(Baseline)	81.1	31.0	42.6	19.9
L_{Base} + L_{CST}	86.2	35.9	46.4	22.7
L_{Base} + L_{ORG}	89.7	38.8	49.6	23.5
L_{Base} + L_{CST} + L_{ORG}	**92.0**	**40.9**	**53.8**	**25.9**

Table 5: Ablation studies of the proposed model on CSCC. L_{Base} denotes $L_{ID} + L_{Tri}^{ID} + L_C$.The best results are in bold.

Feature	Standard		Cloth-changing	
	Rank-1	mAP	Rank-1	mAP
Identity	92.0	**40.9**	**53.8**	**25.9**
Cloth	72.8	30.8	34.4	16.1
Identity + Cloth	**92.4**	38.3	47.9	23.7

Table 6: Ablation studies of different features.

Short-term datasets. As is shown in Tab.4. Our method achieves comparable results on Market-1501[Zheng *et al.*, 2015] compared with short-term methods. This indicates that our method is effective under both cloth-changing and cloth-consistent setting.

5.5 Ablation Studies

We conduct comprehensive and systematic ablation studies on CSCC to evaluate our method.

Effectiveness of losses. As is shown in Tab.5. Firstly, compared with the baseline model, adding the orthogonal loss or the consistent loss can lead to performance improvements, indicating their effectiveness. When they are both added to train the network, the maximum performance is observed.

Influences of identity and clothing feature. We change the retrieval feature to evaluate our model as shown in Tab.6. First, the maximum Rank-1 under the standard setting is obtained when we concatenate clothing and identity feature to retrieve which indicates combining clothing feature is useful when the gallery has cloth-consistent images. Second, the clothing is no more useful under the cloth-changing setting.

Influences of hyper-parameters. There are six hyper-parameters in our model, $\alpha, \beta, \gamma, \delta, \zeta$ in Eq.9 controlling contribution of each item and m in Eq.5 controlling the margin of the triplet loss. We set all parameters to 1 at the beginning,

Methods	Severe set				Moderate set			
	Standard		Cloth-changing		Standard		Cloth-changing	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
ResNet-50[He *et al.*, 2016]	72.9	28.7	32.0	13.1	70.9	40.1	38.2	28.5
HACNN [Li *et al.*, 2018]	77.2	29.4	34.5	14.3	72.0	41.6	39.8	29.0
MuDeep [Qian *et al.*, 2020a]	78.9	30.4	34.3	14.2	76.6	44.1	40.7	31.9
PCB [Sun *et al.*, 2018]	81.5	32.1	37.6	15.5	78.9	45.5	42.9	33.4
SE+CESD [Qian *et al.*, 2020b]	87.5	35.8	47.6	21.9	82.8	52.6	59.2	39.1
ReIDCaps[Huang *et al.*, 2019b]	88.7	36.9	47.3	23.0	84.3	54.9	59.9	39.4
UCAD(Ours)	**92.0**	**40.9**	**53.8**	**25.9**	**86.2**	**56.5**	**61.5**	**41.2**

Table 7: Comparative results on CSCC under moderate, and severe cloth-changing settings. The best results are in bold.

Value	Standard		Cloth-changing	
	Rank-1	mAP	Rank-1	mAP
0	89.7	38.8	49,6	23.5
0.2	91.0	40.0	52.1	25.2
0.4	88.3	37.9	49.1	23.1
0.6	90.1	38.6	49.4	23.5
0.8	**92.0**	**40.9**	**53.8**	**25.9**
1.0	90.4	39.6	51.4	24.5

Table 8: Influences of hyper-parameter δ.

Methods	Standard		Cloth-changing	
	Rank-1	mAP	Rank-1	mAP
LTCC(w/o) $\rightarrow$ CSCC	84.3	33.9	45.2	20.1
LTCC(Universal) $\rightarrow$ CSCC	90.2	38.9	50.1	24.2
w/o UCA Label	89.7	38.1	49.2	23.1
UCA Label	**92.0**	**40.9**	**53.8**	**25.9**

Table 9: Effectiveness of the universal clothing attribute label.

and change from 1 to 0 with intervals of 0.1 one by one with others fixed to find the optimal value. It's worth noting that the best results yield at the same value for five datasets. It suggests these hyper-parameters are robust to different datasets. Tab.8 shows example of the influences of δ.

Effectiveness of the universal clothing attribute label. As shown in Tab.9, w/o universal clothing attribute label (UCA Label) refers to treating one clothing worn by each pedestrian as a distinct category. LTCC(Universal) $\rightarrow$ CSCC refers to the CAE branch trained on the LTCC dataset with the universal clothing attribute labels is used when the whole network is trained on CSCC. Results indicate the effectiveness of the universal clothing attribute label and the CAE branch trained on one dataset with the universal clothing attribute labels can be transferred to another dataset.

Influences of face and resolution. w/o face in Tab.10 refers to using face detector MTCNN [Zhang *et al.*, 2016] to detect the faces and mask the faces. Resize in Tab.10 refers to resizing the images to half their original size (both width and height). We carried out experiments with our method UCAD. It can be observed in Tab.10 that there is no obvious performance degradation when we mask the faces and our method achieve improvements without faces even compared with SE+CESD with faces (w/o face in Tab.10 compared with SE+CESD in Tab.3). This indicates our method focus on all identifiable features (shape, haircut, face, body contour *etc.*) not just some kind of information (shape or face).

Methods	Standard		Cloth-changing	
	Rank-1	mAP	Rank-1	mAP
Resize & w/o face	84.9	30.5	40.7	17.3
Resize	86.4	31.9	43.1	19.1
w/o face	90.1	38.8	49.4	23.7
CSCC	**92.0**	**40.9**	**53.8**	**25.9**

Table 10: Influences of face and resolution.

5.6 Degree of Cloth-Changing: A New Criterion

Some methods [Yang *et al.*, 2021] are specifically designed for the moderate cloth-changing setting which also have promising applications. To promote the research of related works, we provide two variants of our dataset: moderate, and severe subset. The moderate subset contains identities that have cloth-changing both in summer and spring (autumn). Clothes in the moderate subset would not strongly affect a person's contour sketch. It contains 15,870 images of 233 identities. The severe set is complete CSCC with people changing clothes during four seasons. We conduct experiments on these two subsets. The results are listed in Tab.7.

We have the following key findings: (1) Higher performance under standard setting is observed on severe set compared with the moderate set. The reason is that the persons with cloth-changing over four seasons (34 persons in total) are not abundant. Winter clothes they wear are distinct so that models overfit on clothes. (2) Lower performance under cloth-changing setting is observed on severe set compared with the moderate set. The reason is that the images with the same clothes in the gallery are discarded, cloth-changing over four seasons especially heavy winter clothing severely obstructs discernible features. This suggests the severe set is more challenging under cloth-changing setting. Researchers can effectively evaluate their models from various aspects.

6 Conclusion

In this paper, we proposed a novel Universal Clothing Attribute Disentanglement (UCAD) method which can effectively weaken the influence of clothing and force the model to learn identity-related features which are unrelated to the worn clothing. To further study Re-ID, we construct a severe cloth-changing person Re-ID dataset called CSCC containing 36,700 images of 267 identities. In addition, two variants considering moderate and severe cloth-changing setting are provided for researchers to evaluate their models from various aspects. Moreover, experiments on existing cloth-changing datasets including our CSCC and short-term dataset Market-1501 demonstrate the effectiveness of UCAD.

Acknowledgments

This work was supported in part by the Research Project of ZJU-League Research & Development Center, Zhejiang Lab under Grant 2019KD0AB01.

References

[Chen *et al.*, 2019] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. *ICCV*, pages 371–381, 2019.

[Chen *et al.*, 2020] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *ICCV*, 2020.

[Chen *et al.*, 2021] Jiaxing Chen, Xinyang Jiang, Fudong Wang, Jun Zhang, Feng Zheng, Xing Sun, and Wei-Shi Zheng. Learning 3d shape feature for texture-insensitive person re-identification. In *CVPR*, 2021.

[Fan *et al.*, 2020] Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. Learning longterm representations for person re-identification using radio signals. *CVPR*, pages 10696–10706, 2020.

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

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *ICCV*, pages 2980–2988, 2017.

[Huang *et al.*, 2019a] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In *IJCNN*, pages 1–8. IEEE, 2019.

[Huang *et al.*, 2019b] 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. *Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2019.

[Huang *et al.*, 2019c] Yangru Huang, Peixi Peng, Yi Jin, Junliang Xing, Congyan Lang, and Songhe Feng. Domain adaptive attention model for unsupervised cross-domain person re-identification. *ArXiv*, abs/1905.10529, 2019.

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

[Li *et al.*, 2020a] Shuzhao Li, Huimin Yu, Wei Huang, and Jing Zhang. Attributes-aided part detection and refinement for person re-identification. *Pattern Recognit.*, 97, 2020.

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

[Liang *et al.*, 2019] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:871–885, 2019.

[Qian *et al.*, 2020a] Xuelin Qian, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Leader-based multi-scale attention deep architecture for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):371–385, 2020.

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

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

[Wan *et al.*, 2020] Fangbin Wan, Yang Wu, Xuelin Qian, and Yanwei Fu. When person re-identification meets changing clothes. *arXiv preprint arXiv:2003.04070*, 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. *Proceedings of the 26th ACM international conference on Multimedia*, 2018.

[Wang *et al.*, 2020] Kai Wang, Zhi Ma, Shiyan Chen, Jinni Yang, Keke Zhou, and Tao Li. A benchmark for clothes variation in person re-identification. *International Journal of Intelligent Systems*, 35(12):1881–1898, 2020.

[Xu *et al.*, 2021] Wanlu Xu, Hong Liu, Wei Shi, Ziling Miao, and Feihu Chen. Adversarial feature disentanglement for long-term person re-identification. In *IJCAI*, 2021.

[Yaghoubi *et al.*, 2021] Ehsan Yaghoubi, Diana Borza, Bruno Degardin, and Hugo Proença. You look so different! haven't i seen you a long time ago? *Image and Vision Computing*, 115:104288, 2021.

[Yang *et al.*, 2021] Qize Yang, Ancong Wu, and Weishi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:2029–2046, 2021.

[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. *CVPR*, pages 3397–3406, 2020.

[Zhang *et al.*, 2016] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23:1499–1503, 2016.

[Zhang *et al.*, 2020] Jiangning Zhang, Liang Liu, Chao Xu, and Yong Liu. Hierarchical and efficient learning for person re-identification. *ArXiv*, abs/2005.08812, 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. IEEE, 2015.

[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*, 2019.