Learning Shape Representations for Person Re-Identification under Clothing Change

Yu-Jhe Li[1,2] Xinshuo Weng[1] Kris M. Kitani[1]
[1]The Robotics Institute, Carnegie Mellon University
[2]Dept. Electrical and Computer Engineering, Carnegie Mellon University
{yujheli, xinshuow, kkitani}@cs.cmu.edu

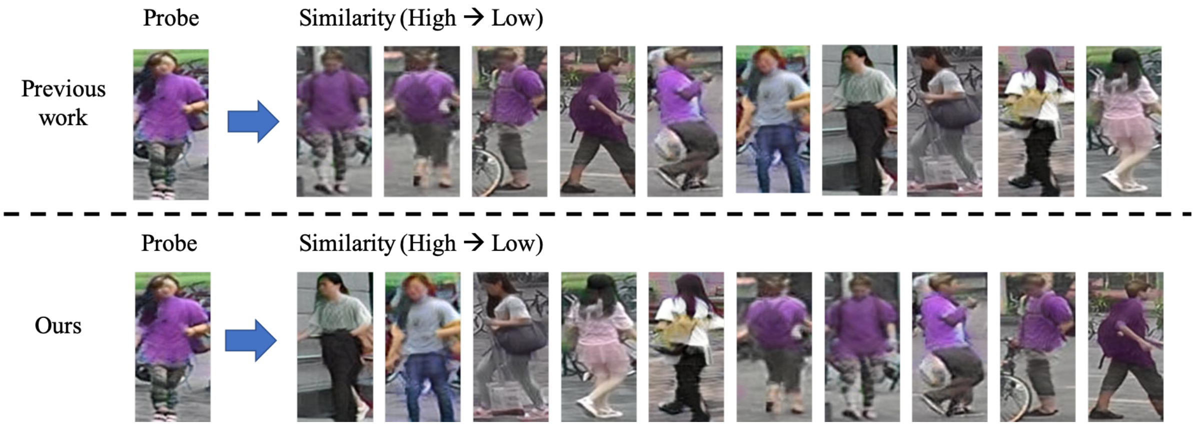


Figure 1: Given a probe image and ten gallery figures from our synthesized dataset (five from same identity and five with same color of clothes as the probe), we aim to prioritize matching the images with same body shape though in different wearings while previous work [59] is dominated by clothing color information.

Abstract

Person re-identification (re-ID) aims to recognize instances of the same person contained in multiple images taken across different cameras. Existing methods for re-ID tend to rely heavily on the assumption that both query and gallery images of the same person have the same clothing. Unfortunately, this assumption may not hold for datasets captured over long periods of time. To tackle the re-ID problem in the context of clothing changes, we propose a novel representation learning method which is able to generate a shape-based feature representation that is invariant to clothing. We call our model the Clothing Agnostic Shape Extraction Network (CASE-Net). CASE-Net learns a representation of a person that depends primarily on shape via adversarial learning and feature disentanglement. Quantitative and qualitative results across 5 datasets (Div-Market, Market1501, three large-scale datasets under clothing changes) show our approach makes significant improvements over prior state-of-the-art approaches.

1. Introduction

Person re-identification (re-ID) [58] aims to recognize the same person across multiple images taken by different cameras at different times. Re-ID is an important component technology in many applications such as person localization [34], multi-object tracking [43, 46, 37, 47, 48, 42], video surveillance [16] and computational forensics [38]. Despite recent advancements, existing methods [10, 21, 62, 29, 3, 28] rely (usually inadvertently) on the assumption that both query and gallery images of the same person will have the *same clothing*. While this is true in many existing re-ID datasets because they capture the data for every person over a very short time, the same clothing assumption does not hold in the real-world since people tend to change their clothes daily. As a result, it is unreasonable for Re-ID approaches to rely on the same clothing assumption and Re-ID approaches should be robust to clothing changes.

As standard re-ID datasets (*e.g.*, Market1501 [57] and DukeMTMC-reID [60, 27]) lack clothing change for the same person, the *clothing-dependence problem* has received little attention in prior work. Recent large-scale datasets started to address this issue [51, 49, 13, 12, 39]. Yang *et al*. [49] proposes the Person Re-id under moderate Clothing Change (PRCC) to study person re-id under clothing change. PRCC uses three non-overlapping cameras to capture every person wearing different clothes at each cam-

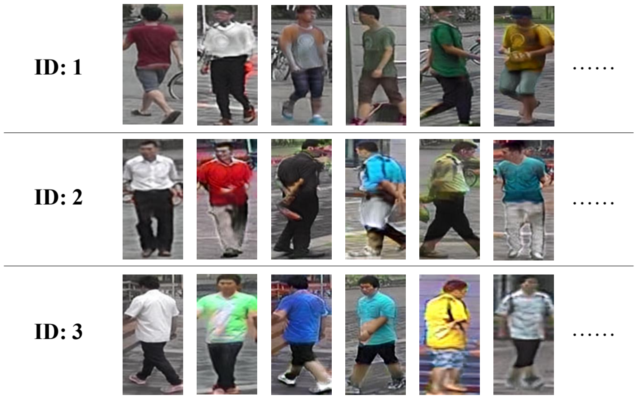


Figure 2: **Examples of our synthesized testing dataset *Div-Market*.** Div-Market is synthesized from Market-1501 with changes in clothing color. The example of each identity is shown in each row.

era. They additionally propose a learning-based spatial polar transformation and an angle specific extractor to improve performance under clothing change. Huang *et al*. [13] also proposes two large-scale re-ID datasets under clothing change: Celeb-reID [13] and Celeb-reID-light [12], which are collected from images of celebrities on the internet. Their proposed module: ReIDCAPS using capsule layers achieved the best result on their datasets. However, these methods require a huge amount of training images under clothing change to learn clothing-invariant features, which is expensive to collect.

To reduce the burden of collecting large-scale clothing change datasets while in the meantime being able to learn clothing-invariant features, one could imagine that [59] is one solution, which learns clothing-invariant features without requiring clothing change data. Specifically, [59] proposed the DG-Net to separately encode each person into an appearance feature representation and a structure feature representation through disentanglement, which improves performance on re-ID datasets without clothing change. Although their idea of disentangling structure (clothing-invariant) and appearance could be applied to address the clothing-dependence problem, we found that their use of the appearance feature to perform re-ID is still dominated by the clothing color. As shown in Fig. 1, given one probe image and ten gallery images (five same identities as the probe but with different clothing color, and five with the same clothing color as the probe but different identities), [59] is learned to match the gallery to the probe if they have similar clothing color, i.e., [59] primarily relies on clothing color for matching while ignoring other clothing-invariant cues such as shape.

In this work, we first introduce a synthetic dataset based on Market-1501 for testing modern re-ID methods and confirming their clothing-dependence problem. Then, in order to address the clothing-dependence problem without requiring large-scale training data with clothing change, we propose a new approach called *Clothing Agnostic Shape Extraction Network (CASE-Net)*, which learns body-structural representations via adversarial learning and structural disentanglement. In particular, we leverage gray-scaled images produced from RGB images to derive visual features of the same distribution across clothing color variations. Moreover, our model achieves structural disentanglement by performing image recovery when observing both gray-scaled and RGB images with pose variations. On the other hand, we synthesize dataset *Div-Market* with diverse changes in clothing across images of the same identity from Market1501 [57] using a generative model proposed by [59]. As depicted in Fig. 2, we synthetically change the color or texture of the original clothing. The goal of the synthesized dataset is to expose the clothing-dependency problem existing in state-of-the-art re-ID approaches in the scenario of changing clothes. We confirm in our experiments that all of our compared re-ID approaches exhibit severe performance drop when evaluated on our introduced dataset with clothing color changes. The contributions of this paper are highlighted below:

1. Quantitative and qualitative results on 5 re-ID datasets (our collected *Div-Market*, one standard benchmark, and three large-scale datasets with clothing change) uncover the weaknesses of prior state-of-the-art approaches in the clothing-changing scenarios.
2. We propose an end-to-end trainable network which learns shape-based representations that are invariant to clothing color and viewing perspective without supervision of data with clothing changes.
3. Without supervision on clothing change data, our model generalizes to the datasets with clothing change.

We believe that re-ID is beneficial if not critical for many applications such as assistive technologies, health activity monitoring, security and safety systems. These systems all require the ability to detect long-term correspondences, and we believe that there is sufficient justification for working on this problem. However, as with any advanced technology, there is a possibility of misuse or abuse. Just as this work reveals the dependence of prior re-ID methods on clothing color, we aim to continue to work towards identifying model bias, data bias and societal impact.

2. Related Works

Person Re-ID. Person re-ID has been widely studied in the literature. Existing methods typically focus on tackling the challenges of matching images with viewpoint and pose variations, or those with background clutter or occlusion presented [19, 4, 21, 15, 29, 2, 18, 22, 44, 31, 3, 28, 41, 36]. For example, Liu *et al*. [22] develop a pose-transferable deep learning framework based on GAN [7] to handle image pose variants. Chen *et al*. [3] integrate conditional random fields (CRF) and deep neural networks with multi-scale similarity metrics. Several attention-based methods [29, 18, 31] are further proposed to focus on learning the discriminative image features to mitigate the effect of background clutter. While promising results have been achieved in standard Re-ID datasets without clothing changes, most existing approaches cannot be easily applied to address the clothing dependence problem and are not robust to clothes or clothing-color changes.

Re-ID under Clothing Change. There are several small-scale [1, 26, 8] datasets and large-scale [51, 49, 13, 12, 39] datasets considering the clothing change problem for person re-ID. For the small ones, they can be categorized into two types: RGB-D and normal RGB video-based datasets. To handle the challenge of cloth changes, the depth information has been leveraged to extract an additional 3D soft-biometric beyond the RGB color cue. Accordingly, RGB-D datasets such as PAVIS [1], BIWI [26], IAS-Lab [26], and DPIT [8] have been proposed using the Kinect camera under controlled environments. However, RGB-D is not the mainstream in surveillance applications. The video-based re-ID dataset such as [53] was also proposed to use gait cue for person re-ID. Still, the scale of this dataset is also too small, and the environment is controlled under an indoor camera. On the other hand, Yang *et al*. [49] has proposed a large-scale re-ID dataset named ”Person Re-ID under moderate Clothing Change” (PRCC). PRCC uses three non-overlapping cameras to capture every person wearing different clothes at each camera. They additionally propose a learning-based spatial polar transformation (SPT) and an angle specific extractor (ASE) to improve the performance under clothing change. [13] also proposes two large-scale re-ID datasets under clothing change: Celeb-reID [13] and Celeb-reID-light [12], which are collected from images of celebrities on the internet. Their proposed module: ReIDCAPS using capsule layers achieves the best result on their datasets. However, these methods require a huge amount of training data with clothing change to learn clothing-invariant perception, which might not be applicable to the real-world scenario.

Disentanglement Re-ID. Recently, a number of models are proposed to better represent specific disentangled features during re-ID [32, 56, 55, 54, 17, 50, 45]. Ma *et al.* [24] generate person images by disentangling the input into foreground, background and pose with a complex multi-branch model which is not end-to-end trainable. Ge *et al.* [6] and Li *et al*. [20] learn pose-invariant features with guided image information. Zheng *et al.* [59] proposes a joint learning framework named DG-Net that couples re-ID learning and data generation end-to-end. Their model involves a generative module that separately encodes each person into an appearance code and a structure code, which leads to improved re-ID performance. However, their appearance encoder used to perform re-ID is still dominated by the clothing-color features corresponding to the input images. Based on the above observations, we choose to learn clothing-color invariant features using a novel and unified model. By disentangling the body shape representation, re-ID can be successfully performed in the scenario of clothing change even if no ground true images containing clothing change are available for training data.

3. CASE-Net

For the sake of the completeness, we define the notations to be used in this paper. In the training stage, we have access to a set of N RGB images $X_{\text{rgb}} = \{x_i^{\text{rgb}}\}_{i=1}^N$ and its corresponding label set $Y_{\text{rgb}} = \{y_i^{\text{rgb}}\}_{i=1}^N$, where $x_i^{\text{rgb}} \in \mathbb{R}^{H \times W \times 3}$ and $y_i^{\text{rgb}} \in \mathbb{N}$ are the i^{th} RGB image and its label, respectively. To allow our model to handle images of different color variations, we generate a gray-scaled image set $X_{\text{gray}} = \{x_i^L\}_{i=1}^N$ by multiplying each image from X_{rgb} with RGB channel summation factors, followed by duplicating the single channel back to the original image size (*i.e.* $x_i^{\text{gray}} \in \mathbb{R}^{H \times W \times 3}$). Naturally, the label set Y_{gray} for X_{gray} is identical to Y_{rgb}. In order to achieve body-shape distilling via image generation, we also sample another set of RGB images $X'_{rgb} = \{x_i'^{rgb}\}_{i=1}^N$, where its corresponding label set $Y'_{rgb} = \{y_i'^{rgb}\}_{i=1}^N$ is same as Y_{rgb} but with different pose and view point.

As depicted in Figure 3, CASE-Net consists of five components:(1) the shape encoder E_S, (2) the color encoder E_C, (3) the feature discriminator D_F, (4) the image generator G, and (5) the image discriminator D_I. We now describe how these models work together to learn a body shape feature which can be used for re-ID in domains that do not use color. Training CASE-Net results in learning a shape encoding and a color encoding of an image of a person. However, we are primarily interested in the body shape feature since it can be re-used for cross-domain (non-color dependent) re-ID tasks.

3.1. Clothing color adaptation in re-ID

Shape encoder (E_S). To utilize labeled information of training data for person re-ID, we employ classification loss on the output feature vector f^s (f_{rgb} and f_{gray}). With person identity as ground truth information, we can compute the negative log-likelihood between the predicted label $\tilde{y} \in \mathbb{R}^K$ and the ground truth one-hot vector $\hat{y} \in \mathbb{N}^K$, and define the identity loss $\mathcal{L}_{id}$ as

$$\begin{aligned} \mathcal{L}_{id} &= -\mathbb{E}_{(x_{\text{rgb}}, y_{\text{rgb}}) \sim (X_{\text{rgb}}, Y_{\text{rgb}})} \sum\nolimits_{k=1}^{K} \hat{y}_k^{\text{rgb}} \log(\tilde{y}_k^{\text{rgb}}) \\ &\quad - \mathbb{E}_{(x_{\text{gray}}, y_{\text{gray}}) \sim (X_{\text{gray}}, Y_{\text{gray}})} \sum\nolimits_{k=1}^{K} \hat{y}_k^{\text{gray}} \log(\tilde{y}_k^{\text{gray}}), \end{aligned} \tag{1}$$

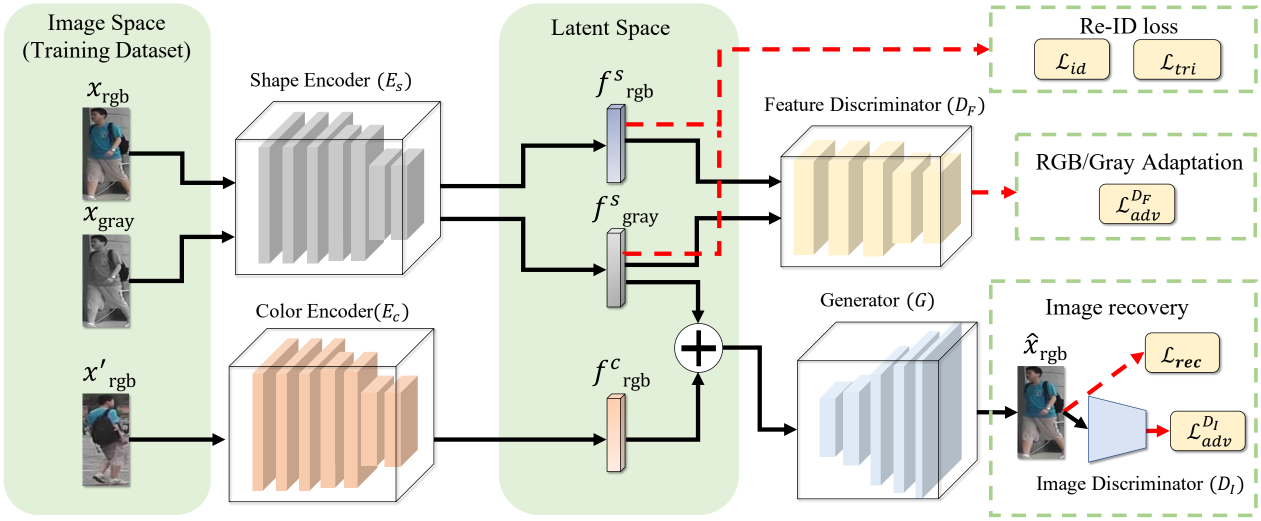


Figure 3: **Clothing Agnostic Shape Extraction Network (CASE-Net)**. The shape encoder E_S encodes input images across different color domains/datasets (x_{rgb} and x_{gray}) and produces color-invariant features f^s (f^s_{rgb} and f^s_{gray}). The color encoder E_C encodes the RGB images ($x\prime_{\text{rgb}}$) and produce color-related feature f^c. Then our feature discriminator D_F is developed to determine whether the input color-invariant features (f^s_{rgb} and f^s_{gray}) are from same distribution. Finally, the generator G jointly takes the color-invariant (f^s_{gray}) derived from gray-scaled image and color related feature (f^c_{rgb}) from RGB inputs, producing the synthesized RGB output images $\hat{x}_{\text{rgb}}$ while jointly training with additional image discriminator (D_I).

where K is the number of identities (classes). To further enhance the discriminative property, we impose a triplet loss $\mathcal{L}_{tri}$ on the feature vector f^s, which would maximize the inter-class discrepancy while minimizing intra-class distinctness. To be more specific, for each input image x, we sample a positive image x_{pos} with the same identity label and a negative image x_{neg} with different identity labels to form a triplet tuple. Then, the following equations compute the distances between x and $x_{\text{pos}}/x_{\text{neg}}$:

$$d_{\text{pos}} = \|f_x - f_{x_{\text{pos}}}\|_2, \tag{2}$$

$$d_{\text{neg}} = \|f_x - f_{x_{\text{neg}}}\|_2, \tag{3}$$

where f_x, $f_{x_{\text{pos}}}$, and $f_{x_{\text{neg}}}$ represent the feature vectors of images x, x_{pos}, and x_{neg}, respectively. With the above definitions, we have the triplet loss $\mathcal{L}_{tri}$ defined as

$$\begin{aligned}\mathcal{L}_{tri} &= \mathbb{E}_{(x_{\text{rgb}},y_{\text{rgb}})\sim(X_{\text{rgb}},Y_{\text{rgb}})} \max(0, m + d^{\text{rgb}}_{\text{pos}} - d^{\text{rgb}}_{\text{neg}}) \\ &+ \mathbb{E}_{(x_{\text{gray}},y_{\text{gray}})\sim(X_{\text{gray}},Y_{\text{gray}})} \max(0, m + d^{\text{gray}}_{\text{pos}} - d^{\text{gray}}_{\text{neg}}),\end{aligned} \tag{4}$$

where $m > 0$ is the margin used to define the distance difference between the distance of positive image pair d_{pos} and the distance of negative image pair d_{neg}.

Feature discriminator (D_F). Next, since our goal is to derive body-shape representations which do not depend on clothing-color, we first learn color-invariant representation by encouraging the content encoder E_S to generate similar feature distributions when observing both X_{rgb} and X_{gray}. To achieve this, we advance adversarial learning strategies and deploy a feature discriminator D_F in the latent *feature space*. This discriminator D_F takes the feature vectors f_{rgb} and f_{gray} as inputs to determine whether the input feature vectors are from X_{rgb} or X_{gray}. To be more precise, we define the feature-level adversarial loss $\mathcal{L}^{D_F}_{\text{adv}}$ as

$$\begin{aligned}\mathcal{L}^{D_F}_{\text{adv}} &= \mathbb{E}_{x_{\text{rgb}}\sim X_{\text{rgb}}}[\log(D_F(f_{\text{rgb}}))] \\ &+ \mathbb{E}_{x_{\text{gray}}\sim X_{\text{gray}}}[\log(1 - D_F(f_{\text{gray}}))],\end{aligned} \tag{5}$$

where $f_{\text{rgb}} = E_S(x_{\text{rgb}})$ and $f_{\text{gray}} = E_S(x_{\text{gray}}) \in \mathbb{R}^d$ denote the encoded RGB and gray-scaled image features, respectively.[1] With loss $\mathcal{L}^{D_F}_{\text{adv}}$, our feature discriminator D_F distinguish the features from two distributions while our shape encoder E_S aligns the feature distributions across color variations, carrying out the learning of color-invariant representations for clothing via adversarial manner.

3.2. Pose guidance for body shape disentanglement

Color encoder (E_C). To ensure our derived feature is body-shape related in clothing-color changing tasks, we need to perform additional body-shape disentanglement during the learning of our CASE-Net. That is, we have the color encoder E_C in Fig. 3 encodes the inputs from RGB images set X'_{rgb} into color-related features f^c_{rgb}. As a result, both gray-scaled body-shape and color features would be produced in the latent space. Inspired by DG-Net [59] using gray-scaled image to achieve body-shape disentanglement across pose variations, we similarly enforce the our generators G to produce the person images conditioned on the en-

[1] For simplicity, we omit the subscript i, denote RGB and gray-scaled images as x_{rgb} and x_{gray}, and represent their corresponding labels as y_{rgb} and y_{gray}.

coded color feature coming from different pose. To be precise, we have the generator take the concatenated shape and color feature pair $(f^s_{\text{gray}}, f^c_{\text{rgb}})$ and output the corresponding image $\hat{x}_{\text{rgb}}$.

Image generator (G). Since we have ground truth labels (i.e., image pair correspondences) from the training data, we can perform an image recovery task given two images x_{rgb} and $x_{\text{rgb}}{}'$ of the same person but with different poses, we expect that they share the same body-shape feature f^c_{gray}. Given the feature pair $(f^s_{\text{gray}}, f^c_{\text{rgb}})$, we then enforce G to output the image $\hat{x}_{\text{rgb}}$ using the body-shape feature f^s_{gray} which is originally associated with x_{rgb}. This is referred to as *Pose guided* image recovery. With the above discussion, image reconstruction loss $\mathcal{L}_{\text{rec}}$ can be calculated as:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{x_{rgb}\sim X_{rgb}, x_{gray}\sim X_{gray}, x'_{rgb}\sim X'_{rgb}}[\|\hat{x}_{rgb} - x_{rgb}\|_1], \tag{6}$$

where $\hat{x}_{rgb}$ denotes $\hat{x}_{rgb} = G(f^s_{gray}, f^c_{rgb})$. Note that we adopt the L1 norm in the above reconstruction loss terms as it preserves image sharpness [11].

Image discriminator (D_I). To further enforce G perform perceptual content recovery, we produce perceptually realistic outputs by having the image discriminator D_I discriminate between the real images x_{rgb} and the synthesized ones $\hat{x}_{rgb}$. To this end, we have both reconstruction loss and perceptual discriminator loss for image recovery. Thus, the image perceptual discriminator loss $\mathcal{L}^{D_I}_{adv}$ as

$$\begin{aligned}\mathcal{L}^{D_I}_{adv} &= \mathbb{E}_{x_{rgb}\sim X_{rgb}}[\log(D_I(x_{rgb}))] \\ &+ \mathbb{E}_{x_{gray}\sim X_{gray}, x'_{rgb}\sim X'_{rgb}}[\log(1 - D_I(\hat{x}_{rgb}))].\end{aligned} \tag{7}$$

To perform person re-ID in the testing phase, our network encodes the query image by E_S for deriving the body shape feature f_s, which is applied for matching the gallery ones via nearest neighbor search (in Euclidean distances). We will detail the properties of each component in the following subsections.

It is important to note that the goal of CASE-Net is to perform re-ID in clothing changing scenario without observing ground true clothing changing training data. By introducing the aforementioned network module, our CASE-Net would be capable of performing re-ID in environments with clothing changes. More precisely, with the joint training of encoders/generator and the feature discriminator, our model allows learning of body-structural representation. The pseudo code for training our CASE-Net using above losses is summarized in Algorithm 1, where λ_{tri} and λ_I are hyper-parameters.

4. Experiments

4.1. Datasets

To evaluate our method, we conduct experiments on three current large-scale image-based re-ID datasets with clothing change: Celeb-reID [13], Celeb-reID-light [12], and PRCC [49]. one of our synthesized datasets Div-Marke, and one benchmark re-ID dataset Market-1501 [57], which is commonly considered in recent re-ID tasks.

Algorithm 1: Learning of CASE-Net

```
Data: Image set: $X_{rgb}$, $X_{gray}$, $X'_{rgb}$; Label set: $Y_{rgb}$,
      $Y_{gray}$
Result: Configurations of CASE-Net
$\theta_{E_S}, \theta_{E_C}, \theta_{D_F}, \theta_G, \theta_{D_I} \leftarrow$ initialize
for Num. of training Iters. do
    $x_{rgb}, x_{gray}, x'_{rgb}, y_{rgb}, y_{gray} \leftarrow$ sample from $X_{rgb}$,
      $X_{gray}, X'_{rgb}, Y_{rgb}, Y_{gray}$
    $f^s_{rgb}, f^s_{gray}, f^c_{rgb} \leftarrow$ obtain by $E_S(x_{rgb})$, $E_S(x_{gray})$,
      $E_C(x'_{rgb})$
    $\mathcal{L}_{\text{id}}, \mathcal{L}_{tri} \leftarrow$ calculate by (1), (4)
    $\theta_{E_S} \xleftarrow{+} -\nabla_{\theta_{E_S}}(\mathcal{L}_{\text{id}} + \lambda_{tri}\mathcal{L}_{tri})$
    $\hat{x}_{rgb} \leftarrow$ obtain by $G(f^s_{gray}, f^c_{rgb})$
    $\mathcal{L}^{D_F}_{\text{adv}}, \mathcal{L}_{rec}, \mathcal{L}^{D_I}_{\text{adv}} \leftarrow$ calculate by (5), (6), (7)
    for Iters. of updating generator do
        $\theta_{E_S} \xleftarrow{+} -\nabla_{\theta_{E_S}}(-\mathcal{L}^{D_F}_{\text{adv}})$
        $\theta_{E_S,E_C,G} \xleftarrow{+} -\nabla_{\theta_{E_S,E_C,G}}(\mathcal{L}_{rec} - \lambda_I\mathcal{L}^{D_I}_{\text{adv}})$
    for Iters. of updating discriminator do
        $\theta_{D_F} \xleftarrow{+} -\nabla_{\theta_{D_F}}\mathcal{L}^{D_F}_{\text{adv}}$
        $\theta_{D_I} \xleftarrow{+} -\nabla_{\theta_{D_I}}\mathcal{L}^{D_I}_{\text{adv}}$
```

Celeb-reID. Celeb-reID [13] is composed of 34036 images of 1052 identities (celebrities) and it is crawled from Google, Bing, and Baidu websites. The dataset is split into two non-overlapping parts: 20,208 images from 632 identities for training and 13978 images from 420 identities for testing. Among the testing dataset, 2972 images are for query while 11006 are for gallery. More than 70% of the images of each person show different clothes on average while a person may wear the same clothes (the ratio is less than 30% within each ID) in Celeb-reID.

Celeb-reID-light. Celeb-reID-light [12] or Celebrities-reID is a light version of Celeb-reID. Unlike Celeb-reID, a person in Celeb-reID-light will not wear the same cloth twice. There are 590 identities: 490 identities with 9,021 images are used for training, and 100 identities with 1,821 images are used for testing. In the testing set, 887 images are used as queries, and 934 images are used as galleries. Although the scale of Celeb-reID-light is smaller than Celeb-reID, it can be used to testify the robustness of re-ID methods when a person is entirely in different clothes.

PRCC. Person Re-ID under moderate Clothing Change (PRCC) [49] dataset consists of 33698 images from 221 identities. Each person in Cameras A and B is wearing the same clothes, but the images are captured in different rooms. For Camera C, the person wears different clothes, and the images are captured in a different day. Following [49], the dataset is randomly split into a training set and a testing set. The training set consist of 150 people while the testing set consist of 71 people. Each image from this dataset has the corresponding contour sketch image for use.

Market-1501. Market-1501 [57] is composed of 32,668 labeled images of 1,501 identities collected from 6 camera views. The dataset is split into two non-over-lapping fixed parts: 12,936 images from 751 identities for training and 19,732 images from 750 identities for testing. In testing, 3368 query images from 750 identities are used to retrieve the matching persons in the gallery.

Div-Market. Div-Market is our small synthesized dataset generated from Market-1501. We use our generative model similar as [59] to change the clothing-color in the images of each identity. It contains a total of 24732 images of 200 identities, each with hundreds of figures. *We only use this dataset for testing*. We would like to note that, the model for generating the images is separate from our designed CASE-Net. The generative model is just used for producing images with different clothing.

4.2. Implementation Details

We implement our model using PyTorch. Following Section 3, we use ResNet-50 pre-trained on ImageNet as our backbone of shape encoder E_S and color encoder E_C. Given an input image x (all images are resized to size $256 \times 128 \times 3$, denoting width, height, and channel respectively.), E_S encodes the input into 2048-dimension content feature f^s. The structure of the generator is 6 convolution-residual blocks similar to that proposed by Miyato *et al.* [25]. The structure of the image discriminator D_I employs the ResNet-18 as backbone while the architecture of shared feature discriminator D_F adopts is composed of 5 convolution blocks in our CASE-Net. These three components (other than E_S and E_C) are randomly initialized. The margin for the $\mathcal{L}_{tri}$ is set as 2.0, and we fix λ_{tri} and λ_I as 1.0 and 0.1, respectively. Note that we do not overfit the datasets by selecting the best parameters through parameter-searching experiments. The performance of our method can be possibly further improved by parameter searching, applying pre/post-processing methods, attention mechanisms, or re-ranking techniques. However, such techniques are not used in all of our experiments.

4.3. Evaluation Settings and Protocol

Table 1: **Quantitative results on the Celeb-reID and Celeb-reID-light dataset.** The first large row shows the approaches for standard re-ID while the second large row indicates the methods for clothing change. To fit our model, * indicates the results are produced by the model additionally trained on the images where the person wears same clothes but with different pose from Celeb-reID.

Method	Celeb-reID [13]			Celeb-reID-light [12]		
	Rank1	Rank5	mAP	Rank1	Rank5	mAP
IDE [58] + DenseNet	42.9	56.4	5.9	10.5	24.8	5.3
Verif-Identif [61]	36.3	54.5	7.8	-	-	-
MLFN [2]	41.4	54.7	6.0	10.6	31.0	6.3
HACNN [18]	47.6	63.3	9.5	16.2	42.8	11.5
Part-aligned [33]	19.4	40.6	6.4	-	-	-
PCB [36]	37.1	57.0	8.2	-	-	-
MGN [40]	49.0	64.9	10.8	21.5	47.4	13.9
ReIDCaps [13]	51.2	65.4	9.8	20.3	48.2	11.2
ReIDCaps+ [13]	63.0	76.3	15.8	**33.5**	**63.3**	**19.0**
Ours	**66.4**	**78.1**	**18.2**	35.1*	66.7*	20.4*

Table 2: **Quantitative results on the PRCC [49] dataset.** The common methods for re-ID and methods designed for clothing change are split into two big rows.

Method	Cross clothes			Same clothes		
	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10
Verif-Identif [61]	19.10	53.83	65.41	76.27	97.24	98.53
VGG16 [30]	18.21	46.13	60.76	71.39	95.89	98.68
ResNet-50 [9]	19.43	52.38	66.43	74.80	97.28	98.85
HACNN [18]	21.81	59.47	67.45	82.45	98.12	99.04
PCB [36]	22.86	61.24	78.27	**86.88**	**98.79**	**99.62**
SketchNet [52]	17.89	43.70	58.62	64.56	95.09	97.84
Deform. Conv. [5]+ASE [49]	25.98	71.67	85.31	61.87	92.13	97.65
STN [14]+ASE [49]	27.47	69.53	83.22	59.21	91.43	96.11
SPT [49]+ASE [49]	34.38	77.30	88.05	64.20	92.62	96.65
Ours	**39.51**	**80.42**	**91.23**	71.20	97.13	98.54

For **Celeb-reID**, **Celeb-reID-light**, and **PRCC**, we train our model on the training set and evaluate the performance on the testing set (query set and gallery set). We note that, these three datasets have images from identities with clothing change. To be specific, we train our color encoder on images from the identities which have different pose while update our shape encoder using all of the images (containing images with different clothing). For **Market-1501** and our synthesized testing set **Div-Market**, we evaluate the models trained only with Market-1501 (without clothing change) on testing set of Market-1501 and Div-Market.

We employ the standard metrics as in most person Re-ID literature, namely the cumulative matching curve (CMC) used for generating ranking accuracy, and the mean Average Precision (mAP). We report rank-1 accuracy and mean average precision (mAP) for evaluation on all datasets.

4.4. Comparison with State-of-the-Art

Celeb-reID and Celeb-reID-light. We compare our proposed method with several common re-ID approaches and the methods for clothing variation. These standard approaches include IDE [58], Verif-Identif [61], MLFN [2], HACNN [18], Part-aligned [33], PCB [36], and MGN [40] while methods for clothing change involve ReIDCaps and ReIDCaps+ [13]. ReIDCaps+ indicates the method is using body parts partition. As the reported results presented

Table 3: **Quantitative results of person re-ID on the Market-1501 and Div-Market dataset.** Note that all the reported results on Div-Market are reproduced using released codes available online. The first large row shows the approaches for standard re-ID while the second large row indicates the methods for clothing change.

Method	Market-1501				Div-Market			
	R1	R5	R10	mAP	R1	R5	R10	mAP
Verif-Identif [61]	79.5	86.0	90.3	61.5	9.2	23.9	34.6	1.0
SVDNet [35]	82.2	92.3	93.9	62.4	9.8	25.1	35.5	1.3
FD-GAN [6]	90.5	96.0	97.7	77.9	14.3	26.4	36.5	1.6
Part-aligned [33]	93.8	97.7	98.3	79.9	14.9	27.4	36.1	1.8
PCB [36]	93.2	97.3	98.2	81.7	15.7	27.0	39.5	1.7
DG-Net [59]	94.4	98.4	98.9	85.2	19.7	30.1	47.5	2.2
Bag of tricks [23]	**95.0**	-	-	**88.2**	20.5	30.8	49.0	2.2
ReIDCaps [13]	89.0	-	-	72.7	14.0	25.5	35.7	1.5
ReIDCaps+ [13]	92.8	-	-	78.0	15.0	27.7	37.5	1.9
Ours	94.6	98.9	99.1	85.7	**56.2**	**61.5**	**69.2**	**13.5**

Table 4: **Ablation study of the loss functions on the Div-Market dataset.** We note that, each row indicates the model with only one loss excluded.

Method	Rank 1	Rank 5	Rank 10	mAP
Ours (full model)	**56.2**	**61.5**	**69.2**	**13.5**
Ours w/o $\mathcal{L}_{\text{adv}}^{D_I}$	55.7	60.4	66.8	10.1
Ours w/o $\mathcal{L}_{\text{tri}}$	54.0	58.1	66.3	8.7
Ours w/o $\mathcal{L}_{\text{id}}$	50.5	57.6	65.5	8.2
Ours w/o $\mathcal{L}_{\text{adv}}^{D_F}$	49.8	51.5	64.1	7.3
Ours w/o $\mathcal{L}_{\text{rec}}$	46.5	50.1	61.5	6.9

on the left side of Table 1, our proposed CASE-Net outperforms all the compared methods on Celeb-reID. For Celeb-reID-light, since this dataset does not contain person wearing same clothes for learning the disentanglement, we only compare the result when we train our model on images with same clothes from Celeb-reID. While the comparison on Celeb-reID-light may not be fair, we still report the results for reference.

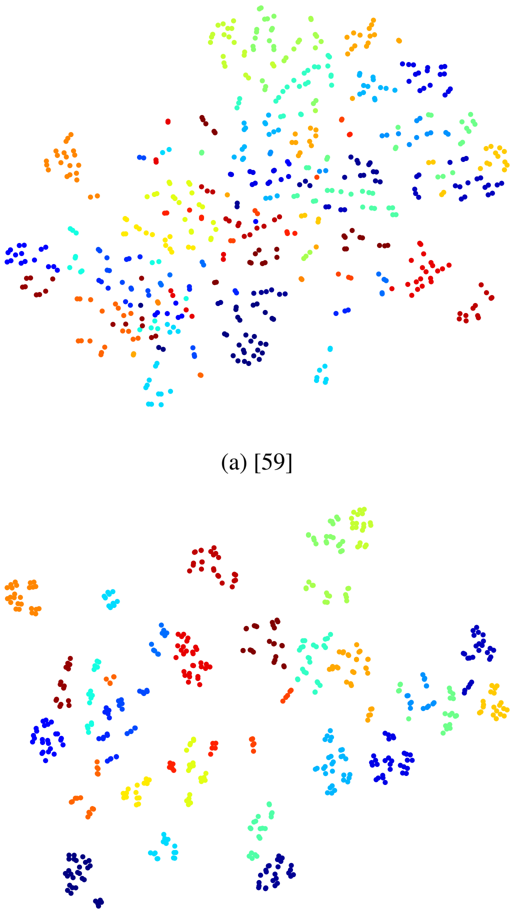
PRCC. We also compare our method with five common deep approaches and four methods for clothing variation specifically on this dataset. These common approaches include Verif-Identif [61], backbones with VGG16 [30] and ResNet-50 [9], HACNN [18], and PCB [36] while methods for clothing change on this dataset involve SketchNet [52], Deform. Conv. [5]+ASE [49], STN [14]+ASE [49], and SPT [49]+ASE [49]. As the reported results presented on the left side of Table 2, our proposed CASE-Net outperforms all the compared methods under clothing change. In addition, some phenomenons can also be observed. First, we found severe performance drops under clothing change in all the common approaches, which indicates standard re-ID approaches all suffer from *clothing-color/clothes* mismatch problems. Second, our method is more stable comparing with all of the four methods for clothing change when testing under the same clothing. Third, five common deep approaches seem to be dominated by clothing when they outperform the competitors when testing under the same clothing.

Market-1501 and **Div-Market.** We compare our proposed method with seven current standard re-ID approaches and one method for clothing variation. We reported the results on both Market-1501 and Div-Market. These standard approaches include Verif-Identif [61], SVDNet [35], Part-aligned [33], FD-GAN [6], PCB [36], DG-Net [59], and Bag of tricks [23] while method for clothing change involve ReIDCaps and ReIDCaps+ [13]. We report all the results in Table 3 and several phenomenons can be observed which we summarized as three folds. Firstly, state-of-the-arts methods outperform our approach by a small margin but suffer severe performance drop on the Div-Market, which shows their vulnerability to clothing variations and weak generalization when *they train to overfit on the clothing color.* Second, our proposed CASE-Net outperforms all the methods on Div-market, which demonstrates that our ability to derive body shape representation without supervision on clothing change.

4.5. Ablation Studies

Loss functions. To further analyze the importance of each introduced loss function, we conduct an ablation study shown in Table 4. Firstly, the feature adversarial loss $\mathcal{L}_{rec}$ is shown to be vital to our CASE-Net, since we observe 10% drops on Div-Market when the loss was excluded. This is caused by no explicit supervision to guide our CASE-Net to generate human-perceivable images with body shape disentanglement, and thus the resulting model would suffer from image-level information loss. Secondly, without the feature adversarial loss $\mathcal{L}_{\text{adv}}^{D_F}$, our model would not be able to perform feature-level color adaptation, causing failure on learning clothing color invariant representation and resulting in the re-ID performance drop (about 7%). Thirdly, when either $\mathcal{L}_{id}$ or $\mathcal{L}_{tri}$ is turned off, our model is not able to be supervised using two re-ID losses, indicating that jointly use of two streams of supervision achieve best results. Lastly, the image adversarial loss $\mathcal{L}_{\text{adv}}^{D_I}$ is introduced to our CASE-Net to mitigate the perceptual image-level information loss.

Feature-level visualization. We visualize feature vectors f^s on our Div-Market in Figures 4 via t-SNE. It is worth emphasizing that, in our synthesized Div-Market, the same identity can have different wearings while different q identities can have the same wearing. We select 30 different persons, each of which is indicated by a color. From Fig. 4a and



(a) [59]

(b) Ours

Figure 4: Visualization of structure feature vectors f^s on Div-Market via t-SNE. We visualize 30 different identities, each of which is shown in a unique color and compare our results with [59].

Fig. 4b, we observe that our projected feature vectors are well separated when it compared with DG-Net [59], which suggests that sufficient re-ID ability can be exhibited by our model. Although the idea of DG-Net [59] is also disentangling structure (shape) and appearance (color), their use of appearance to perform re-ID is still dominated by clothing color information. This shows that existing methods are trained to focus only on matching clothing color while ignoring other identity-related cues such as shape.

Image-level visualization. We visualize the reconstructed images $\hat{x}_{\text{rgb}}$ given only one fixed x_{rgb} as input and different x'_{rgb} as input. As shown in the Fig. 5, the clothing of output $\hat{x}_{\text{rgb}}$ depends on the input x'_{rgb} of the color encoder while the body-shape of the output depends on the input x_{rgb} of the shape encoder. This infers that our shape encoder has successfully learned the description of body shape as feature representations while our color encoder does derive the clothing from the input image. By using images with different poses, our CASE-Net indeed performs the body-shape disentanglement via image generator and discriminator.

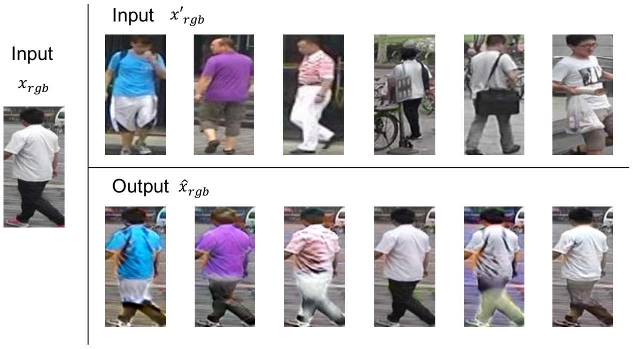


Figure 5: **Examples of our body-shape disentanglement.** The body-shape of the output depends on the input x_{rgb} of the shape encoder which is invariant to clothing (input x'_{rgb})

5. Conclusions

In this paper, we unfolded a challenging yet significant person re-identification task which has been long ignored in the past. We collected a re-ID dataset (*Div-Market*), which contains changes in clothes or clothing-color. To address clothing changes in re-ID, we presented a novel Clothing Agnostic Shape Extraction Network (CASE-Net) which learns body shape representation without training on data containing clothing change. By advancing the adversarial learning and body shape disentanglement, our model resulted in satisfactory performance on three recent re-ID datasets with clothing change, the collected dataset (Div-Market), and one re-ID benchmark (Market-1501). Qualitative results also confirmed that our model is capable of learning body shape representation, which is clothing-color invariant. Qualitative results showed our approach had exhibited body-shape disentanglement while it was learned without supervision under the clothing change.

Acknowledgements. This project was sponsored in part by IARPA (D17PC00340).

References

[1] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with rgb-d sensors. In *ECCV*. Springer, 2012.

[2] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018.

[3] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *CVPR*, 2018.

[4] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016.

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017.

[6] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NeurIPS*, 2018.

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

[8] Albert Haque, Alexandre Alahi, and Li Fei-Fei. Recurrent attention models for depth-based person identification. In *CVPR*, 2016.

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

[10] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. In *arXiv preprint*, 2017.

[11] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

[12] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019.

[13] 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.

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, 2015.

[15] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018.

[16] Furqan M Khan and François Brémond. Person re-identification for real-world surveillance systems. *arXiv*, 2016.

[17] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.

[18] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

[19] Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. Recover and identify: A generative dual model for cross-resolution person re-identification. In *ICCV*, 2019.

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

[21] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. In *arXiv preprint*, 2017.

[22] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *CVPR*, 2018.

[23] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 2019.

[24] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018.

[25] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *ICLR*, 2018.

[26] Matteo Munaro, Alberto Basso, Andrea Fossati, Luc Van Gool, and Emanuele Menegatti. 3d reconstruction of freely moving persons for re-identification with a depth sensor. In *ICRA*. IEEE, 2014.

[27] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[28] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *CVPR*, 2018.

[29] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, 2014.

[31] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018.

[32] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *CVPR*, 2017.

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

[34] Xi Sun, Xinshuo Weng, and Kris Kitani. When We First Met: Visual-Inertial Person Localization for Co-Robot Rendezvous. *IROS*, 2020.

[35] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *arXiv preprint*, 2017.

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

[37] Siyu Tang and Max Planck. Multiple People Tracking by Lifted Multicut and Person Re-Identification. *CVPR*, 2017.

[38] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 2013.

[39] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPRW*, 2020.

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

[41] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, 2018.

[42] Yongxin Wang, Xinshuo Weng, and Kris Kitani. Joint Detection and Multi-Object Tracking with Graph Neural Networks. *arXiv:2006.13164*, 2020.

[43] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards Real-Time Multi-Object Tracking. *ECCV*, 2020.

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

[45] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*, 2017.

[46] Xinshuo Weng, Yongxin Wang, Yunze Man, and Kris Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. *CVPR*, 2020.

[47] Xinshuo Weng, Yongxin Wang, Yunze Man, and Kris Kitani. Graph Neural Network for 3D Multi-Object Tracking. *ECCVW*, 2020.

[48] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple Online and Realtime Tracking with a Deep Association Metric. *ICIP*, 2017.

[49] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *TPAMI*, 2020.

[50] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *TIP*, 2019.

[51] 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*, 2020.

[52] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *CVPR*, 2016.

[53] Peng Zhang, Qiang Wu, Jingsong Xu, and Jian Zhang. Long-term person re-identification using true motion from videos. In *WACV*, 2018.

[54] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.

[55] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *CVPR*, 2017.

[56] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. In *arXiv preprint*, 2017.

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

[58] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. In *arXiv preprint*, 2016.

[59] 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.

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

[61] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2018.

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