Learning 3D Shape Feature for Texture-insensitive Person Re-identification

Jiaxing Chen[1,5#], Xinyang Jiang[3#], Fudong Wang[3], Jun Zhang[3], Feng Zheng[4], Xing Sun[3], Wei-Shi Zheng[1,2*]
[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Youtu Lab, Tencent
[4]CSE, Southern University of Science and Technology
[5]Pazhou Lab, Guangzhou, China
chenjx228@mail2.sysu.edu.cn,xinyangj9024@gmail.com,{fudongwang,bobbyjzhang}@tecent.com
zhengf@sustech.edu.cn,winfredsun@tencent.com,wszheng@ieee.org

Abstract

*It is well acknowledged that person re-identification (person ReID) highly relies on visual texture information like clothing. Despite significant progress has been made in recent years, texture-confusing situations like clothing changing and persons wearing the same clothes receive little attention from most existing ReID methods. In this paper, rather than relying on texture based information, we propose to improve the robustness of person ReID against clothing texture by exploiting the information of a person's 3D shape. Existing shape learning schemas for person ReID either ignore the 3D information of a person, or require extra physical devices to collect 3D source data. Differently, we propose a novel ReID learning framework that directly extracts a texture-insensitive 3D shape embedding from a 2D image by adding 3D body reconstruction as an auxiliary task and regularization, called 3D Shape Learning (3DSL). The 3D reconstruction based regularization forces the ReID model to decouple the 3D shape information from the visual texture, and acquire discriminative 3D shape ReID features. To solve the problem of lacking 3D ground truth, we design an adversarial self-supervised projection (**ASSP**) model, performing 3D reconstruction without ground truth. Extensive experiments on common ReID datasets and texture-confusing datasets validate the effectiveness of our model.*

1. Introduction

The aim of person ReID is to find the target person among an existing set of persons captured by a distributed camera system. Some works [7, 34, 44, 46] have demonstrated that person ReID largely depends on clothing appearance textures, and most of existing methods decline a lot when clothing texture is confusing. Considering clothing texture-confusing situations (see Figure 1) that people might change their clothing [44] or different people wear very similar clothing [46], clothing texture would become unreliable for ReID. Situations like suspects wearing different clothes, or different people wearing similar uniforms in hospitals or schools are ubiquitous.

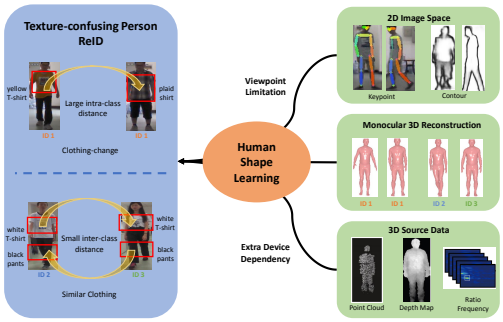


Figure 1. Illustration of texture-confusing person ReID. Human shape information is crucial when clothing texture is misleading. However, modeling shape upon 2D measurement could not capture the intact shape perception and collection of 3D source data relies on auxiliary devices. Single-view 3D human reconstruction could help to learn 3D shape feature in surveillance environments.

To extend the scalability of real-world person ReID, in this paper, we explicitly model discriminative clues beyond human clothing textures, i.e., human shape representations. Existing works try to learn shape-related features in two ways: *2D image space* [7, 34, 44] and *3D source data* [21, 31, 41]. 2D-based methods mainly make attempt to extract shape feature based on visual statistics, such as contour [44] and keypoint [34], or via adversarial feature disentanglement [21]. These methods only utilize the structure and shape information in 2D space while 3D information like

# Equal Contribution.
* Corresponding Author.

depth or relative 3D position is ignored. 3D-based source data could be collected from kinect cameras [41] or ratio signals [7], which has the potential to capture the integrated shape from an all-around horizon. However, collecting 3D data might be infeasible in a surveillance environment.

In order to learn 3D shape representation without extra 3D devices, we propose a novel feature learning schema combining 3D human reconstruction from a single image [1, 2, 4, 5, 14, 18, 32]. Instead of extracting ReID features from imprecise reconstructed 3D meshes [50], we train a ReID model that extracts texture-insensitive 3D features directly from the original 2D images by adding 3D reconstruction as an auxiliary task and regularization to the ReID feature learning. The 3D reconstruction based regularization is able to force the ReID model to decouple the 3D shape information from the visual texture, and acquire discriminative 3D shape ReID features that are more reliable for texture-confusing persons. In practice, a multi-task framework is adopted, and ReID feature is supervised by both identification losses (e.g., softmax loss and triplet loss) and 3D reconstruction losses.

One of the troublesome obstacles for training 3D human reconstruction lies in the lack of 3D ground truth. To overcome the data limitation, following the literature of 3D reconstruction [14, 32], we design a purely unsupervised framework called Adversarial Self-Supervised Projection (**ASSP**). We first utilize external unlabeled 3D data [26] to train a discriminator distinguishing the reconstruction results from real 3D parameters in an adversarial way. This could prevent abnormal poses and shapes in a coarse level. Then, we introduce a self-supervised learning loop that reprojects 3D reconstruction results back to the 2D plane and minimize the reconstruction error compared with 2D observations (e.g., keypoints and silhouettes). This could further fit personalized 3D bodies in a fine level.

3D human reconstruction tends to obtain a mean shape representation, and thus a global 3D shape feature is not discriminative enough. To enhance the discriminative ability of ReID, we propose the Multi-Granularity Shape feature (**MGS**) learning to combine both global and part shape features. In **MGS**, the global 3D shape feature corresponds to the global shape parameter estimation and part 3D shape features are used to estimate subtle local shape displacements. This could help to capture 3D shape features in different scopes and enrich the diversity of features.

We summarize our contributions as follows:

- We propose a novel end-to-end architecture combining person ReID and 3D human reconstruction to learn texture-insensitive 3D shape embedding. We further propose a multi-granularity shape (**MGS**) learning to enhance discriminative ability for person ReID.

- To address the problem of lacking 3D ground truth, we design the Adversarial Self-supervised Projection (**ASSP**) combining adversarial learning and self-supervised projection, validating that 3D reconstruction is capable to promote ReID in a unified training schema even without 3D ground truth.

The experimental results in common person ReID datasets (Market1501 [48], DukeMTMC-ReID [51]) and texture-confusing datasets (PRCC [44], VC-Clothes [39], LTCC [34], FGPR [46]) have demonstrated the effectiveness of the proposed model.

2. Related Work

2.1. Person ReID

Person ReID has been well advanced [9, 22, 29, 42, 45]. However, some works [7, 34, 44, 46] have argued that most person ReID methods including hand-crafted methods [11, 20, 38] and deep models [22, 29] degenerate a lot in performance when clothing texture is misleading due to the lack of the specifically designed mechanisms.

In this paper, we target at clothing texture-confusing ReID situations where the same identity might change clothing or different identities might wear similar clothing. There are several directions in the ReID literature learning texture-insensitive representation. Attribute-based methods [23, 36] overcome texture bias to some extent. Yu et al. utilize the given description to perform retrieval from a database of predefined clothing templates [47]. However, the above methods require auxiliary annotation and cost extra labour. Another routine devotes to directly capturing identity-invariant shape-related representation beyond clothing texture [7, 21, 31, 34, 41, 44]. Existing methods rely on either *2D image space* [7, 34, 44] or *3D source data* [21, 31, 41] to extract shape features. The former seeks to model a shape representation based on visual characteristics [34, 44] or via adversarial feature disentanglement [7]. For example, Yang et al. transform contours in polar coordinates for shape learning [44]. This kind of approaches could only capture partial shape representation, limited by viewpoint change and pose variation. The latter focuses on directly characterizing shape concept based on 3D source data, like the depth map [41], ratio frequency [21] and 3D skeletons [31]. Although 3D source data could reflect the full-view shape representation, it is usually hard to collect such data in real applications. In this paper, we utilize the invariance of 3D shape to break the limitations of 2D-based shape and get rid of extra 3D devices because what we use is only a single image.

2.2. Single-view 3D Human Reconstruction

Many methods leverage parametric models, such as SCAPE [3] and SMPL [25], to digitize 3D human representations. Most recent methods estimate pose and shape parameters of the SMPL model under the supervision of 3D ground truth [5, 14, 19, 30, 32]. To cap-

ture finer details, models [1, 2] extend the basic SMPL to "SMPL+Displacement".

Some works have made attempt to perform person ReID with the help of single-view 3D human reconstruction and the SMPL model. Zheng et al. directly use the reconstructed 3D meshes as inputs to perform ReID [50]. Li et al. rely on the SMPL model to synthesize 3D data for ReID [21]. Since the extracted 3D reconstruction meshes are not precise enough, such processing might lead to corrupted feature learning. Different from the above works, we integrate ReID and 3D reconstruction in a unified end-to-end training framework, which could adaptively learn more robust 3D shape features and reduce information loss.

3. Method

3.1. Overview

In this paper, we propose a novel framework to solve clothing texture-confusing person ReID, which can be specified as two cases: (1) the same identity changes clothing [44]; (2) different identities wear the same uniforms [46].

Our main contribution lies in the 3D shape learning (**3DSL**) branch, which combines person ReID and human 3D reconstruction in an end-to-end training framework for the first time. The goal of **3DSL** is to learn a 3D shape feature which could not only distinguish different identities but also estimate shape-related parameters of the 3D model SMPL [25]. Specifically, we train a sequence of deep networks E_{3D} and E_{shape} that extract a 3D shape feature from an image, denoted as F_{shape}. As shown in Figure 2, apart from supervised by ReID losses (i.e., softmax and triplet losses), F_{shape} is also the input of the 3D reconstruction sub-network that predicts the shape parameters of the 3D human model. In this way, the 3D human reconstruction task is added as an auxiliary regularizer to force F_{shape} to focus on 3D shape information.

Furthermore, to overcome the lack of 3D ground truth (e.g., 3D skeletons, 3D point clouds), similar to the literature of human reconstruction [14, 32], the 3D reconstruction branch is trained in a self-supervised framework, called Adversarial Self-Supervised Projection (**ASSP**). As shown in Figure 2, two kinds of supervisions are conducted in **ASSP**. Firstly, the discriminator D trained on extra unlabeled 3D parameters is used to distinguish 3D SMPL parameters estimated based on F_{shape} from real 3D parameters. Secondly, **ASSP** re-projects the reconstructed 3D meshes to the 2D plane and computes the 2D reconstruction error with 2D keypoints and silhouettes obtained from the original RGB image.

Since some RGB features like faces and attributes could also be texture-insensitive, we construct an extra network branch to learn these useful RGB features. Here, we propose a sampling strategy for the triplet loss [12] specifically designed for different clothing texture-confusing ReID tasks. Finally, the texture-insensitive 3D shape feature F_{shape} and RGB feature F_{rgb} are pooled and concatenated to form the final ReID feature.

3.2. 3D Shape Learning

The branch of 3D shape learning (**3DSL**) essentially trains person ReID and 3D human reconstruction in an end-to-end network.

3.2.1 3D Parametric Model

We choose the parametric 3D model SMPL [25] as a base model to carry out human reconstruction. Thanks to the SMPL's prior manifold, extra prior knowledge on 3D body shape can be transferred into the ReID model and the integrity of the reconstructed results will be better attained even when ground truth is unavailable. Moreover, different groups of parameters in SMPL contain specific semantics (i.e., shape-related, pose-related). It helps us to extract specific features for each group of parameters and decouple identity relevant shape features from identity irrelevant pose ones. SMPL is modeled as a function of the pose parameter $\theta \in \mathbb{R}^{24\times3}$ and the shape parameter $\beta \in \mathbb{R}^{10}$ returning $N_V = 6890$ vertices and $N_F = 13776$ faces. However, the shape parameter has only 10 dimensions and does not have enough capacity to construct a discriminative 3D model to represent diverse human shapes. Hence, we introduce vertice-wise displacement values denoted as $\delta \in \mathbb{R}^{6890\times3}$ into the 3D modeling for accommodating to the ReID learning:

$$\mathcal{M}(\beta, \theta, \delta) = W(T(\beta, \theta, \delta), J(\beta), \theta, \mathbb{W}), \quad (1)$$

where W is a linear blend-skinning function applied to the rest pose $T(\beta, \theta, \delta)$ and the skeleton joints $J(\beta)$. Please refer to [25] for the detailed implementation of W.

Denoting the 3D ReID shape feature as F_{shape}, in our method the SMPL's shape parameter θ and displacement δ are estimated with a sub-network from F_{shape}. In this way, we could also make F_{shape} decouple with pose interference and become pose-invariant.

3.2.2 3D Shape Feature Extraction

In this section we introduce how to use 3D human reconstruction to facilitate extracting 3D shape ReID features.

As shown in the green branch in Figure 2, the general feature containing all 3D information is extracted by a base network E_{3D}. The output of E_{3D} is then fed into the 3D reconstruction network. There are two groups of 3D model parameters: shape irrelevant parameters (i.e., pose parameters θ, camera parameters ψ) and shape relevant parameters (i.e., shape parameters β, vertice-wise displacements δ). 3D reconstruction models [14] predict different groups

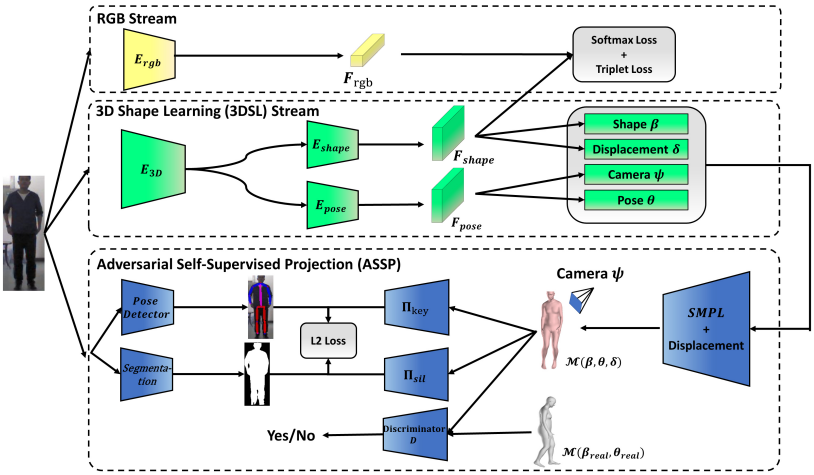


Figure 2. The overview of the proposed model. The model consists of two branches. One is for learning 3D shape features under the regularization of 3D human reconstruction, named 3D Shape Learning (**3DSL**, see Section 3.2). The another branch is for learning texture-insensitive RGB features from original images via metric learning, which is introduced in Section 3.3.

of parameters as a holistic distribution, which would hinder the extraction of identity-specific features since shape irrelevant information is an interference factor for ReID. In order to decouple the information of different parameter groups, separate estimation sub-networks E_{pose} and E_{shape} are adopted. The output feature map of E_{shape} contains all the 3D shape related information, which is exactly what we need for the 3D shape ReID feature, denoted as F_{shape}.

Besides receiving supervision from 3D reconstruction, F_{shape} is also supervised by ReID losses including the softmax and triplet losses. In this way, F_{shape} is both 3D shape-related and identity discriminative.

3.2.3 Adversarial Self-Supervised Projection

In the literature of 3D human reconstruction, training a model commonly requires high-quality 3D ground truth (e.g., 3D keypoints, 3D model parameters, 3D point clouds). However, in the surveillance videos, we do not have any 3D annotations for training. Following the literature of 3D reconstruction [14, 32], we design a purely unsupervised pipeline called adversarial self-supervised projection (**ASSP**). As shown in the blue branch in Figure 2, ASSP combines adversarial learning [10] and self-supervised projection from 3D to 2D, which train 3D reconstruction in coarse and fine levels, respectively.

Adversarial Learning for 3D reconstruction. The goal of adversarial learning is to generate reasonable 3D human models and avoid abnormal poses and shapes at a coarse level. Specifically, we train a discriminator network D (as shown in Figure 2) to distinguish the 3D reconstruction parameters estimated by E_{3D}, E_{pose} and E_{shape} from the extra real 3D human parameters.

Following [14], we transform the 3-dim rotation vector of each joint into 3×3 rotation matrix via the *Rodrigues formula*. That is to say, the pose parameter $\theta \in \mathbb{R}^{24\times3}$ would be transformed to $\mathbb{R}^{24\times3\times3}$. Then we put the transformed pose parameter and β as the input of the discriminator D, where the architecture of D is the same as [14]. A large-scale dataset of SMPL parameters [26] is used as the real human body data, denoted as θ_{real} and β_{real}. The adversarial loss for the discriminator D could be formulated as:

$$\mathcal{L}_{adv}(D) = \mathbb{E}[(1 - D(\theta_{real}, \beta_{real}))^2] + \mathbb{E}[D(\theta, \beta)^2], \tag{2}$$

The adversarial loss for the estimation network E_{3D}, E_{pose} and E_{shape} is defined as:

$$\mathcal{L}_{adv}(E_{3D}, E_{pose}, E_{shape}) = \mathbb{E}[(1 - D(\theta, \beta))^2], \tag{3}$$

Self-Supervised Projection from 3D to 2D. The goal of self-supervised projection is to reconstruct 3D meshes that fit to the original 2D image at a fine level. This is done by projecting the estimated 3D meshes back to the original 2D plane, making the projections consistent with 2D observations (e.g., keypoints, silhouettes) predicted from original input images. In this way, **3DSL** is trained under self-generated supervision signals in an end-to-end loop, which has been widely applied [5, 14, 19, 30, 32].

Here we choose keypoints and silhouettes as intermediary to bridge 2D and 3D spaces. We utilize the off-the-shelf detector [8] to predict keypoint locations $K \in \mathbb{R}^{P\times2}$ from original input images. For silhouettes M, we follow the processing in [28] and apply GrabCut [35] for prediction. Since the projection from the 3D space to 2D image space requires the 3D position of the camera, we simultaneously estimate the camera position parameter $\psi \in \mathbb{R}^3$. We adopt the same camera model and parameters as [14]. For the keypoint projection from 3D to 2D, it is a sparse mapping through a projection matrix derived from ψ and the camera

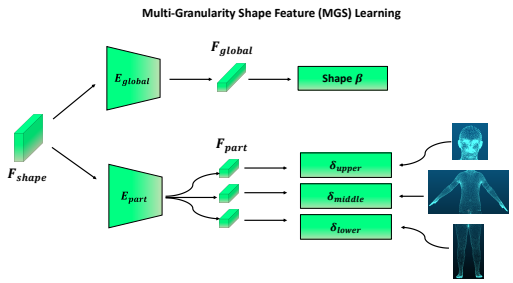


Figure 3. The illustration of Multi-Granularity Shape (**MGS**) learning. In **MGS**, there is a global shape feature F_{global} to estimate the global shape parameter β and part shape features F_{part} to estimate part displacements. We partition the 3D vertices into typically 3 parts. E_{global} and E_{part} both consists of shallow convolution and fully-connected layers.

model. The keypoint projection loss is defined as:

$$\mathcal{L}_{key} = \left\| K - \hat{K} \right\|_2^2, \tag{4}$$

where $\hat{K} = \Pi_{key}(\beta, \theta, \delta, \psi)$ is the projected keypoints from the reconstructed 3D meshes as shown in Figure 2.

The silhouette projection should be carried out with the help of a differentiable renderer to make it end-to-end training. We choose the neural renderer [15]. The silhouette projection could be expressed as: $\hat{M} = \Pi_{sil}(\beta, \theta, \delta, \psi)$. The silhouette projection loss is formulated as:

$$\mathcal{L}_{sil} = \left\| M - \hat{M} \right\|_2^2 + \|\delta\|_2, \tag{5}$$

where we restrict the value of δ to avoid recovering clothing details and keep the reconstructed meshes smooth.

3.2.4 Multi-Granularity Shape Feature

In order to enhance the ReID discriminative ability, we introduce the part-based ReID paradigm and propose Multi-granularity Shape (**MGS**) feature learning, which extracts local 3D shape features from different parts of human bodies. The **MGS** feature is used as the input of a sub-network to predict the vertice-wise displacements for a body part, which forces this feature to contain local shape information.

Specifically, as shown in Figure 3, we partition the vertices of the 3D SMPL model into P parts, and each part corresponds to displacements of a sub-group of vertices, denoted as $\delta_i \in \mathbb{R}^{p_i \times 3}$, where p_i is the vertice number of i-th part. Taking F_{shape} as input, several shallow sub-networks are applied to predict the global shape feature F_{global} and part shape feature F_{part}, respectively. F_{global} is responsible to estimate the global shape parameter β. F_{part} is first partitioned into several stripes like the part-based model [38] and each stripe F_{part}^i is responsible for predicting vertice-wise displacements δ_i of the i-th part. Both F_{global} and F_{part} would be supervised by the losses of person ReID.

3.3. Texture-insensitive RGB Features

Besides 3D shape features, there are some other useful texture-insensitive RGB features, which also play an important role in clothing texture-confusing situations, such as face features and some other local attributes (e.g., glasses).

An extra network branch (E_{rgb} in Figure 2) is constructed to excavate valuable information mentioned above. To force this branch to pay more attention to areas irrelevant to clothing texture, we adaptively modify the sampling strategy of triplet loss [12] via provided clothing information [34, 39, 44, 46], based on the characteristics of specific tasks. For datasets with clothing change, images with the same identity but different clothes are selected as positive pairs. For the situation that different identities wear similar uniforms, images wearing similar clothes will be selected as negative pairs

Our two-branch model could also be adapted to solve the common person ReID problem [48, 51]. In this situation, positive and negative pairs are randomly sampled to train E_{rgb}. The 3D shape learning branch could serve as supplement of texture information because in this situation texture is more powerful.

4. Experiment

4.1. Datasets

We conduct experiments on 4 clothing texture-confusing person ReID datasets, i.e., PRCC [44], VC-Clothes [39], LTCC [34], FGPR [46], and 2 common datasets, i.e., Market1501 [48], DukeMTMC-ReID [51], which demonstrates the effectiveness of our model in different situations.

Clothing texture-confusing person ReID datasets. There are two types of clothing texture-confusing benchmarks. The first type of datasets are collected for evaluating the performance when the same identity would change clothing, such as PRCC [44], LTCC [34], VC-Clothes [39]. The PRCC dataset is captured under 3 disjoint camera views and samples of identities dressed in different clothing are collected under different camera views. There are 33698 images in the PRCC, with 150 identities in the training set and 71 identities in the testing set. The LTCC contains 17138 images of 152 identities. In the training set with 77 identities, 46 people appears in different clothing and the other 31 people do not change clothing while the testing set consists of 45 clothing-change identities and 30 clothing-consistent ones. The VC-Clothes dataset is a virtual benchmark synthesized by game engines under 4 camera views. VC-Clothes has 9449 images of 256 identities in the training set and 9611 images of 256 identities in the testing set. The second type of datasets are used for solving the problem of different identities wearing similar clothes, e.g., FGPR [46]. There are 115106 images and 245 identities, which are split into "blue" and "white" groups. 10 train/test splits are conducted and for each split, 150 identities are divided

Figure 4. Samples of datasets.

for training and 95 identities for testing. We use the video-based setting for both training and testing and we apply average pooling on features obtained on every sequence.

Common person ReID datasets. We utilize two large-scale benchmarks to validate that our method also achieves comparable performance on common person ReID situations. Market1501 [48] includes 1501 identities and 32688 images collected in 6 non-overlapping cameras. 12936 images of 751 identities form the training set and the other 750 identities form the query set (3368 images) and gallery set (19734 images). DukeMTMC-ReID [51] covers 8 disjoint camera views. There are 702 identities and 16522 images for training, while the testing set contains 702 identities, with 2228 query samples and 16522 gallery samples.

We apply the evaluation protocol of the PRCC dataset the same as [44] and that of the FGPR dataset the same as [46]. Notably, the evaluation of the above datasets is single-shot, so we only report the *cumulative match characteristic (CMC)* curve. For other datasets, we use the *CMC* curve and *mean average precision (mAP)* for evaluation. For PRCC [44], LTCC [34], VC-Clothes [39] and FGPR [46], we only report the performance involving clothing texture-confusing situations as the original papers.

4.2. Implementation details.

We leverage ResNet50 [11] for E_{rgb}. E_{3D} is the part of ResNet50 [11] before *res_conv4*. E_{pose} and E_{shape} both consists of the sub-networks *res_conv4* & *conv5* in ResNet50. E_{global} and E_{part} consists of two 1×1 convolution layers, a global average pooling layer and two fully-connected layers. We resize images to 256×128 for training and testing. The batch size is set as 64 with the number of identities $T = 16$ and the sample number of each identities $S = 4$. The optimizer is Adam [16]. The total epoch is set as 120. The initial learning rates of E_{3D} and E_{pose} are set as 0.0001 while those of E_{rgb}, E_{shape}, E_{global}, E_{part} and D are set as 0.0008. Learning rates would be decayed by 10 after 40 and 90 epochs. The weight decay is set as 0.0005.

4.3. Comparison with state-of-the-art

We compare our model with the state-of-the-art ReID methods separately on clothing texture-confusing and common ReID datasets, including: **(1)** *hand-crafted* feature representations, e.g., LOMO [22], GOG [27] and *metric learning*, e.g., XQDA [22], KISSME [17]; **(2)** state-of-the-art deep models on common ReID datasets (image-based or video-based), e.g., MGN [40], PCB [38], AGRL [43]; **(3)** state-of-the-art deep models on clothing texture-confusing ReID datasets, e.g., SPT+ASE [44], SE+CESD [34]. The *Baseline* model in our experiment is a plain two-branch model that combines two original ResNet50 networks, which are trained separately and features are concatenated for evaluation.

From Table 1, Table 2 and Table 3, we could observe that our model significantly exceeds other competitors on clothing texture-confusing situations. For example, our model achieves a rank-1/mAP improvement of more than **10.5%/13.9%** over those models on VC-Clothes. For FGPR where different identities might wear similarly, our model outperforms other methods by about **2.5%** in rank-1. Simultaneously, our model achieves comparable performance on Market-1501 and DukeMTMC-ReID.

Comparison on clothing texture-confusing datasets. As shown in Table 1, on datasets that the same identity might change clothing, part-based models like MGN [40] achieve relatively better performance than the basic ReID feature [11] by considering local ReID features, but still could not capture effective clues for ReID. The 2D shape-based methods (e.g., SE+CESD [34]) exceed the RGB based model on some of the clothing change datasets while they are limited by the ambiguity of 2D shape representations. Compared with the above competitors, our method achieves the highest performance on all three clothing changing datasets thanks to the modeling of effective 3D shape embedding.

Table 2 shows the performance comparison on the dataset that different identities would wear similar clothing. Note that FGPR is a video based ReID dataset and we make comparison with video-based ReID methods. Our method still archives the highest performance in terms of rank-1. This verifies that 3D shape embedding has not only inherent invariance to identify the same person but also better discriminability to distinguish different identities. The **MGS** emphasizes on both global and local 3D shapes, and captures shape differences in different granularities.

Comparison on common datasets. On common ReID datasets, person clothing texture is the most crucial clue. As shown in Table 3, although our method does not put much effort on improving the RGB based feature, our method still achieves comparable results with the state-of-the-art common ReID models. The experimental results reveal that the 3D shape feature embedding could help to improve the performance of the baseline method. Additionally, we could observe that the performance of state-of-the-art ReID models on the cloth-confusing benchmark is extremely unstable. For example, MGN [40] achieves 95.7% rank-1 in Market1501 while 47.2% in PRCC. In comparison, our model

Table 1. Performance on clothing change person ReID datasets. The best and second best results are indicated by Red and Blue, respectively. "†" indicates that we carry out experiments with open codes by ourselves. Performance is measured by %.

Model	PRCC		Model	LTCC		Model	VC-Clothes (cam3&cam4)	
	rank-1	rank-10		rank-1	mAP		rank-1	mAP
LOMO[22]+XQDA[22]	14.5	43.6	LOMO[22]+XQDA[22]	11.0	5.6	LOMO[22]+XQDA[22]	34.5	30.9
LOMO[22]+KISSME[17]	18.6	49.8	LOMO[22]+KISSME[17]	11.0	5.3	GOG[27]+XQDA[22]	35.7	31.3
ResNet [11]†	44.8	81.2	ResNet [11]	20.1	9.0	ResNet [11]	36.4	32.4
PCB [38]†	45.6	82.3	PCB [38]	23.5	10.0	PCB [38]	62.0	62.2
MGN [39]†	47.2	84.3	OSNet [52]	24.0	10.8	MDLA [33]	59.2	60.8
SPT+ASE [44]	34.4	77.3	SE+CESD [34]	26.2	12.4	Part-aligned [37]	69.4	67.3
Baseline	45.6	83.0	Baseline	25.0	9.7	Baseline	73.6	71.5
Our Model	51.3	86.5	Our Model	31.2	14.8	Our Model	79.9	81.2

Table 2. Performance (%) results of our method and other compared methods on FGPR. The "All groups", "Blue group" and "White group" settings are the same as [46]. "†" indicates that we carry out experiments with open codes by ourselves. "∗" indicates the model performance reported in the original paper [46] where the dataset used is different with the released one. We list it here just for reference.

Models	All groups		Blue group		White group	
	rank-1	rank-5	rank-1	rank-5	rank-1	rank-5
LOMO[22]+XQDA[22]†	20.6	35.1	22.1	38.5	24.1	40.0
GOG[27]+XQDA[22]†	22.6	37.1	23.5	37.9	26.5	42.2
ResNet[11]†	83.0	92.0	84.2	94.0	85.1	94.7
STMP[24]†	85.0	93.4	86.2	94.2	87.5	96.4
AGRL[43]†	85.5	94.5	87.9	96.7	86.5	94.9
FGPR[46]∗	87.1	95.2	93.6	97.2	99.0	100.0
Baseline	84.6	93.2	85.8	93.8	86.0	94.2
Ours	88.0	95.0	88.3	96.0	88.9	95.9

Table 3. Performance (%) results of our method and other compared methods on Market1501 (under the single-query setting) and DukeMTMC-ReID.

Models	Reference	Market1501		DukeMTMC-ReID	
		rank-1	mAP	rank-1	mAP
HA-CNN [20]	CVPR 2018	91.2	75.7	80.5	63.8
PCB [38]	ECCV 2018	92.3	77.4	81.8	66.1
MGN [40]	ACMMM 2018	95.7	86.9	88.7	78.4
DGNet [49]	CVPR 2019	94.8	86.0	86.6	74.3
MHN [6]	ICCV 2019	93.6	83.6	87.5	75.2
OSNet [52]	ICCV 2019	94.8	86.0	88.6	73.5
SAN [13]	AAAI 2020	96.1	88.0	87.9	75.5
Baseline		94.7	84.8	86.8	74.0
Ours		95.0	87.3	88.2	76.1

adaptively captures optimal features in different situations and achieves more favourable trade-offs.

4.4. Ablation Study

In ablation study, we carry out experiments to demonstrate: (1) the effectiveness of 3D shape features and the combination with texture-insensitive RGB features; (2) the effectiveness of different components of learning 3D shape features including losses in **ASSP** and **MGS**. Please refer to Section 4.3 for the implementation of the baseline.

Comparison with 3D shape features and RGB Features. As shown in Table 4, two types of RGB features are trained in this ablation study, i.e., RGB features with random triplet sampling and RGB features with triple sampling described in Section 3.3. We observe that compared with the baseline, adding 3D shape learning achieves a 4.8% rank-1 improvement on PRCC and 2.5% mAP improvement on Market1501 in total. The texture-insensitive RGB feature learned with specifc triplet sampling strategy outperforms random sampling. Single 3D shape features outperform RGB features on PRCC but achieve worse performance on Market1501, demonstrating that clothing texture information is more important on common ReID datasets.

Table 4. Performance (%) comparison of combining different features in our methods. Rank-1, rank-5 and mAP are reported. "†" indicates evaluation with the single feature of one branch. Others are evaluated with concatenation of features from two branches.

3D Shape	RGB Feature		PRCC		Market1501	
	Random	Strategy	rank-1	rank-5	rank-1	mAP
×	√	×	45.6	83.0	94.7	84.8
√	√	×	50.4	85.7	95.0	87.3
√	×	√	51.3	86.5	-	-
×†	×	×	44.9	83.2	94.5	84.2
√†	×	√	49.2	84.2	93.2	83.7

Table 5. Performance (%) comparison of different components in **3DSL**. Rank-1, rank-5 and mAP are reported.

ASSP			**MGS**	PRCC		Market1501	
$\mathcal{L}_{adv}$	$\mathcal{L}_{key}$	$\mathcal{L}_{sil}$		rank-1	rank-5	rank-1	mAP
×	×	×	×	45.6	83.0	94.7	84.8
√	√	√	×	50.1	83.6	94.2	86.7
√	√	√	√	51.3	86.5	95.0	87.3
√	×	×	√	47.0	82.9	93.9	84.5
×	√	√	√	49.8	84.0	94.5	86.7
√	√	×	√	49.0	83.5	93.9	85.5
√	×	√	√	49.5	83.9	94.1	86.0

Comparison of components in 3DSL. The losses used in **ASSP** determine the accuracy of 3D reconstruction and whether we could capture the intrinsic 3D shape embedding. In Table 5, we observe that the version with only adversarial learning (i.e., $\mathcal{L}_{adv}$) achieves limited performance since this only ensures a coarse 3D model without sufficient discriminative ability. The introduction of self-supervised projection (i.e., $\mathcal{L}_{key} + \mathcal{L}_{sil}$) could bring remarkable improvement because it fits to fine-level body and captures distinguishable body shapes. The combination of all losses integrates multi-level reconstruction, and thus performs the best. Moreover, **MGS** gives 1.2% rank-1 improvement on PRCC and 0.6% mAP improvement on Market1501.

Figure 5. Rank list of the baseline and our model in PRCC [44]. Green boxes indicate matched samples while red ones indicate mismatched. Notably, evaluation of PRCC is single-shot, which means that there is only one matched sample in the gallery set.

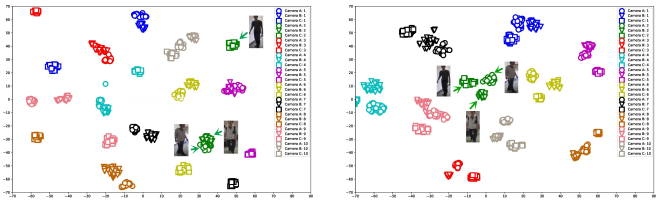
4.5. Further Analysis

Visualization of rank list. To further have an insight on how 3D shape feature makes sense in texture-confusing situations, we do some visualization work in this section. As shown in Figure 5, when applying the baseline model, the most similar identities are those who have similar clothing texture or color with the query identity, which typically illustrates the texture bias that we target at in this paper (e.g., biase to the red clothing pattern in the first line). On the other hand, as our model uses texture-insensitive 3D shape feature, we could capture the inherent invariance for each identity and overcome the distraction of clothing texture.

Visualization of feature distribution. We visualize feature distributions to better understand the effectiveness of the 3D shape feature. In Figure 6a, we could observe that for each identity, features of samples with the same clothing (circles and triangles) flock together while samples with different clothing (squares) keep away from the other two kinds of features. This phenomenon reveals that for general deep models, the main obstacle lies in large intra-class distance, which essentially originates from the excessive attention on clothing texture. In Figure 6b, the intra-class distance is obviously reduced, illustrating the invariance of 3D shape. Features of different identities also remain at certain distance mainly thanks to the discriminative ability of **3DSL**.

Visualization of 3D reconstruction. In Figure 7, we visualize 3D human reconstruction to evaluate the influence of coarse-level adversarial learning and fine-level self-supervised projection in **ASSP**. The results with only "$\mathcal{L}_{adv}$" could ensure valid human models based on extra auxiliary information but could not accurately fit to corresponding shapes and poses. In contrast, reconstructions with only "$\mathcal{L}_{key}+\mathcal{L}_{sil}$" might generate abnormal poses and shapes. For example, joint rotations of "ID 2" are unreasonable and the body size of "ID 3" is malformed. With the combination of coarse and fine loss constraints, the proposed **ASSP** could best fit to specific identities.

Notably, distinction of different body shapes could be reflected by the reconstruction results. For example, reconstructed under the same viewpoint, "ID 1" and "ID 3" in Figure 7 are obviously distinguishable in body size.



(a) Baseline (b) Our Model

Figure 6. t-SNE visualization of different feature distributions of randomly selected identities in PRCC [44]: (a) features of baseline model; (b) features of our model. Under Camera A (circles) and Camera B (triangles), the same identity would not change clothing while under Camera C (square) clothing change occurs.

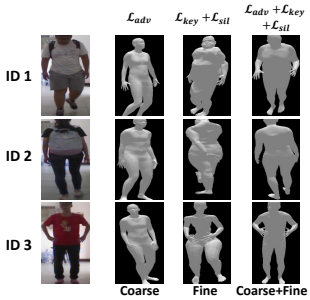


Figure 7. Visulizaiton of 3D human reconstruction under different losses in **ASSP**.

5. Conclusion

In this paper, we propose to learn a texture-insensitive 3D shape representation and demonstrate the effectiveness in the situations when clothing texture becomes confusing. Specifically, we propose a novel framework to capture 3D shape ReID features by combining person ReID and 3D human reconstruction in an end-to-end training manner. To solve the problem of lacking ground truth to train 3D reconstruction, we introduce an unsupervised module called Adversarial Self-Supervised Projection (**ASSP**) to ensure coarse body manifolds via adversarial learning and fit the fine body details via self-supervised projection from 3D to 2D. To enhance the discriminative ability of 3D shape features, we propose Multi-Granularity Shape (**MGS**) learning to capture part 3D shapes and increase the feature diversity. Experimental results on either clothing texture-confusing ReID benchmarks or common ReID benchmarks have illustrated the effectiveness of the proposed modules.

Acknowledgement

This work was supported partially by the NSFC(U1911401,U1811461), Guangdong NSF Project (No. 2020B1515120085, 2018B030312002), Guangzhou Research Project (201902010037), and Research Projects of Zhejiang Lab (No. 2019KD0AB03), and the Key-Area Research and Development Program of Guangzhou (202007030004).

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *3DV*, pages 98–109. IEEE, 2018. 2, 3

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, pages 8387–8397, 2018. 2, 3

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH*, pages 408–416. 2005. 2

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. *ECCV*, 2020. 2

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, and Peter Gehler. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, pages 561–578. Springer, 2016. 2, 4

[6] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *ICCV*, pages 371–381, 2019. 7

[7] Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. Learning longterm representations for person re-identification using radio signals. In *CVPR*, 2020. 1, 2

[8] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, pages 2334–2343, 2017. 4

[9] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8295–8302, 2019. 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 4

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 6, 7

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3, 5

[13] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *AAAI*, pages 11173–11180, 2020. 7

[14] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 2, 3, 4

[15] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, pages 3907–3916, 2018. 5

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 6

[17] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295. IEEE, 2012. 6, 7

[18] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, pages 2252–2261, 2019. 2

[19] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, pages 6050–6059, 2017. 2, 4

[20] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 2, 7

[21] Yu-Jhe Li, Zhengyi Luo, Xinshuo Weng, and Kris M Kitani. Learning shape representations for clothing variations in person re-identification. *arXiv preprint arXiv:2003.07340*, 2020. 1, 2, 3

[22] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015. 2, 6, 7

[23] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *PR*, 95:151–161, 2019. 2

[24] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, volume 33, pages 8786–8793, 2019. 7

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 3

[26] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019. 2, 4

[27] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, pages 1363–1372, 2016. 6, 7

[28] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *CVPR*, pages 7023–7034, 2020. 4

[29] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *PR*, 29(1):51–59, 1996. 2

[30] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 2, 4

[31] Pietro Pala, Lorenzo Seidenari, Stefano Berretti, and Alberto Del Bimbo. Enhanced skeleton and face 3d data for person re-identification from depth cameras. *Computers & Graphics*, 79:69–80, 2019. 1, 2

[32] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape

from a single color image. In *CVPR*, pages 459–468, 2018. 2, 3, 4

[33] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *ICCV*, pages 5399–5408, 2017. 7

[34] 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. *arXiv preprint arXiv:2005.12633*, 2020. 1, 2, 5, 6, 7

[35] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004. 4

[36] Arne Schumann and Rainer Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *CVPRW*, pages 20–28, 2017. 2

[37] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, pages 402–419, 2018. 7

[38] 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 480–496, 2018. 2, 5, 6, 7

[39] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPRW*, pages 830–831, 2020. 2, 5, 6, 7

[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*, pages 274–282, 2018. 6, 7

[41] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Robust depth-based person re-identification. *IEEE TIP*, 26(6):2588–2603, 2017. 1, 2

[42] Shangxuan Wu, Ying-Cong Chen, Xiang Li, An-Cong Wu, Jin-Jie You, and Wei-Shi Zheng. An enhanced deep feature representation for person re-identification. In *WACV*, pages 1–8. IEEE, 2016. 2

[43] Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, Qi Tian, and Xue Zhou. Adaptive graph representation learning for video person re-identification. *IEEE TIP*, 29:8821–8830, 2020. 6, 7

[44] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE TPAMI*, 2019. 1, 2, 3, 5, 6, 7, 8

[45] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *CVPR*, pages 1389–1398, 2019. 2

[46] Jiahang Yin, Ancong Wu, and Wei-Shi Zheng. Fine-grained person re-identification. *IJCV*, 2020. 1, 2, 3, 5, 6, 7

[47] 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. 2

[48] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *CVPR*, pages 1116–1124, 2015. 2, 5, 6

[49] 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. 7

[50] Zhedong Zheng and Yi Yang. Parameter-efficient person re-identification in the 3d space. *arXiv preprint arXiv:2006.04569*, 2020. 2, 3

[51] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *CVPR*, pages 3754–3762, 2017. 2, 5, 6

[52] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3702–3712, 2019. 7