GEFF: Improving Any Clothes-Changing Person ReID Model using Gallery Enrichment with Face Features

Daniel Arkushin*[1] Bar Cohen*[2] Shmuel Peleg[1] Ohad Fried[2]
[1]The Hebrew University of Jerusalem [2]Reichman University
http://www.vision.huji.ac.il/reface/
{daniel.arkushin, peleg}@mail.huji.ac.il bar.cohen01@post.runi.ac.il ofried@runi.ac.il

Abstract

*In the Clothes-Changing Re-Identification (CC-ReID) problem, given a query sample of a person, the goal is to determine the correct identity based on a labeled gallery in which the person appears in different clothes. Several models tackle this challenge by extracting clothes-independent features. However, the performance of these models is still lower for the clothes-changing setting compared to the same-clothes setting in which the person appears with the same clothes in the labeled gallery. As clothing-related features are often dominant features in the data, we propose a new process we call Gallery Enrichment, to utilize these features. In this process, we enrich the original gallery by adding to it query samples based on their face features, using an unsupervised algorithm. Additionally, we show that combining ReID and face feature extraction modules alongside an enriched gallery results in a more accurate ReID model, even for query samples with new outfits that do not include faces. Moreover, we claim that existing CC-ReID benchmarks do not fully represent real-world scenarios, and propose a new video CC-ReID dataset called **42Street**, based on a theater play that includes crowded scenes and numerous clothes changes. When applied to multiple ReID models, our method (GEFF) achieves an average improvement of 33.5% and 6.7% in the Top-1 clothes-changing metric on the PRCC and LTCC benchmarks. Combined with the latest ReID models, our method achieves new SOTA results on the PRCC, LTCC, CCVID, LaST and VC-Clothes benchmarks and the proposed **42Street** dataset.*

1. Introduction

Person re-identification (ReID) aims to match the same people appearing at different times and locations. Given samples of people of interest — commonly referred to as a *gallery*, and unlabeled query samples, the goal is to predict the correct label (*i.e*. person ID) for every query sample based on the given gallery. Existing ReID models tend to perform poorly when re-identifying the same people over a prolonged time due to appearance changes such as different clothes and hairstyles [49]. Moreover, the performance of models that try to extract clothes-independent features such as body shape [5], contours [27,48], or gait [33,52], is subpar compared to same-clothes settings, as clothes are often the most dominant features [16,49]. To address the clothes-changing problem we introduce a simple process which we refer to as *Gallery Enrichment*. In this process, we use the gallery data to automatically add to it parts of the query data where people appear in different outfits. Extending the gallery in this manner results in an enriched gallery that increases the chances of finding a correct match. This is done by an unsupervised algorithm that uses the similarity between the faces in the query and gallery samples.

As face features provide an accurate prediction for the query samples that include faces, this algorithm results in an enriched gallery with minimal errors (Fig. 1). Once enriched, the gallery allows the ReID model, which relies on appearance-related features, to correctly predict the identity of a person with previously unseen outfits, even if the query sample does not include a face. In addition to using faces in the gallery enrichment process, we claim that integrating a face feature extraction module during inference is beneficial for the results of ReID, and introduce a new method that combines pre-trained face features extraction and ReID modules alongside an enriched gallery. We call this method GEFF — Gallery Enrichment with Face Features.

We claim that current video CC-ReID benchmarks do not include enough cases of occlusions, various illumination conditions, and multiple clothes and hairstyle changes. Therefore, we introduce the ***42Street*** dataset, curated from a theater play, as many theater plays include these challenges.

Extensive experiments show that GEFF improves the performance of the evaluated ReID models, on 5 CC-ReID benchmarks and the ***42Street*** dataset.

*These authors contributed equally

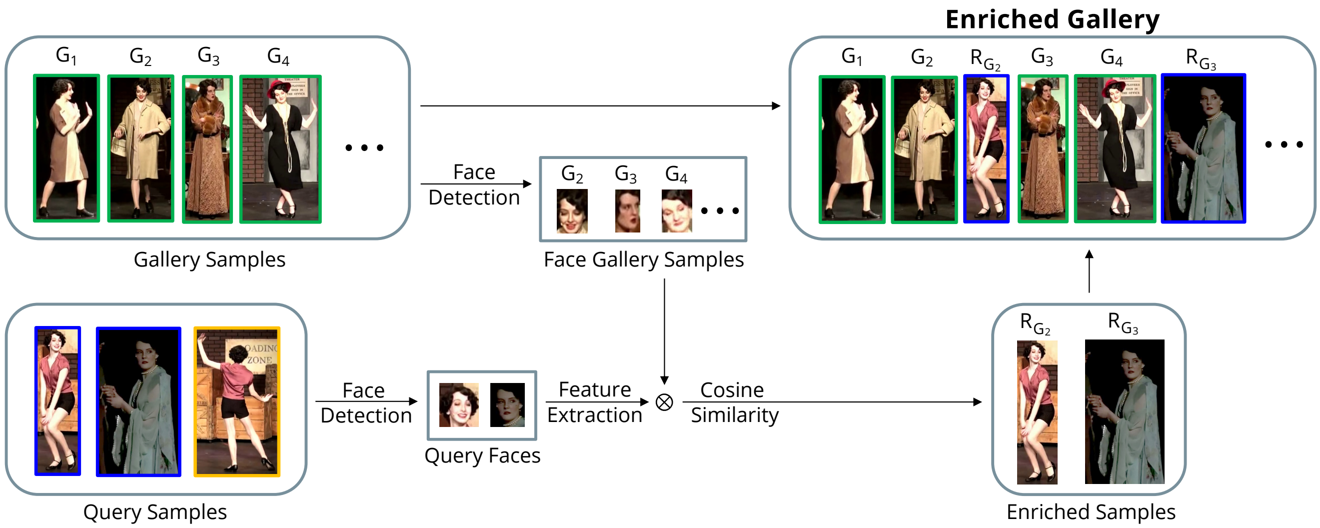

Figure 1. **The Gallery Enrichment Process.** For every gallery sample in which a face was detected, a face feature vector is extracted to create a face gallery. Then, for every query sample in which a face was detected, a reference to the closest sample in the gallery is saved, based on face similarity. Finally, an enriched gallery is created by extending the original gallery with the enriched samples. Colored frames were added for visualization purposes; green indicates a gallery sample and blue a query sample that is added to the enriched gallery. Notice that query samples that do not include faces (yellow frame) will not be added to the gallery. However, as an outcome of the enrichment process, during the ReID inference (not illustrated) it will be more likely to find a correct match for such queries.

2. Related Work

2.1. Person Re-Identification

The common inference process of person ReID can be seen as an instance retrieval problem. Given gallery and query samples, the goal is to classify each query sample correctly. First, feature vectors for all gallery samples are extracted by applying a feature-extraction model. Next, given an unseen query sample, the distances between its feature vector and the gallery feature vectors are computed. Finally, the predicted label of the query sample is defined as the label of the gallery feature vector that is closest to the query.

Image-based and Video-based ReID In image-based ReID datasets, every data sample is a single-person crop. Over the years, multiple image-based models have been developed [20, 29, 30, 46], and they achieve impressive results on same-clothes image-based benchmarks. In video-based ReID datasets, every data sample is a sequence of crops, *i.e.* a track, and the video ReID model produces a single feature vector to represent the entire sequence by using the spatio-temporal information in the sequence [13, 16, 21, 33].

Clothes-changing ReID In this setting, a person in the query set might wear different clothes from the gallery set. Some models try to extract clothes-independent features by modeling body shape and gait using skeletons [35], silhouettes [5], and contour sketches [48]. M2Net [27] uses contour images and human-parsing images to extract meaningful features. CAL [16], proposes a Clothes-based Adversarial Loss to mine clothes-independent features, and uses the video input to extract spatio-temporal patterns. AIM [49] utilizes a causality-based auto-intervention model to mitigate clothing bias and CCFA [19] implicitly augments clothes-changing data in the feature space. A concurrent work, IGCL [14], applies vision transformers to provide direct supervision to learn identity-specific features. Since the accuracy of these models is lower under the clothes-changing setting compared to the same-clothes setting, we suggest a process that partially converts the clothes-changing setting into the same-clothes setting, by building an enriched gallery and using face features during inference.

2.2. Face Feature Extraction

Face feature extraction is the process of detecting and identifying specific facial features from images or videos. Early works in the field such as Viola-Jones [41] and Local Binary Pattern [1] laid the foundations for more recent methods such as Face Attention Network [43] and the Insightface [9–12] library, which use deep learning to extract facial features and surpass human performance [44].

Using Face Features for ReID Several studies attempted utilizing face features for the task of person ReID, using various deep learning techniques [3, 15, 24, 25]. While these works try to predict an identity based on face features only, as we show in our work, face features are insufficient on their own since not all query samples contain faces. Therefore, we propose a model that combines both face and ReID modules. Another work, 3APF [42], combines a holistic feature extractor (ReID part) and a local face feature extractor (face part) on the feature vectors space. In our method, we propose to create a score vector for each model separately and then combine them into a final score vector. Additionally, we use face features to build the enriched gallery. In our experiments, we show that we outperform 3APF on VC-Clothes [42], the dataset they publish.

3. Method

In order to address the clothes-changing ReID problem, our method enriches the gallery using an unsupervised process (Sec. 3.1) and combines a pre-trained ReID module together with a pre-trained face feature extraction module (Sec. 3.2). Both elements work in conjunction and define our method which can be applied to any ReID model and works with image-based and video-based settings.

3.1. Creating an Enriched Gallery

The objective of most ReID models is to produce a feature extraction function that generates a feature vector for each data sample. Given two different data samples of the same person, the feature extraction function is expected to generate feature vectors with a lower distance compared to two samples of different persons. The richer the gallery is with samples that are similar to the query set, the more likely it is to find a correct gallery match. Hence, we propose the following unsupervised algorithm to enrich a given gallery from the query using face features, when available (Fig. 1). Given gallery and query samples:

1. We first apply a face detector on all gallery samples. We then build a face gallery by applying a face feature extractor on every sample in which a face was detected.

2. For every query sample in which a face was detected, we save a reference to the most similar gallery sample by computing its face feature vector and comparing it to the face gallery from Step 1 using cosine-similarity.

3. We create an enriched gallery by extending the original gallery with the queries from Step 2.

During evaluation, the references to the original gallery samples are used to determine the predicted identity of a given query sample.

3.1.1 Matching Face to Pose Estimation

A person crop is a part of an image that aims to capture a single individual. However, in crowded scenes multiple people might appear in the background, making it difficult to determine which face belongs to the main person in the crop. Therefore, when predicting an identity by using face features, it is crucial to verify that the detected face indeed belongs to the targeted individual. To achieve this, we utilize a pose estimator [7] (which we limit to provide a single pose estimation) to confirm that the eyes and nose keypoints of the main person in the crop fall within the face bounding box. Therefore, for datasets curated from crowded scenes (like the proposed ***42Street*** dataset), Steps 1 and 2 should include an additional step of face-to-pose matching. Examples are shown in Fig. 2.

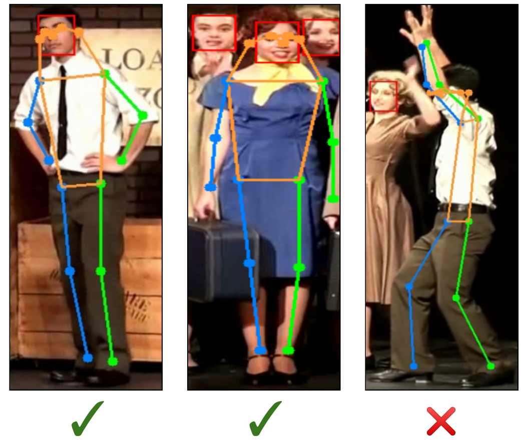

Figure 2. **Matching Face Detection (red bounding box) and Pose Estimation (colored skeleton)** *Left:* Detected face matches the pose estimation. *Middle:* Irrelevant detected faces are ignored as pose estimation matches a single face. *Right:* Detected face does not match the pose estimation.

3.2. Combining ReID and Face Modules

Following is a detailed description of our method for the video-based setting, where every data sample is a person track. The prediction process for the image-based setting is treated as a special case of the video-based setting, in which a data sample, *i.e.* a single image, is a track of length 1. In our method we propose to use a face feature extraction module and a ReID module to compute face and ReID score vectors respectively. These score vectors represent the confidence of each module that the given track belongs to each of the possible identities. Our method combines these two score vectors into a final score vector which is used to predict the identity of the person in a given data sample.

Predicting an Identity of a Track Given a gallery with a set of identities I, we first build an enriched gallery, $G_{enriched}$, as described in Sec. 3.1, and a face gallery denoted G_{face}, using face features extracted in the gallery enrichment process. Then, we use the ReID and face feature extraction modules on the track to create score vectors of size $|I|$, $v_{ReID} \in \mathbb{R}^{|I|}$ and $v_{face} \in \mathbb{R}^{|I|}$, respectively, which represent the confidence of each module that the given track belongs to each identity in I. This process is inspired by works such as CTL [46] and MCTL [2] that use the identity of each sample during inference to calculate centroids. Finally, we combine v_{ReID} and v_{face} into a single score vector. The process is illustrated in Fig. 3 and detailed next.

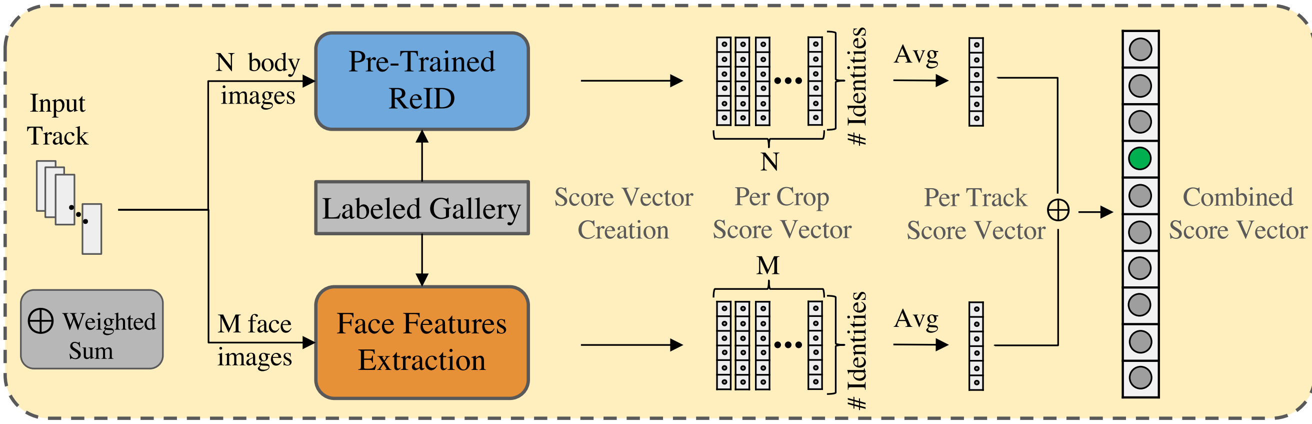

Figure 3. **Track Identity Prediction Overview.** After receiving an input track, the ReID and face modules generate a track score vector that indicates the likelihood that the track belongs to each identity based on all track images. Subsequently, these score vectors are combined to form a conclusive score vector, and the track is identified by its highest score (indicated by a green circle).

ReID and Face Score Vectors To compute v_{ReID}, for every crop q in a track of size N, we compute the feature vector using the ReID module. Then, for each identity $i \in I$, we determine the confidence that image q belongs to identity i by finding the maximum cosine-similarity between the feature vector of q and the feature vectors of all gallery samples of identity i in $G_{enriched}$. The ReID score for identity i, *i.e.* $v_{ReID}[i]$, is the mean confidence for identity i across all images in the track.

To compute v_{face} we follow the same procedure as above while changing the input images and gallery. In this case, the input images for the procedure are M detected face images from the original person track. Those are created by applying a face-detector on every image in the original track and comparing them to the gallery G_{face}, while verifying that the pose matches the main person in the image (Sec. 3.1.1).

Combining Score Vectors The combined score vector is defined as:

$$v_{pred} = \alpha \cdot v_{ReID} + (1 - \alpha) \cdot v_{face} \tag{1}$$

With α as a hyper-parameter determining the weight of each module. In the supplementary material, we examine different α values. The final prediction for the entire track is given by taking the identity with the highest score in v_{pred}.

4. The 42Street Dataset

Widely used clothes-changing ReID benchmarks (e.g. LTCC [35], PRCC [48], LaST [37], VC-Clothes [42]) do not capture crowded scenes that include multiple clothes changes per identity with various scale and illumination conditions. Moreover, CCVID [17], a video-based clothes-changing ReID benchmark, was curated from a gait-recognition dataset (FVG [52]), in which the people are captured while walking towards the camera with clearly visible faces. On this dataset, a simple face-recognition model achieves superior results compared to ReID models, as shown in Tab. 6. For these reasons, we publish a new video-based clothes-changing dataset — the ***42Street*** dataset. While the theater-play based dataset is attractive since it addresses the challenges described above, there are not enough people-of-interest to create both training and evaluation sets. Hence, we publish it as an evaluation dataset only, which can be used as a new benchmark to test the generalization ability of ReID models.

Name	Type	ID	Gallery	Query	Enrich
PRCC	Img	71	3384	7416	2792
LTCC	Img	75	7050	493	40
LaST	Img	5807	125353	10176	4609
VC-Clothes	Img	256	8591	1020	618
CCVID	Vid	151	1074	834	734

Table 1. Statistics of the evaluation set of multiple clothes-changing ReID benchmarks. *Enrich* shows the number of query samples used in the gallery enrichment process of each dataset. For the CCVID, the numbers indicate the number of sequences.

The dataset is created using a public recording of the 42Street theatre play [32]. The play is ∼1.5 hours long and we split it into 5 equally long parts of ∼20 minutes each, with various clothes changes between the different parts. From *Part 1* we label samples from which a gallery is built. Parts *2–3* and *4–5* are used for validation and test, respectively. From these parts we randomly extract 5 validation videos and 10 test videos, each 17 seconds long.

5. Experiments

5.1. Experiments on Existing Benchmarks

In this section, we show the potential improvement of several ReID models when applying GEFF, across all benchmarks detailed in Tab. 1. For each benchmark, we apply GEFF on the most recent ReID model for which we were able to reproduce similar results to the results reported in the original paper. Additionally, as our method relies on faces, we also apply a face-recognition model — InsightFace [9–12], to all the evaluated benchmarks as a face module baseline.

5.1.1 Evaluation Protocol

Following most CC-ReID works, *Top-1* accuracy and *mAP* are used as evaluation metrics in three test scenarios:

- *General:* All query and gallery samples are used to calculate accuracy.
- *Clothes-Changing:* Uses only query samples that have matching gallery samples with different clothes. Additionally, gallery samples with the same identity and the same clothes are discarded.
- *Same-Clothes (SC):* Uses only query samples that have matching gallery samples with the same clothes. Additionally, gallery samples with the same identity and different clothes are discarded.

Under all settings, gallery samples with the same identity and the same camera id are discarded.

5.1.2 Results — Existing Benchmarks

In Tab. 2, we apply GEFF to three ReID models on the PRCC and LTCC benchmarks showing an average improvement to the *Top-1* and *mAP* metrics, under most evaluated settings. Specifically, our method achieves an average improvement of 33.5% and 6.7%, under the clothes-changing setting, respectively. Tabs. 3 to 5 show that when applying GEFF on the SOTA models on the PRCC, LTCC, LAST, and VC-Clothes datasets, new SOTA results are achieved. On the CCVID benchmark, as the SOTA model DCR-ReID requires pre-processing of the dataset which is yet to be published, we apply our method on the second best model — CAL. Tab. 6 shows that applying GEFF achieves a significant improvement (outperforming even the DCR-ReID model). Notice that for this dataset, the baseline face-recognition model, InsightFace, achieves superior results under the general setting. We attribute this success to the fact that the CCVID dataset was curated from a gait recognition dataset, in which every image includes a clearly visible face. While the face-recognition model is successful on this dataset, the results on the other datasets, suggest that it is insufficient by itself.

Cross-Dataset Results In this experiment, we analyze the generalization ability of different models when training on one dataset but testing on another. Tab. 7 shows that applying GEFF to three ReID models increases their generalization abilities. While training the ReID models on the PRCC dataset and evaluating them on the LTCC dataset, applying GEFF achieves an average *Top-1* improvement of 4.7% and 6.4% on the General and Clothes-Changing settings, respectively. While training the ReID models on the LTCC dataset and evaluating them on the PRCC dataset, applying GEFF achieves an average *Top-1* improvement of 5.2% and 44.5% on the Same-Clothes and Clothes-Changing settings, respectively. Note that in both experiments, CTL was trained on the DukeMTMC [36] dataset.

5.2. Experiments on The 42Street Dataset

5.2.1 Evaluation Protocol

Given the gallery created from *part 1* of the play, we apply our method to the CTL ReID model and measure its performance on the evaluation videos of the ***42Street*** dataset. In our evaluation protocol, all models are evaluated without any training, as training data is not provided in this dataset.

Evaluation Metrics Similar to most ReID works, we measure the performance of a model using the top-1 metric. Since the query tracks are of different lengths, we measure the top-1 accuracy of both *Per-Image* and *Per-Track* accuracy.

- *Per-Image Accuracy*: the number of correctly identified person crops in a video, divided by the total number of person crops in the video, across all evaluation videos.
- *Per-Track Accuracy*: the number of correctly classified tracks, *i.e.* whether the model's single prediction on the entire track is correct, divided by the total number of tracks in the video, across all evaluation videos.

We observe that the image-based models we assess generate individual predictions for each image and do not offer a single prediction for an entire track. To overcome this shortcoming, we establish a single prediction based on the majority vote for the entire track. Additionally, since the primary focus of this study is not on enhancing tracking capabilities, we exclude tracks with less than 10 frames from our evaluation calculations, as they are more likely to be tracking errors. We note that the evaluation protocol of the ***42Street*** dataset concerns only the detected person crops. The person detector of ByteTrack [51] which we used, achieved an IDF1 score of 80.5 on MOT16 [31].

5.2.2 Results — 42Street Dataset

Tab. 8 summarizes the results on the ***42Street*** dataset. Similarly to the results shown in Sec. 5.1.2, since the evaluated models have a limited generalization ability, they perform poorly on this dataset as they are not being trained on

Method	*GE*	*FF*	PRCC				LTCC			
			Same-Clothes		Clothes-Changing		General		Clothes-Changing	
			top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
CTL [46]			81.9	73.0	28.1	23.4	36.7	11.1	11.7	4.8
	✓		93.8 (+11.9)	68.0 (-5.0)	78.5 (+50.4)	33.8 (+10.4)	38.9 (+2.2)	11.2 (+0.1)	16.3 (+4.6)	4.9 (+0.1)
	✓	✓	97.3 (+15.4)	76.6 (+3.6)	80.4 (+52.3)	42.3 (+18.9)	40.8 (+4.1)	12.0 (+0.9)	19.9 (+8.2)	5.5 (+0.7)
CAL [16]			100	99.8	55.7	56.3	74.4	41.2	39.3	19.0
	✓		99.7 (-0.3)	99.5 (-0.3)	82.2 (+26.5)	59.3 (+3.0)	75.1 (+0.7)	41.2 (0)	45.4 (+6.1)	19.2 (+0.2)
	✓	✓	99.6 (-0.4)	99.4 (-0.4)	83.5 (+27.8)	64.0 (+7.7)	75.5 (+1.1)	41.8 (+0.6)	46.4 (+7.1)	20.2 (+1.2)
AIM [49]			100	99.8	58.2	58.0	75.9	41.7	40.8	19.2
	✓		99.7 (-0.3)	99.4 (-0.4)	82.2 (+24.0)	60.4 (+2.4)	76.3 (+0.4)	41.7 (0)	45.2 (+4.4)	19.3 (+0.1)
	✓	✓	99.8 (-0.2)	99.1 (-0.7)	82.5 (+24.3)	64.7 (+6.7)	76.3 (+0.4)	42.3 (+0.6)	45.7 (+4.9)	20.3 (+1.1)
	✓		+3.7	-1.9	+33.6	+5.2	+1.1	0	+3.2	+0.1
Avg.	✓	✓	+4.9	+0.8	+34.6	+11.1	+1.8	+0.7	+6.7	+1.0

Table 2. **Applying GEFF on 3 ReID models over the PRCC and LTCC benchmarks.** In green are improvements of at least +1.0%. The first row of every evaluated model is a result reproduced by us, done in order to create a fair comparison between a ReID model that we trained and the improvement achieved by applying GEFF on it. The second and third rows show the improvement achieved by the gallery enrichment (GE) and by applying GEFF (Enriched Gallery + Face Module), respectively. The last rows (Avg.) show the average improvement of applying an Enriched Gallery and GEFF across all models.

it. However, when applying GEFF to the CTL and CAL ReID models, whilst not requiring any further training on the dataset, strong results are achieved, significantly outperforming the baseline ReID models. Fig. 4 visualizes the performance of the various models on a single frame from a test video. Interestingly, face-recognition by itself achieves mediocre results, as the face detector detected faces only in 74% of the total person crops, some of which are not of sufficient quality to be recognized correctly.

	Method	Year	PRCC				LTCC			
			Same-Clothes		Clothes-Changing		General		Clothes-Changing	
			top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
Non-CC-Models	*HACNN* [26]	2018	82.5	-	21.8	-	60.2	26.7	21.6	9.3
	PCB [39]	2018	99.8	97.0	41.8	38.7	65.1	30.6	23.5	10.0
	ISP [54]	2020	92.8	-	36.6	-	66.3	29.6	27.8	11.9
	InsightFace [9–12]	2020	95.6	70.0	78.0	54.7	27.4	9.2	24.5	8.3
	CTL [46]	2021	81.9	73.0	28.1	23.4	36.7	11.1	11.7	4.8
CC-Models	*FSAM* [22]	2021	98.8	-	54.5	-	73.2	35.4	38.5	16.2
	GI-ReID [23]	2022	86.0	-	33.3	-	63.2	29.4	23.7	10.4
	UCAD [47]	2022	96.5	-	45.3	-	74.6	34.8	32.5	15.1
	CAL [16]	2022	**100**	99.8	55.2	55.8	74.2	40.8	40.1	18.0
	ACID [50]	2023	99.1	99.0	55.4	**66.1**	65.1	30.6	29.1	14.5
	CCFA [19]	2023	99.6	98.7	61.2	58.4	75.8	**42.5**	45.3	**22.1**
	DCR-ReID [8]	2023	100	99.7	57.2	57.4	76.1	42.3	41.1	20.4
	AIM [49]	2023	**100**	**99.9**	57.9	58.3	**76.3**	41.1	40.6	19.1
	AIM + GEFF	2023	99.8	99.1	**82.5**	64.7	**76.3**	42.3	**45.7**	20.3

Table 3. **Results on the LTCC and PRCC benchmarks.** *CC-Models* (*Non-CC-Models*) are ReID models that were (not) designed specifically for the clothes-changing challenge. *AIM + GEFF* is the AIM model combined with a gallery enrichment and face module. Our method introduces an improvement over the Clothes-Changing setting and achieves comparable result on the Same-Clothes setting.

Method	top-1	mAP
CTL [46]	20.1	3.2
InsightFace [9–12]	57.8	31.1
OSNet [53]	63.8	20.9
BoT [28]	68.3	25.3
mAPLoss [37]	69.9	27.6
CAL [16]	73.7	28.8
CAL + GEFF	**78.0**	**37.2**

Table 4. **Results on LaST.** GEFF introduces a significant improvement when applied to the CAL model.

Method	General		SC		CC	
	top-1	mAP	top-1	mAP	top-1	mAP
InsightFace [9–12]	83.8	61.8	92.7	89.2	63.1	34.4
PCB [39]	87.7	74.6	94.7	94.3	62.0	62.2
MDLA [34]	88.9	76.8	94.3	93.9	59.2	60.8
3APF [42]	90.2	82.1	-	-	-	-
Part-align [38]	90.5	79.7	93.9	93.4	69.4	67.3
FSAM [22]	-	-	94.7	94.8	78.6	78.9
3DSL [6]	-	-	-	-	79.9	81.2
CAL [16]	92.9	87.2	95.1	95.3	81.4	81.7
CAL+GEFF	**94.9**	**88.9**	**96.5**	**96.3**	**86.7**	**84.4**

Table 5. **Results on the VC-Clothes benchmark.** GEFF introduces a significant improvement under all settings when applied to the CAL model.

Method	General	Clothes-Changing
	top-1	top-1
CTL [46]	71.8	69.3
I3D [4]	79.7	78.5
Non-Local [45]	80.7	79.3
AP3D [18]	80.9	80.1
TCLNet [40]	81.4	80.7
CAL [16]	82.9	81.9
DCR-ReID [8]	84.7	83.6
InsightFace [9–12]	**95.3**	73.5
CAL + GEFF	89.2	**90.5**

Table 6. **Results on the CCVID benchmark.** GEFF introduces a significant improvement when applied to the CAL model. The face model baseline (InsightFace) achieves a superior result as most tracks in this dataset include a clearly visible face image. As we explain in the supplementary material, mAP is not computed for video-based benchmarks when using our method.

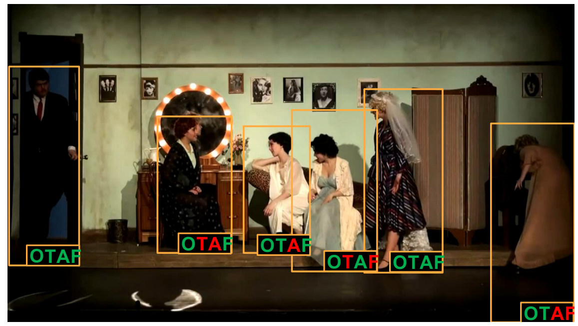

Figure 4. **Performance Visualization on the 42Street Dataset.** For each bounding box, the abbreviation 'OTAF' refers to the different models — O (Ours), T (CTL), A (CAL), F (InsightFace). The green and red colors correspond to a correct and incorrect prediction of each model on the bounding-box.

6. Ablation Study

In this section, we evaluate the impact of each component of GEFF on the overall performance of a ReID model, both on the existing benchmarks and the ***42Street*** dataset. From Tabs. 2 and 9 we conclude that the *Gallery Enrichment Process* introduces a significant improvement compared to using only the original gallery. In supplementary material, we further discuss the influence of additional raw data in the gallery enrichment process on the performance of the ReID model. Moreover, combining ReID and face modules using score vectors (even without gallery enrichment), significantly improves the results of ReID models. Finally, although the face module achieves solid results on some benchmarks (showing the significance of face features for ReID tasks), it is an insufficient model by itself as biometric information such as faces is often unavailable in ReID problems.

7. Ethical Considerations

New person ReID and tracking datasets raise privacy concerns as individuals may appear in them without consent. In this work, we use publicly available videos from the ***42 Street*** theatre play and only utilize face features for image retrieval and distance measurement, without identity matching. Our dataset is intended for academic use only. Moreover, we condemn the usage of ReID methods with nefarious intent and publish this work to progress academic research in this field.

8. Limitations

In order to enrich the gallery with samples of an unseen clothes-set (of a single identity), the gallery enrichment process relies on the assumption that at least one sample with these clothes includes a clearly visible face. For datasets where this assumption does not hold on multiple clothes-

Method	PRCC → LTCC				LTCC → PRCC			
	General		Clothes-Changing		Same-Clothes		Clothes-Changing	
	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
CTL	36.7	11.1	11.7	4.8	81.9	73.0	28.1	23.4
CTL + GEFF	40.8 (+4.1)	12.0 (+0.9)	19.9 (+8.2)	5.5 (+0.7)	97.3 (+15.4)	76.6 (+3.6)	80.4 (+52.3)	42.3 (+18.9)
CAL	21.9	6.1	7.4	3.3	99.6	96.4	37.7	35.4
CAL + GEFF	25.2 (+3.3)	6.1 (0)	11.5 (+4.1)	3.3 (0)	99.7 (+0.1)	94.8 (-1.5)	80.7 (+43.0)	48.9 (+13.5)
AIM	22.1	6.3	6.1	3.0	99.6	95.8	40.7	38.4
AIM + GEFF	29.2 (+7.1)	7.3 (+1.0)	13.0 (+6.9)	3.8 (+0.8)	99.7 (+0.1)	94.6 (-1.2)	81.0 (+40.3)	50.5 (+12.1)
Avg.	+4.7	0.6	+6.4	+0.5	+5.2	0.3	+45.2	+18.8

Table 7. **Cross-Dataset Generalization.** $X \rightarrow Y$ means that the model was trained on dataset X and evaluated on dataset Y, with the exception of CTL that was trained on DukeMTMC. These results suggest that the generalization ability of ReID models increases significantly when applying the proposed GEFF method. The last row (Avg.) shows the average improvement of applying GEFF.

Method	top-1	
	Image	Track
InsightFace	54.8	62.2
CAL	22.2	21.5
CAL + GEFF	74.1 (+51.9)	66.7 (+42.2)
CTL	31.1	26.7
CTL + GEFF	91.9 (+59.1)	81.8 (+51.1)

Table 8. **Results on the 42Street dataset.** All models are pre-trained on other datasets and are not fine-tuned on this dataset. *CTL, CAL* and *InsightFace* are image-based models, for which we apply a majority vote in order to calculate per-track accuracy.

Method	top-1	
	Image	Track
ReID	31.1	26.7
+ Enriched	80.1	71.1
Face	54.8	62.2
ReID + Face	81.3	66.7
*+ Enriched (**GEFF**)*	**90.5**	**77.8**

Table 9. **Ablation study of GEFF on the 42Street dataset.** The used ReID module is CTL.

sets (e.g. LTCC), applying GEFF would only introduce a slight improvement, as only a limited amount of query samples will be added to the original gallery. That said, we believe our assumption holds for many real-world scenarios and as such can introduce a significant improvement when applied to ReID models, as we showed on multiple datasets.

Additionally, to apply our work to real-world applications we use a tracking module to extract person tracks. Therefore, we inherit all the tracker's limitations, such as missed detection and mid-track identity switches. We use the tracking module without applying any changes to it, thus we do not deal with these potential tracking mistakes.

Finally, the proposed ***42Street*** dataset does not include a training set with a separate identity set, as the number of identities in the data was limited. Hence, it should be used for evaluation purposes only. Moreover, the dataset does not conform to the standard CC-ReID dataset settings, as it does not provide clothes and camera ids labels. However, we note that the gallery and query samples are taken from different (non-overlapping) parts of the play, captured with dynamic camera settings (various scene cuts, angles, and scales), and multiple clothes sets per identity. Therefore, we find this dataset a valid CC-ReID benchmark and an important contribution to the field, especially since the number of publicly available video CC-ReID datasets is limited.

9. Conclusion

In this work, we show a simple yet effective approach to address the clothes-changing ReID challenge by creating an enriched gallery from the given query and gallery samples. By applying GEFF on existing ReID models, new SOTA results are achieved, both on the existing clothes-changing ReID benchmarks and on the real-world clothes-changing dataset we publish, ***42Street***. Furthermore, we showed that by using GEFF, the generalization ability of existing ReID models increases, without requiring any further training.

10. Acknowledgements

This project was partially supported by ISF grant No. 1574/21

References

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In Tomás Pajdla and Jiří Matas, editors, *Computer Vision - ECCV 2004*, pages 469–481, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. 2

[2] Alaa Alnissany and Yazan Dayoub. Modified centroid triplet loss for person re-identification. *Journal of Big Data*, 10(1):74, 2023. 3

[3] Yogameena Balasubramanian, Nagavani Chandrasekaran, Sangeetha Asokan, and Saravana Sri Subramanian. Deep-facial feature-based person reidentification for authentication in surveillance applications. *IntechOpen*, 2019. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 7

[5] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8126–8133, 2019. 1, 2

[6] Jiaxing Chen, Xinyang Jiang, Fudong Wang, Jun Zhang, Feng Zheng, Xing Sun, and Wei-Shi Zheng. Learning 3d shape feature for texture-insensitive person re-identification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8142–8151, 2021. 7

[7] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020. 3

[8] Zhenyu Cui, Jiahuan Zhou, Yuxin Peng, Shiliang Zhang, and Yaowei Wang. Dcr-reid: Deep component reconstruction for cloth-changing person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 6, 7

[9] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the IEEE Conference on European Conference on Computer Vision*, 2020. 2, 5, 6, 7

[10] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2, 5, 6, 7

[11] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 2, 5, 6, 7

[12] Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *IJCV*, 2018. 2, 5, 6, 7

[13] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12036–12045, 2021. 2

[14] Zan Gao, Shenxun Wei, Weili Guan, Lei Zhu, Meng Wang, and Shenyong Chen. Identity-guided collaborative learning for cloth-changing person reidentification. *arXiv preprint arXiv:2304.04400*, 2023. 2

[15] Artur Grudzień, Marcin Kowalski, and Norbert Pałka. Face re-identification in thermal infrared spectrum based on thermalfacenet neural network. *2018 22nd International Microwave and Radar Conference (MIKON)*, pages 179–180, 2018. 2

[16] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *CVPR*, 2022. 1, 2, 6, 7

[17] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *CVPR*, 2022. 4

[18] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 228–243. Springer, 2020. 7

[19] Ke Han, Shaogang Gong, Yan Huang, Liang Wang, and Tieniu Tan. Clothing-change feature augmentation for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22066–22075, June 2023. 2, 6

[20] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 2

[21] Tianyu He, Xin Jin, Xu Shen, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Dense interaction learning for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1490–1501, 2021. 2

[22] Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10508–10517, 2021. 6, 7

[23] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14278–14287, 2022. 6

[24] Pei Li, Joel Brogan, and Patrick J. Flynn. Toward facial re-identification: Experiments with data from an operational surveillance camera plant. In *8th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2016, Niagara Falls, NY, USA, September 6-9, 2016*, pages 1–8. IEEE, 2016. 2

[25] Pei Li, Maria Loreto Prieto, Patrick J Flynn, and Domingo Mery. Learning face similarity for re-identification from real surveillance video: A deep metric solution. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 243–252. IEEE, 2017. 2

[26] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018. 6

[27] Mengmeng Liu, Zhi Ma, Tao Li, Yanfeng Jiang, and Kai Wang. Long-term person re-identification with dramatic appearance change: Algorithm and benchmark. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6406–6415, 2022. 1, 2

[28] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 7

[29] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, 2019. 2

[30] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. Aggregating deep pyramidal representations for person re-identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1544–1554, 2019. 2

[31] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 5

[32] SOTA music. 42 street - the complete musical, 2013. 4

[33] Athira Nambiar, Alexandre Bernardino, and Jacinto C. Nascimento. Gait-based person re-identification: A survey. *ACM Comput. Surv.*, 52(2), apr 2019. 1, 2

[34] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5399–5408, 2017. 7

[35] 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. 2, 4

[36] Ergys Ristani, Francesco Solera, Roger S. Zou, R. Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshops*, 2016. 5

[37] Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4390–4403, 2021. 4, 7

[38] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 402–419, 2018. 7

[39] 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 *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018. 6, 7

[40] Chao Tan. Tclnet: Learning to locate typhoon center using deep neural network. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 4600–4603. IEEE, 2021. 7

[41] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001. 2

[42] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 830–831, 2020. 2, 3, 4, 7

[43] Jianfeng Wang, Ye Yuan, and Gang Yu. Face attention network: An effective face detector for the occluded faces. *arXiv preprint arXiv:1711.07246*, 2017. 2

[44] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. 2

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 7

[46] Mikolaj Wieczorek, Barbara Rychalska, and Jacek Dabrowski. On the unreasonable effectiveness of centroids in image retrieval. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV 28*, pages 212–223. Springer, 2021. 2, 3, 6, 7

[47] Yuming Yan, Huimin Yu, Shuzhao Li, Zhaohui Lu, Jianfeng He, Haozhuo Zhang, and Runfa Wang. Weakening the influence of clothing: Universal clothing attribute disentanglement for person re-identification. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1523–1529. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track. 6

[48] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2029–2046, 2019. 1, 2, 4

[49] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1472–1481, June 2023. 1, 2, 6

[50] Zhengwei Yang, Xian Zhong, Zhun Zhong, Hong Liu, Zheng Wang, and Shin'Ichi Satoh. Win-win by competition: Auxiliary-free cloth-changing person re-identification. *IEEE Transactions on Image Processing*, 32:2985–2999, 2023. 6

[51] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 5

[52] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition

via disentangled representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4710–4719, 2019. 1, 4

[53] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 7

[54] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 346–363. Springer, 2020. 6