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

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

The supplementary material includes the following:

1. Code for existing benchmarks experiments. Due to the size limitation of the supplementary material, the ***42Street*** dataset and database are not included. However, those will be published alongside the paper.
2. Example videos - performance evaluation of different models on the ***42Street*** dataset.
3. The supplementary material document:
   (a) Using Raw-Data For Gallery Enrichment(Sec. 1)
   (b) Construction of the ***42Street*** Database (Sec. 2)
   (c) Thresholds (Sec. 3)
   (d) Implementation Details (Sec. 4)
   (e) Open-set Settings in the 42Street Dataset (Sec. 5)
   (f) Calculating mAP (Sec. 6)

1. Using Raw-Data For Gallery Enrichment

As explained in the ***42Street*** dataset creation, the test data of this dataset consists of two parts from the play, each around 20 minutes long. Out of these parts, we extract short videos, 17-seconds each, for evaluation. In this section, in addition to enriching the gallery with the query data (from the evaluation videos), we analyze the impact of using more raw-data (from the test part) for the gallery enrichment process. We start by using only the evaluation videos for gallery enrichment, and gradually add more randomly sampled raw data from the test parts. Fig. 1 shows the accuracy of the compared models as a function of the amount of raw-data used for the gallery enrichment process. Notice that the initial gallery enrichment with query data only, already introduces a significant improvement compared to not using enrichment at all, and that the accuracy increases as we add more raw-data. This is true for the Image-based ReID model, Track-based ReID model and our method, on both the closed and open set settings, with the best results achieved by our method. Even though raw-data is not available in the standard ReID task, we argue that such data is common in real-world scenarios like the application we presented, and show how it can be used to boost the performance of our method.

*These authors contributed equally

2. Construction of the *42Street* Dataset

2.1. Database Structure

We save each labeled crop from the ***42Street*** dataset in a designated database, which includes spatial and temporal information for each crop. Every entry in the database includes the following information:

- *label*.
- *im_name*: unique entry.
- *frame_num*: frame number of crop in video.
- *x1,y1,x2,y2*: top-left and bottom-right coordinates of the crop's bounding box.
- *conf*: person detector's confidence that a person exists in the crop.
- *vid_name*: the name of the video.
- *track_id*: the track number given by the tracker.
- *crop_id*: crop number within the track.
- *invalid*: boolean value set to true for crops that do not represent a clear person.

2.2. Annotating Evaluation Videos

To annotate an evaluation video with ground truth labels, a tracker, ByteTrack [16], is applied to automatically extract tracks of detected people, followed by a manual annotation of every track. These ground-truth labels, as well as person bounding boxes, track ids, video names, and more are saved in a designated database published alongside the dataset. In the supplementary material, we provide more details about the database and how it can be used for further research and to reproduce our results.

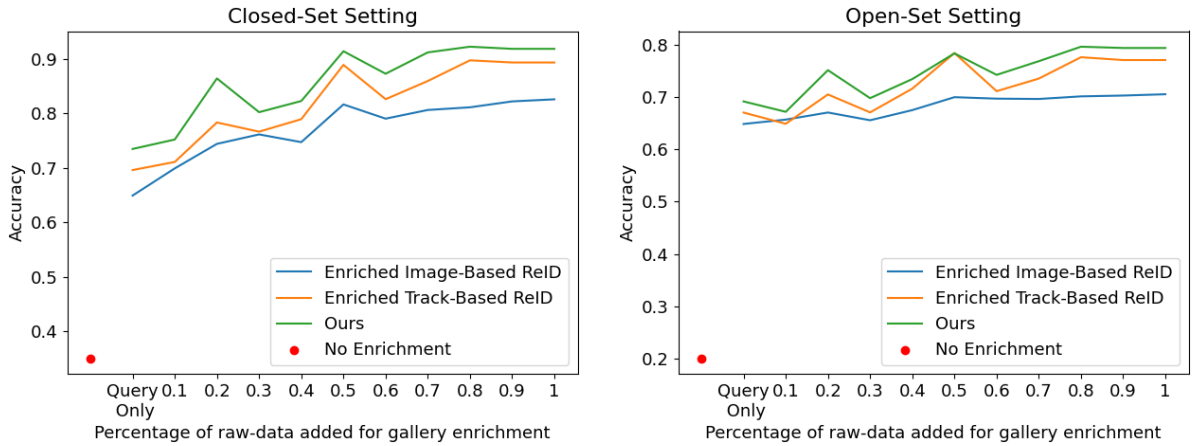

Figure 1. **Accuracy vs. Additional Raw Data for Gallery Enrichment in *42Street*** show the relationship between accuracy and the percentage of additional raw data utilized for gallery enrichment in the *42Street* dataset. The impact of ReID module on closed-set (left) and open-set (right) settings is demonstrated by enriching the gallery with extra raw data. The X-axis indicates the percentage of raw-data used for gallery enrichment. The red point represents image-based ReID module results without any enrichment. The *Query Only* denotes using data solely from evaluation videos for gallery enrichment. Each step shows the percentage of extra data from the raw-video. The findings suggest that enriching the gallery with the *Query Only* data significantly enhances the ReID module compared to no enrichment. Additionally, using more raw-data improves the overall accuracy of all compared modules.

3. Threshold Details

In Tab. 2 we detail the detection and similarity thresholds used by our method for the existing benchmarks. Recall from Sec. 3 that the values of the thresholds were defined based on the training data of each benchmark. Moreover, for the ***42Street*** dataset, we used the evaluation data to extract the following thresholds:

- *Detection Threshold*: For the gallery enrichment process we used a threshold of 0.8 and during inference, we used 0.7. The reason we use a higher threshold for the gallery enrichment process is that we are interested in creating a highly accurate gallery. On the other hand, during inference, we want to use more images with faces, and we are more tolerant of mistakes.
- *Similarity Threshold*: During the gallery enrichment process we use a threshold of 0.4 for the cosine similarity between an unlabeled sample and the labeled gallery. That is, if the maximal cosine similarity between the feature vector of the unlabeled sample and the labeled gallery is below this threshold, this sample will not be added to the enriched gallery.
- *Rank Difference*: In addition to detecting only clearly visible faces with high similarity to the labeled gallery, we want to use only samples for which the face model was confident about their identity. We measure this confidence as the difference between the similarity score of the top-1 and top-2 predictions of the model. In the ***42Street*** dataset, if the difference is below 0.1, we do not add the sample to the enriched gallery.

Under the open-set setting, to label people who are not in the people-of-interest set as "Unknown", we set another threshold. This threshold asserts that a person will be labeled as "Unknown" if the similarity to all people-of-interest is below 0.3. This means that whilst having high detection confidence, the model is not confident of the identity.

4. Implementation Details

4.1. Setting Hyper-Parameters

A typical face-detection model produces a confidence score that a face exists in a given image. Additionally, cosine distances between the face feature vector of the input image and a face gallery are calculated. These distances can be used as the confidence that the given input image belongs to a certain identity. To achieve the best results, we set thresholds for both confidence scores. Given a dataset, we apply our gallery enrichment method to the training set to find the best combination of these two thresholds. Ideally, we would like to find the optimal combination that achieves the highest prediction accuracy based on face features, whilst maintaining a high detection accuracy per person. Meaning, that we would like the model to predict at least one image of each identity in order to create an enriched gallery with examples of all identities. However, there exists a trade-off between the per-identity detection accuracy and the prediction accuracy: for a higher detection threshold, the number of unique predicted identities decreases. At the same time, since the detected faces are of better quality, the prediction accuracy increases. Vice versa,

α	PRCC				LTCC			
	Same-Clothes		Clothes-Changing		General		Clothes-Changing	
	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
0	93.7	64.5	71.3	47.7	15.6	6.5	13.3	6.3
0.25	99.6	82.8	84.3	62.5	76.3	37.0	46.4	18.8
0.5	99.8	95.7	83.4	65.2	76.3	40.6	45.9	19.8
0.75	99.8	99.1	82.5	64.7	76.3	42.3	45.7	20.3
1	99.7	99.4	82.2	60.4	76.3	41.7	45.2	19.3

Table 1. **Ablation study on the impact of different α values**. Alpha values of 0 and 1 are equivalent to using only the face and ReID modules, respectively. We conclude that an α of 0.75 presents a good balance between the weight given to the ReID and Face modules.

Threshold	Detection	Similarity
CCVID	0.5	0.75
LTCC	0.85	0.5
PRCC	0.7	0.65
LaST	0.7	0.45

Table 2. **Threshold values for existing benchmarks.** The different detection and similarity thresholds used for the gallery enrichment process in each of the existing benchmarks.

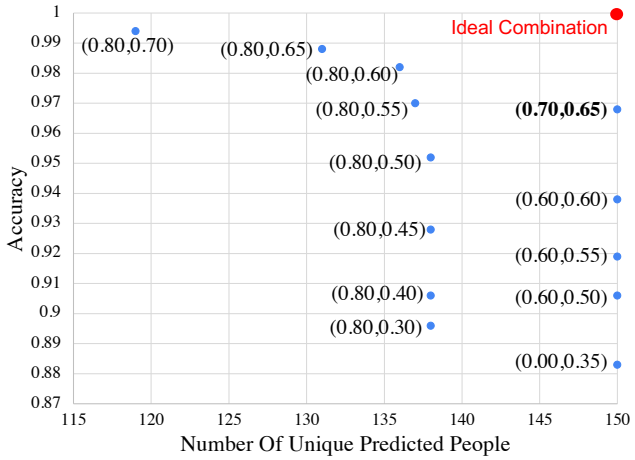

Figure 2. **Prediction Accuracy vs. Number of Unique People.** Example of the thresholds trade-off on the PRCC benchmark. Each data tuple represents the values of the detection threshold (left) and similarity threshold (right). We look for the combination that achieves the highest prediction accuracy while predicting the highest number of unique people (bold).

a lower detection threshold, leads to a larger number of predicted identities while achieving lower prediction accuracy. Fig. 2 demonstrates this trade-off for the PRCC benchmark.

4.2. Used Face and ReID Modules

For our face module, we use *InsightFace* [3–6] for face detection, alignment, and feature extraction. For the ReID module, we examine both CTL [13], pre-trained on DukeMTMC [11], CAL [7] and AIM [15]. The hyper-parameter α used for combining the face score vector and the ReID score vector is set to 0.75, giving more weight to the ReID module. An ablation model of the impact of α is presented in Tab. 1

The detection and similarity thresholds used by our face module (Sec. 4.1) are determined according to the training set for the existing benchmarks, and the validation set for the ***42Street*** dataset. In Sec. 3, we detail all the thresholds used for the different datasets.

For ***42Street***, the entire raw videos of test-set parts are used as query input for the gallery enrichment process. To apply our method on the real-world application described in the paper, we use ByteTrack [16] as our tracking module as well as the MMCV [1] and MMTRACK [2] frameworks. To deal with some of the tracking limitations, as the original play includes many scene cuts, i.e. abrupt changes of the camera angle or zoom, we apply a scene cut detection algorithm, and split tracks accordingly if necessary.

4.3. Hardware

First, we note, that our method does not require any training and uses only pre-trained ReID and face modules. However, since the evaluated ReID models do not release their trained checkpoints, we reproduced the results using the original code published by the authors. In this work, we used two GPUs: NVIDIA TITAN V with 12GB memory and Quadro RTX 5000 with 16GB.

5. Open-set Settings

Open-Set and Closed-Set A closed-set setting assumes that every identity in the query set has at least one corresponding sample in the gallery set [8]. In contrast, in an open-set setting, some identities in the query may not be present in the gallery. While closed-set ReID can be seen as an instance retrieval problem, open-set ReID is usually formulated as a person verification problem. In this formulation, the model is required to discriminate whether two

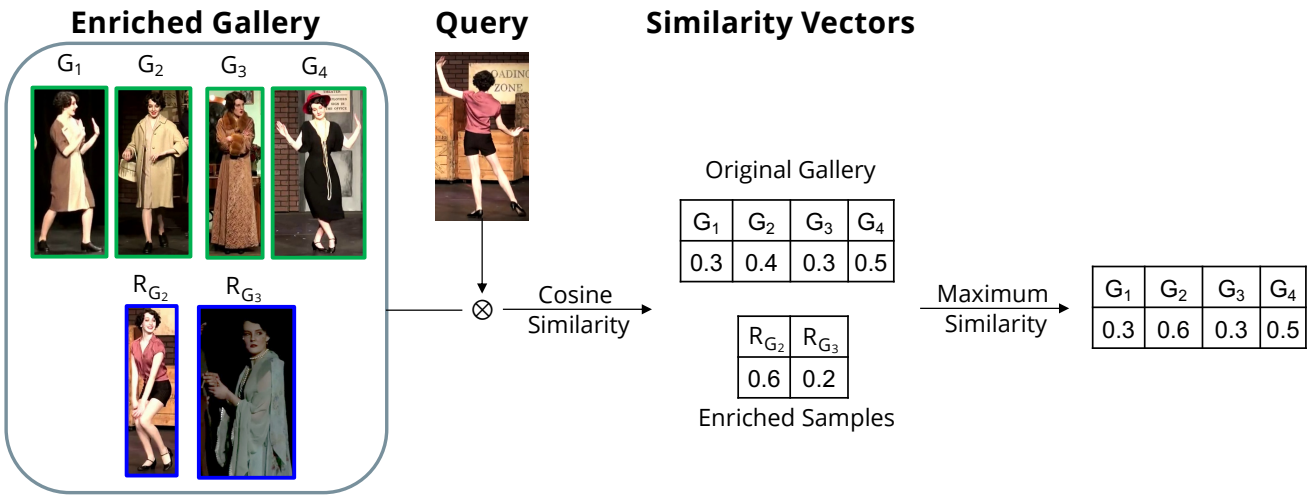

Figure 3. **Computing Similarities from an Enriched Gallery.** Given an enriched gallery and a query sample, we first compute similarity vectors between the original gallery samples (green frames) and the query (top similarity vector), and the enriched samples (blue frames) and the query (bottom similarity vector). Then, we combine the two similarity vectors into a final similarity vector in the size of the original gallery, by taking the maximum between each original sample similarity and the enriched samples that used it as a reference. For example, note that the similarity of gallery sample G_2 was replaced with the similarity of the enriched sample R_{G_2} that used it as a reference in the gallery enrichment process.

person images belong to the same identity [12, 17]. Models that address the open-set scenario [9, 10, 14] typically learn to discriminate between a given query and gallery images according to their similarity [8]. In this work, we use the gallery enrichment process to classify previously unseen people as "Unknown". This allows us to recognize such query samples during inference and effectively enables models that originally addressed only the closed-set setting, to address the open-set setting.

Addressing the Open-Set Challenge In the open-set setting, we approach the task of labeling an out-of-gallery query sample by utilizing the gallery enrichment process. During this process, we label a given query sample as "Unknown" and add it to the enriched gallery if it fulfills the following criteria:

1. A face was detected.
2. The cosine similarity between the feature vector of the detected face and each of the labeled face gallery feature vectors is below a certain threshold.
3. The difference between the closest and second-closest predictions in the face score vector is below a certain threshold.

We evaluate this capability on the 42Street data set in Tab. 3. We limit this evaluation to the CTL model and our proposed dataset since, to the best of our knowledge, currently available CC-ReID benchmarks and ReID models operate under the closed-set setting.

Method	Closed-Set		Open-Set	
	Per Image	Per Track	Per Image	Per Track
CTL	31.3	26.7	20.5	15.5
CTL + GEFF	**91.9**	**81.8**	**80.5**	**65.2**

Table 3. **Results on the 42Street dataset under the closed/open set settings.** Applying our method to the pre-trained *CTL* model, significantly improves the results of the model under both settings.

6. Calculating mAP

In instance retrieval tasks, given a query sample, the goal of the model is to rank all gallery samples from the most similar to the least similar. The mAP metric measures the rank of all "positive" gallery samples for each query, i.e. the positions of all gallery samples with the same label as the query, compared to the positions of all other gallery samples. A model with a 100% mAP score would rank the "positive" samples before all other gallery samples. During the gallery enrichment process, we add samples to the gallery, hence resulting in a larger gallery than the original one. Therefore, in order to provide a fair comparison with previous works, during evaluation, we have to reduce the size of the gallery to its original size. For each query sample, a similarity vector is computed, holding the similarities between the query and all original gallery samples. Similarly, a similarity vector is computed between the query

and all enriched samples. Finally, the similarity vectors are combined, resulting in a similarity vector of the same size as the original gallery. The combination is done by iterating over every original gallery sample and examining the group of all enriched samples that used it as a reference (i.e. this sample was the most similar gallery sample based on face features similarity) during the enrichment. Then, we set the similarity of the gallery sample in the similarity vector, as the maximum similarity between the query and the samples in this group including the original gallery sample. This process is illustrated in Fig. 3. We note that for the evaluated video benchmarks (CCVID, 42Street) our method utilizes score vectors to predict the identity of an entire track. The score vector holds a score per identity, and not ranking on the entire gallery. Therefore, mAP is not computed on these benchmarks.

References

[1] MMCV Contributors. MMCV: OpenMMLab computer vision foundation. https://github.com/open-mmlab/mmcv, 2018. 3

[2] MMTracking Contributors. MMTracking: OpenMMLab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking, 2020. 3

[3] 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. 3

[4] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 3

[5] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 3

[6] 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. 3

[7] 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. 3

[8] Qingming Leng, Mang Ye, and Qi Tian. A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1092–1108, 2020. 3, 4

[9] Xiang Li, Ancong Wu, and Wei-Shi Zheng. Adversarial open-world person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 280–296, 2018. 4

[10] Giuseppe Lisanti, Niki Martinel, Alberto Del Bimbo, and Gian Luca Foresti. Group re-identification via unsupervised transfer of sparse features encoding. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2468–2477, 2017. 4

[11] 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. 3

[12] Hanxiao Wang, Xiatian Zhu, Tao Xiang, and Shaogang Gong. Towards unsupervised open-set person re-identification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 769–773, 2016. 4

[13] 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. 3

[14] Hao Xiao, Weiyao Lin, Bin Sheng, Ke Lu, Junchi Yan, Jingdong Wang, Errui Ding, Yihao Zhang, and Hongkai Xiong. Group re-identification: Leveraging and integrating multi-grain information. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 192–200. ACM, 2018. 4

[15] 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. 3

[16] 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. 1, 3

[17] Xiatian Zhu, Botong Wu, Dongcheng Huang, and Wei-Shi Zheng. Fast open-world person re-identification. *IEEE Transactions on Image Processing*, 27(5):2286–2300, 2018. 4