Deep Compact Person Re-Identification with Distractor Synthesis via Guided DC-GANs

Víctor Ponce-López[0000−0002−4662−5722], Tilo Burghardt, Yue Sun, Sion Hannuna, Dima Damen[0000−0001−8804−6238], and Majid Mirmehdi[0000−0002−6478−1403]

Dept. of Computer Science, University of Bristol, Bristol, BS8 1UB, UK
{v.poncelopez@bristol.ac.uk}

Abstract. We present a dual-stream CNN that learns both appearance and facial features in tandem from still images and, after feature fusion, infers person identities. We then describe an alternative architecture of a single, lightweight ID-CondenseNet where a face detector-guided DC-GAN is used to generate distractor person images for enhanced training. For evaluation, we test both architectures on FLIMA, a new extension of an existing person re-identification dataset with added frame-by-frame annotations of face presence. Although the dual-stream CNN can outperform the CondenseNet approach on FLIMA, we show that the latter surpasses all state-of-the-art architectures in top-1 ranking performance when applied to the largest existing person re-identification dataset, MSMT17. We conclude that whilst re-identification performance is highly sensitive to the structure of datasets, distractor augmentation and network compression have a role to play for enhancing performance characteristics for larger scale applications.

Keywords: Person Re-ID · GANs · Distractor Synthesis · Deep Face Analysis.

1 Introduction

Visual person re-identification (Re-ID) is tasked with linking people's identities across multiple acquisition scenarios usually comprising disjoint fields of view. Given this highly variable operational environment, real-world Re-ID constitutes a particularly challenging sub-domain in computer vision due to inherent viewpoint and illumination changes, partial occlusions, limitations on resolution, and significant appearance alterations, such as changes in clothing [9,14]. These exigent visual conditions and the presence of facial occlusions render unimodal approaches, such as face recognition systems, *on their own* inadequate – and that is despite their human-level performance on favourable, well-known datasets, *e.g.* [33,16].

The emergence of deep learning techniques such as Convolutional Neural Networks (CNNs), streamed network designs, and large scale datasets [35,30,3] all have significantly evolved the field of Re-ID and addressed some of the issues mentioned above, with significant impact on applications including outdoor

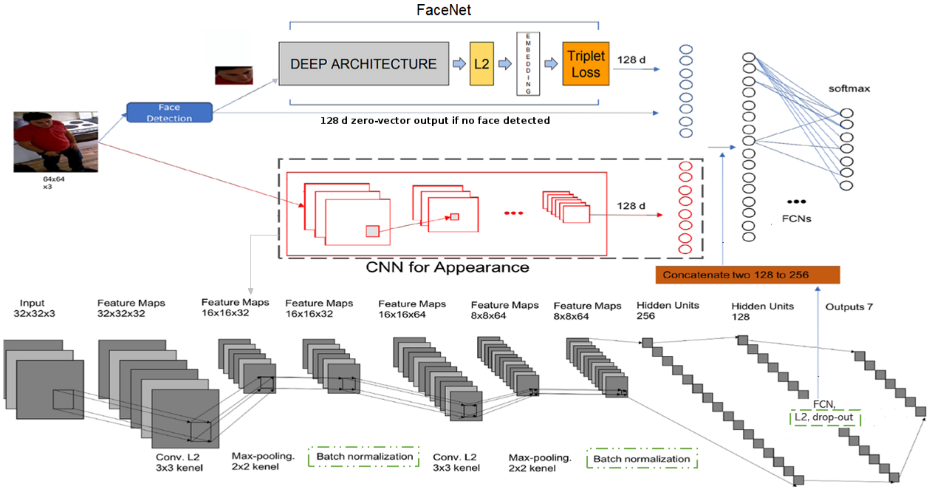

Fig. 1. Utilised Dual-Stream Architecture. Dual-stream CNN subdivided into appearance and facial feature streams using late feature fusion to map from frames to person identities. The appearance CNN network is based on LeNet-5 [15].

CCTV surveillance [7] and indoor e-health systems [1]. Whilst CNN-based representation learning excels at generating discriminative feature stacks that map inputs to compact identity clusters in embedding space, obtaining cross-referenced ground truth over long term [27], realising deployment of inexpensive inference platforms, and establishing visual identities from very limited data, remain challenging. In particular, the dependency of most deep learning paradigms on high computational requirements and on vast annotated training data pools appear as significant challenges to the field of person Re-ID.

In this paper, we explore the problem of ineffective training and heavy network footprints by proposing a generative-discriminative framework that generates images of a distractor class for enhancing the training of a discriminative ID-network – one which is lightweight and compact to deploy.

Initially, we describe a traditional two-stream CNN architecture (see Figure 1 split into appearance and facial feature streams that map in a conventional way, after late feature fusion, from still images to person identities. This network follows a regular streaming architecture deploying one visual task per stream before combined inference. Then, we propose to utilise the facial stream of this architecture to aid a setup where a single compact CondenseNet [11] is trained to perform Re-ID. Critically, training data is enhanced via a Deep Convolutional Generative Adversarial Network (DC-GAN) [20] generating a large set of distractor images semantically guided by facial semantics (see Figure 2). Note that synthesised distractor person images are generated by training input from across all identities; the synthesised content is thus *not* identical to given images of *any one* identity. Conceptually, adding such a distractor class as an extra identity to

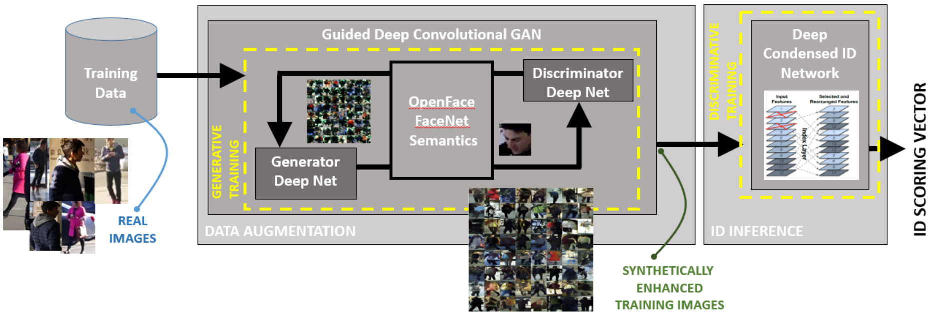

Fig. 2. Guided DC-GAN Compact Architecture. CondenseNet training is enhanced via distractor data generated by a DC-GAN which is semantically guided by a face detector.

the given identities for training the identification network enforces differentiation of persons from visually nearby distractors.

For evaluation, we introduce Facial-LIMA (FLIMA), which is an extension of the Long-term Identity-aware Multi-target multi-camerA dataset (LIMA) [14], by way of added frame-wise annotations of occurrence of faces. For an evaluation in a second, very different scenario, comparative experiments on the large Multi-Scene Multi-Time (MSMT17 [30]) person Re-ID dataset are presented. This comparison includes the dual-stream architecture and different settings of the proposed Guided DC-GAN trained compact CondenseNet against other reported results of the state-of-the-art on this dataset. Due to differences in the standard evaluation protocols, to sensitivity to the presence of detectable faces, and to resolution differences, we report on the varying efficacy of the tested approaches.

2 Related Work

The transition from hand-crafted features and small-scale evaluation to deep learning systems [36] with large-scale training datasets has fundamentally changed the way Re-ID systems are designed and operated. Looking back, early sliding window algorithms that made use of Histograms of Oriented Gradients (HOG) [6] or Haar-like Features [29] together with Eigenfaces [23] or Support Vector Machines (SVM) [5] were used to first detect and then classify persons or faces based on finding and categorizing a relevant image patch. However, these approaches' reliance on manually crafted features render them suboptimal in many application scenarios.

Deep Learning – Deep representation learning, on the other hand, avoids manual feature crafting entirely and has achieved significant improvements in image classification tasks compared to traditional methods. Space Displacement Neural Networks (SDNN) [15] demonstrated that neural nets can be effective for scale-invariant object detection too as shown, for instance, for face location [28], and detection and tracking [18] in videos. More recently, object detection has been addressed by region-focussed architectures such as R-CNN, Fast R-CNN, and Faster R-CNN [21] by integrating region proposal generation and classification by sharing convolutional features. With respect to person Re-ID, various CNN-centered approaches have been introduced recently, *e.g.* [31,24], including two-stream Siamese CNNs [4] providing pairwise class equivalences. Often, however, it is not the network design alone, but the availability of a large, learning-relevant training data corpus that makes the difference in effective network training.

Adversarial Synthesis – Generative Adversarial Networks (GANs) [8] have been applied widely and successfully to create large, learning-relevant training data via augmentation – building on their ability to construct a latent space that underpins the sparser training data, and then to sample from it to produce further training information. DC-GANs [20] pair the GAN concept with compact convolutional operations to synthesise visual content more efficiently. The DC-GAN's ability to organise the relationship between a latent space and an actual image space associated to the GAN input has been shown in a variety of applications, including face and pose analysis [20,17]. In these and other domains, latent spaces have been constructed that can convincingly model and parameterise object attributes, and hence dramatically reduce the amount of data needed for conditional generative modeling of complex image distributions. Some recent examples are face frontalisation [32] and identity preservation via generative modelling [26,34]. For instance in [34], Dual-Agent GANs (*i.e.* DA-GANs) were introduced to synthesise profile face images with varying poses.

Despite the deep learning revolution, the utilisation of *both* facial and person appearance features has remained a fundamental challenge in long-term monitoring [14,19]. Thus, in Section 4.1 we employ a two-stream CNN architecture (see Figure 1) split into appearance and facial feature streams. We then compare it in Section 4.2 to a single compact CondenseNet [11], which has access to both facial and overall appearance information, where training data is enhanced via a DC-GAN [20] performing distractor image generation. These models are then explored and results are presented and discussed in Section 5. We begin by introducing the datasets used.

3 Datasets: LIMA, FLIMA and MSMT17

The LIMA dataset [14] consists of $188,427$ frames of 7 manually labeled identities associated to person bounding box tracklets estimated by OpenNI NiTE. Identities refer to 6 person identities and 1 'unknown' label, which represents one distractor class that acts as an umbrella to capture any non-identity including noise or multiple people in the same bounding box tracklet. The whole dataset is recorded in various indoor environments and split into 13 sessions. According to previous works for long-term analysis [14,19], one fundamental evaluation protocol is to perform a leave-one-out performance evaluation with a train-test ratio of $12:1$ to validate the generalization capability over the different periods.

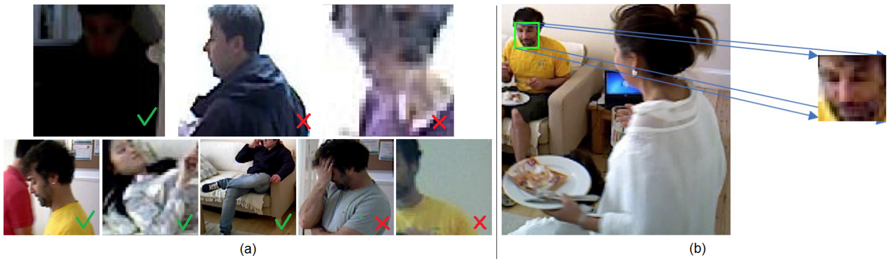

Fig. 3. FLIMA Data Annotation. (a) Examples of challenging face annotations and one example **(b)** where 2 faces are contained in the bounding box.

The FLIMA[1] dataset extends LIMA and assigns to every person bounding box an additional tag indicating the presence or absence of a face. Note that if a bounding box contains more than one face, the box will still just be labelled as 'face'. In general, well resolved frontal-to-profile facial occurrences are labeled as a 'face'. By contrast, faces that are mostly occluded or non-visible are considered as 'non-face'. Figure 3 provides some examples from the FLIMA dataset. Overall, $60,939$ bounding boxes are annotated as containing faces.

Beyond FLIMA, we also consider the MSMT17[2] dataset [30], as it is the largest person Re-ID dataset available. It contains $126,441$ bounding boxes of $4,101$ identities taken by 15 cameras during 4 days.

4 Proposed Methods

4.1 Dual-Stream Architecture

We propose a two-steam network as shown in detail in Figure 1. The fundamental design contains two separate streams for full person and facial appearance, respectively, which are combined through a fully connected layer that utilises Softmax activation plus a categorical cross-entropy cost function. Adam [13] is used as optimizer for network training.

The first stream deals with overall person appearance and a modified version of the LeNet-5 [15] architecture is utilised to implement it. Different to the standard implementation, (i) the input tensors are reshaped to $s = 64\times64\times3$, (ii) additional batch normalization layers [12] are introduced after the max-pooling layers to speed-up training, and (iii) L2-regularization and drop-out are added to the last fully connected layers in order to reduce over-fitting and stabilize training.

The second stream deals with facial information exclusively. It starts out by applying a face detector [21] to the input patch containing a detected person. If a face is found then the facial region is fed into FaceNet [22] based on OpenFace [2], which is adjusted to output a 128-D feature vector (or all zeros if no face

[1] FLIMA dataset will be made available at https://data.bris.ac.uk/data.

[2] MSMT17 dataset is online at https://www.pkuvmc.com/publications/msmt17.html

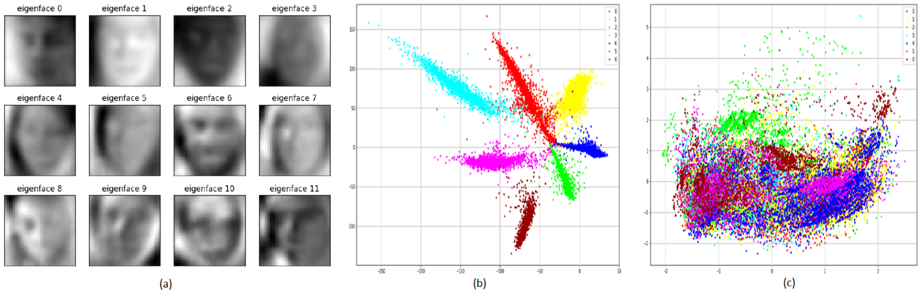

Fig. 4. Facial Feature Representation. (a) Eigenface components, (b) CNN features, (c) Eigenface features.

is found). These OpenFace features separate identities significantly better than traditional approaches, such as Eigenfaces [10] in tandem with a Radial Basis Function Support Vector Machine (RBF-SVM) and grid-search. Figure 4 illustrates the supremacy of deep features over the traditional approach on FLIMA face data. The experiments of our dual-stream network lasted 36 hours for training 1000 epochs on the FLIMA dataset with a Geforce Quadro K4100M running on 4GB RAM. We stabilised the training using the same parameters as in [15], but with a learning rate of 0.001 and a dropout probability of 0.4.

4.2 DC-GAN trained Compact CondenseNet

We argue that, instead of a classic dual-stream solution, a single compact CondenseNet [11] can perform Re-ID equally well or better as long as synthetic training can be effectively leveraged. The idea is to semantically guide an adversarial generative process that utilises the facial stream of the dual-stream architecture as a guidance network. As described in the original DC-GAN paper [20], a discriminator D and a generator G network are trained in tandem, the former learning to distinguish between generated and real input, the latter learning to produce outputs ever closer to the real inputs. The adversarial training loss of this process is, in agreement with [8]:

$$min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1-D(G(\mathbf{z})))], \tag{1}$$

where the data space in $\mathbf{x}$ and latent space in $\mathbf{z}$ are sampled for optimisation. One can understand (1) as a combination of losses, such that the global discriminator loss for the real and generated images is:

$$\mathcal{L}_D = \mathcal{L}_{D_{\mathbf{x}}} + \mathcal{L}_{D_{\mathbf{z}}}, \tag{2}$$

where $\mathcal{L}_{D_{\mathbf{x}}}$ is the discriminator loss for real images and $\mathcal{L}_{D_{\mathbf{z}}}$ the discriminator loss for the generated images, as:

$$\mathcal{L}_{D_{\mathbf{x}}} = \frac{1}{m}\sum_{i=1}^{m}\left[\log\left(D\left(\mathbf{x}^{(\mathbf{i})}\right)\right)\right], \tag{3}$$

$$\mathcal{L}_{D_{\mathbf{z}}} = \frac{1}{m}\sum_{i=1}^{m}\left[\log\left(1 - D\left(G\left(\mathbf{z}^{(i)}\right)\right)\right)\right]. \tag{4}$$

Based on this fundamental layout, we design a training regime that gives particular emphasis to high quality real training images – those which are well resolved and thus contain detectable facial features. These should ideally be modelled as producing a smaller discriminator loss compared to other training images. Following this paradigm, we introduce a penalisation term to our adversarial training loss for all training images where faces are *not* detected, and modify the discriminator losses from Eq. (3) and (4) to be:

$$\mathcal{L}'_{D_{\mathbf{x}}} = \frac{1}{m}\sum_{i=1}^{m}\left[\log\left(D\left(\mathbf{x}^{(\mathbf{i})}\right)\right) + \lambda_1\left(\mathit{\Delta}\left(\mathbf{x}^{(\mathbf{i})}\right)\right)\right], \tag{5}$$

$$\mathcal{L}'_{D_{\mathbf{z}}} = \frac{1}{m}\sum_{i=1}^{m}\left[\log\left(1 - D\left(G\left(\mathbf{z}^{(i)}\right)\right)\right) - \lambda_2\left(\mathit{\Delta}\left(G\left(\mathbf{z}^{(\mathbf{i})}\right)\right)\right)\right], \tag{6}$$

where $\mathit{\Delta}(.) = 1$ when there is no face detectable in the argument, and $\mathit{\Delta}(.) = 0$ otherwise. The two constants $\lambda_{1,2}$ are penalisation factors. Note that practically, this penalisation factor will be multiplied by $n \leq m$ according to n face-detected images within the current batch of m images.

Once the training procedure ends, $48,000$ synthetic training images are generated by the DC-GAN and used as an additional (distractor) class for training. We follow the framework of [19] to train a CondenseNet as a person ID-inference network, using 100 epochs for training the DC-GAN and $1,500$ epochs for the CondenseNet training processes, respectively. We use the same parameters as [19] and different values for the penalisation factors, *e.g.* $\lambda_1, \lambda_2 = [0, 0.025, 0.05]$, of the discriminator and generator, respectively.

5 Results

5.1 FLIMA Results

Table 1 shows results of the application of various architectures to the FLIMA dataset. The first row reports the Re-ID performance when only the $4,531$ facial patches detected by Faster RCNN are processed by an RBF-SVM applied to Eigenfaces. Both precision and recall are poor due to the method's reliance on a basic methodology and well-resolved facial features. In contrast, the second row shows comparative results of the method in [19], which utilises full person imagery. The third row depicts performance details of the DC-GAN trained

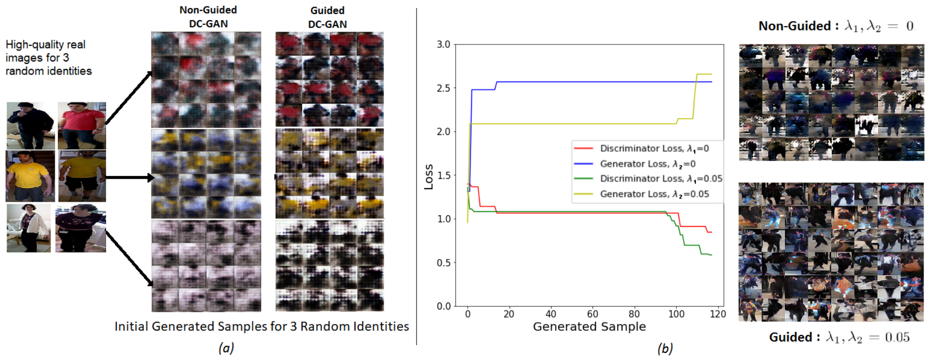

Fig. 5. Augmentation with Guided DC-GANs. (a) Augmentation samples after training DC-GANs with and without face detector guidance on FLIMA instances (3 individual identities shown). Note the improved quality of samples with guidance. (b) Training of the DC-GAN process on all identity samples of the MSMT17 dataset as used for the generation of distractors. We plot the loss values for the initial generated samples. We also show samples of global distractors for different $0 \leq \lambda_{1,2} \leq 0.05$ values. Again, note improvements when activating the guidance with a value above 0.

CondenseNet. The fourth row gives the recognition performance of the LeNet5 stream of the dual-stream architecture that deals with person appearance features only. The final row shows a considerably increased performance for Recall when deploying the full dual-stream architecture. Here, in a dataset with a small number of individuals and good facial resolution, a dual stream approach is advantageous, noticing similar F1-scores for appearance-only CNN stream and an appearance-based CondenseNet approach.

5.2 MSMT17 Results

Comparative performance measures, on what is currently the largest person Re-ID dataset (MSMT17), are provided in Table 2. This dataset has lower resolution facial content than FLIMA, uses a different evaluation scheme [31], and deals with far greater numbers of identities. We apply two metrics to quantify performance: correct classification rate of the top ranked individual (Rank@1) and mean Average Precision (mAP). Our dual-stream architecture and DC-GAN trained CondenseNet results are shown alongside four other approaches,

Table 1. Recognition performance on FLIMA

Method	# Test Images	Precision	Recall	F1-score
RCNN and RBF-SVM-Eigenface (Faces only)	4,531	0.56	0.52	0.47
Selective Augmentation Approach [19]	14,494	0.75	0.74	0.74
Our Guided DC-GAN trained CondenseNet	14,494	0.85	0.85	0.85
Our Appearance-Stream only	14,494	0.92	0.81	0.86
Our Full Dual-Stream	14,494	**0.93**	**0.90**	**0.91**

Table 2. Person Re-ID performance on MSMT17 dataset for single queries.

Method	Rank@1	mAP
Dual-Stream Architecture	4.89	5.91
GoogLeNet [25]	47.6	23.0
PDC [24]	58.0	29.7
GLAD [31]	61.4	**34.0**
Selective Augmentation Approach [19]	61.5	15.01
Our Guided DC-GAN ($\lambda_1, \lambda_2 = 0.05, 0.025$) trained CondenseNet	**63.85**	16.64
Our Guided DC-GAN ($\lambda_1, \lambda_2 = 0.05, 0$) trained CondenseNet	**65.51**	18.57

i.e. GoogLeNet [25], a Pose-driven Deep Convolutional model (PDC) [24], a Global-Local-Alignment Descriptor approach (GLAD) [31], and the Selective Augmentation Approach [19]. It can be seen that whilst GLAD outperforms all other methods with respect to mAP performance, our DC-GAN trained CondenseNet approach provides a significant improvement in Rank@1 performance for single-queries. This is a 4% performance increase above the next best performing method and 27% over GoogLeNet without using expensive and time-consuming training of very-deep multi-stream networks that benefit the mAP metric. Further, one has to consider that this increment is achieved with a significantly smaller footprint of the inference network – the produced CondenseNet carries 8× fewer parameters.

Given its very simple appearance CNN streams, the dual-stream architecture relies on features extracted from the facial stream. Compared to FLIMA, MSMT17 contains lower resolution facial patches and, most importantly, it has an evaluation scheme where the training set contains all different identity-classes to those from the test set. This renders the learning of specific identities completely ineffective and explains the poor performance of the dual-stream approach bound to learned facial features. The increased performance results with our guided DC-GAN trained compact CondenseNet on MSMT17 are based on leveraging distractor synthesis which remains highly relevant in this setting.

6 Conclusion

In this paper we investigated potential approaches for person Re-ID based on the exploitation of facial and person appearance representations, as well as an integration that semantically guides the image synthesis of DC-GAN training. First, we presented a traditional dual-stream architecture to learn *both* relevant appearance and facial features in combination from still images to infer person identities. We then described a second alternative architecture of a single, lightweight ID-CondenseNet, where a DC-GAN is used to generate distractor person images for enhanced training guided by the face detector leveraged from the face stream of our dual-stream CNN architecture. We introduced the FLIMA dataset with well-resolved facial content where we showed that the dual-stream approach performs superior. However, we then reported improvements in top-1 ranking performance compared to all tested state-of-the-art architectures on

MSMT17 when using our proposed CondenseNet system. We therefore conclude that re-identification performance is highly sensitive to the structure of datasets and evaluation metrics. As shown on MSMT17, distractor augmentation and network compression may nevertheless have a role to play for enhancing performance characteristics.

Acknowledgements

This work was performed in the SPHERE IRC funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/K031910/1.

References

1. Acampora, G., Cook, D.J., Rashidi, P., Vasilakos, A.V.: A Survey on Ambient Intelligence in Healthcare. Proceedings of the IEEE **101**(12), 2470–2494 (2013)
2. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: OpenFace: A General-Purpose Face Recognition Library with Mobile Applications. Tech. rep., CMU-CS-16-118 (2016)
3. Barbosa, I.B., Cristani, M., Caputo, B., Rognhaugen, A., Theoharis, T.: Looking Beyond Appearances: Synthetic Ttraining Data for Deep CNNs in Re-Identification. CVIU **167**, 50 – 62 (2018)
4. Chung, D., Tahboub, K., Delp, E.J.: A Two Stream Siamese Convolutional Neural Network for Person Re-Identification. In: ICCV (2017)
5. Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning **20**(3), 273–297 (1995)
6. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: CVPR. vol. 1, pp. 886–893 vol. 1 (2005)
7. Filković, I., Kalafatić, Z., Hrkać, T.: Deep metric learning for person Re-identification and De-identification. In: MIPRO. pp. 1360–1364 (2016)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: NIPS, pp. 2672–2680 (2014)
9. Haghighat, M., Abdel-Mottaleb, M.: Low Resolution Face Recognition in Surveillance Systems Using Discriminant Correlation Analysis. In: FG. pp. 912–917 (2017)
10. Halko, N., Martinsson, P., Tropp, J.: Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. SIAM Review **53**(2), 217–288 (2011)
11. Huang, G., Liu, S., van der Maaten, L., Weinberger, K.: CondenseNet: An Efficient DenseNet using Learned Group Convolutions. CoRR **abs/1711.09224** (2017)
12. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: ICML. vol. 37, pp. 448–456 (2015)
13. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. ICLR (2015)
14. Layne, R., Hannuna, S., Camplani, M., Hall, J., Hospedales, T.M., Xiang, T., Mirmehdi, M., Damen, D.: A Dataset for Persistent Multi-target Multi-camera Tracking in RGB-D. In: CVPR Workshops. pp. 1462–1470 (2017)
15. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based Learning applied to Document Recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
16. Lu, C., Tang, X.: Surpassing Human-level Face Verification Performance on LFW with Gaussian Face. In: AAAI. pp. 3811–3819 (2015)

17. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose Guided Person Image Generation. In: NIPS, pp. 406–416 (2017)
18. Nowlan, S.J., Platt, J.C.: A Convolutional Neural Network Hand Tracker. In: NIPS. pp. 901–908 (1995)
19. Ponce-López, V., Burghardt, T., Hannuna, S., Damen, D., Masullo, A., Mirmehdi, M.: Semantically Selective Augmentation for Deep Compact Person Re-Identification. In: ECCV Workshops. pp. 551–561 (2018)
20. Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. ICLR (2015)
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. TPAMI **39**(6), 1137–1149 (2017)
22. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823 (2015)
23. Sirovich, L., Kirby, M.: Low-Dimensional procedure for the Characterization of Human Faces. JOSA-A **4**(3), 519–524 (1987)
24. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-Driven Deep Convolutional Model for Person Re-identification. ICCV pp. 3980–3989 (2017)
25. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going Deeper with Convolutions. CVPR (2015)
26. Tran, L., Yin, X., Liu, X.: Disentangled Representation Learning GAN for Pose-Invariant Face Recognition. In: CVPR (2017)
27. Twomey, N., Diethe, T., Kull, M., Song, H., Camplani, M., Hannuna, S., Fafoutis, X., Zhu, N., Woznowski, P., Flach, P., Craddock, I.: The SPHERE Challenge: Activity Recognition with Multimodal Sensor Data. CoRR **abs/1603.00797** (2016)
28. Vaillant, R., Monrocq, C., Le Cun, Y.: Original Approach for the Localization of Objects in Images. IEE-VISP **141**(4), 245–250 (8 1994)
29. Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: CVPR. vol. 1, pp. I–511–I–518 vol.1 (2001)
30. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. In: CVPR (2018)
31. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval. CoRR **abs/1709.04329** (2017)
32. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Towards Large-Pose Face Frontalization in the Wild. In: ICCV. vol. 2017-October, pp. 4010–4019 (2017)
33. Yu, S.I., Meng, D., Zuo, W., Hauptmann, A.: The Solution Path Algorithm for Identity-Aware Multi-object Tracking. In: CVPR. pp. 3871–3879 (2016)
34. Zhao, J., Xiong, L., Jayashree, K., Li, J., Zhao, F., Wang, Z., Pranata, S., Shen, S., Yan, S., Feng, J.: Dual-Agent GANs for Photorealistic and Identity Preserving Profile Face Synthesis. In: NIPS, pp. 66–76 (2017)
35. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable Person Re-identification: A Benchmark. In: ICCV (2015)
36. Zheng, L., Yang, Y., Hauptmann, A.G.: Person Re-identification: Past, Present and Future. CoRR (2016)