







# Semantically Selective Augmentation for Deep Compact Person Re-Identification

Víctor Ponce-López<sup>(✉)</sup>, Tilo Burghardt, Sion Hannunna, Dima Damen,  
Alessandro Masullo, and Majid Mirmehdi

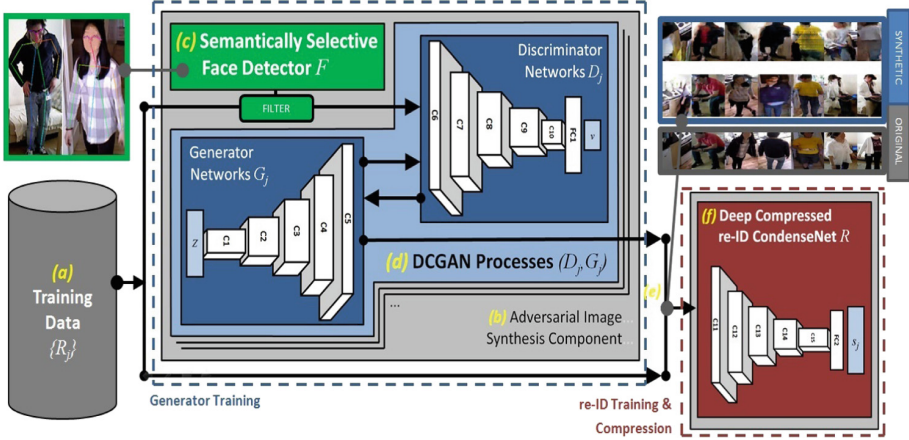
Department of Computer Science, Faculty of Engineering, University of Bristol,  
Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, UK  
vponcelop@gmail.com, {v.poncelopez, tb2935, sh1670, dima.damen,  
a.masullo, m.mirmehdi}@bristol.ac.uk

**Abstract.** We present a deep person re-identification approach that combines semantically selective, deep data augmentation with clustering-based network compression to generate high performance, light and fast inference networks. In particular, we propose to augment limited training data via sampling from a deep convolutional generative adversarial network (DCGAN), whose discriminator is constrained by a semantic classifier to explicitly control the domain specificity of the generation process. Thereby, we encode information in the classifier network which can be utilized to steer adversarial synthesis, and which fuels our CondenseNet ID-network training. We provide a quantitative and qualitative analysis of the approach and its variants on a number of datasets, obtaining results that outperform the state-of-the-art on the LIMA dataset for long-term monitoring in indoor living spaces.

**Keywords:** Person re-identification · Selective augmentation · Face filtering · Adversarial synthesis · Deep compression

## 1 Introduction

Person re-identification (Re-ID) across cameras with disjoint fields of view, given unobserved intervals and varying appearance (*e.g.* change in clothing), remains a challenging subdomain of computer vision. The task is particularly demanding whenever facial biometrics [29] are not explicitly applicable, be that due to very low resolution [7] or non-frontal shots. Deep learning approaches have recently been customized, moving the domain of person Re-ID forward [1] with potential impact on a wide range of applications, for example, CCTV surveillance [5] and e-health applications for living and working environments [23]. Yet, obtaining cross-referenced ground truth over long term [17, 27], realising deployment of inexpensive inference platforms, and establishing visual identities from strongly limited data – all remain fundamental challenges. In particular, the dependency of most deep learning paradigms on vast training data pools and high computational requirements for heavy inference networks appear as significant challenges to many person Re-ID settings.



**Fig. 1.** Framework overview. Visual deep learning pipeline at the core of our approach: inputs (dark gray) are semantically filtered via a face detector (green) to enhance adversarial augmentation via DCGANs (blue). Original and synthetic data are combined to train a compressed CondenseNet (red) for light and fast ID-inference. (Color figure online)

In this paper, we introduce an approach for producing high performance, light and fast deep Re-ID inference networks for people - built from limited training data and not explicitly dependent on face identification. To achieve this, we propose an interplay of three recent deep learning technologies as depicted in Fig. 1: deep convolutional adversarial networks (DCGANs) [21] as class-specific sample generators (in blue); face detectors [25] used as semantic guarantors to steer synthesis (in green); and a clustering-based CondenseNet [10] as a compressor (in red). We show that the proposed face-selective adversarial synthesis allows to generate new, semantically selective and meaningful artificial images that can improve subsequent training of compressive ID networks. Whilst the training cost of our approach can be significant due to the adversarial networks' slow and complicated convergence process [6], our parameter count of final CondenseNets is approximately one order of magnitude smaller than those of other state-of-the-art systems, such as ResNet50 [33]. We provide a quantitative and qualitative analysis over different adversarial synthesis paradigms for our approach, obtaining results that outperform the highest achievements on the LIMA dataset [14] for long-term monitoring in indoor living environments.

## 2 Related Work

Performing person Re-ID is a popular and long-standing research area with considerable history and specific associated challenges [32]. Low-resolution face recognition [7], gait and behaviour analysis [26], as well as full-person,

appearance-based recognition [32] all offer routes to performing ‘in-effect’ person ID or Re-ID. Here we will review particular technical aspects most relevant to the work at hand, *i.e.* looking specifically at recent augmentation and deep learning approaches for appearance-based methods.

**Augmentation.** Despite improvements in methods for high-quality, high-volume ground truth acquisition [17, 19], input data augmentation [18] remains a key strategy to support generalisation in deep network training generally. The use of synthetic training data presents several advantages, such as the ability to reduce the effort of labeling images and to generate customizable domain-specific data. It has been noted that combining synthetic and measured input often shows improved performance over using synthetic images only [24]. Recent examples of non-augmented, innovative approaches in the person Re-ID domain include feature selection strategies [8, 12], anthropometric profiling [2] using depth cameras, and multi-modal tracking [19], amongst many others. Augmentation has long been used in Re-ID scenarios too, for instance in [1], the authors consider structural aspects of the human body by exploiting mere RGB data to fully generate semi-realistic synthetic data as inputs to train neural networks, obtaining promising results for person Re-ID. Image augmentation techniques have also demonstrated effectiveness in improving the discriminative ability of learned CNN embeddings for person Re-ID, especially on large-scale datasets [1, 3, 33].

**Adversarial Synthesis.** Generative Adversarial Networks (GANs) [6] in particular have been widely and successfully applied to deliver augmentation – mainly building on their ability to construct a latent space that underpins the training data, and to sample from it to produce new training information. DCGANs [21] pair the GAN concept with compact convolutional operations to synthesise visual content more efficiently. The DCGAN’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 wide variety of applications, including face and pose analysis [16, 21]. In these and other domains, latent spaces have been constructed that can convincingly model and parameterise object attributes such as scale, rotation, and position from unsupervised models, and hence dramatically reduce the amount of data needed for conditional generative modeling of complex image distributions.

**Compression and Framework.** Given ever-growing computational requirements for very-deep inference networks, recent research into network compression and optimisation has produced a number of approaches capable of compactly capturing network functionality. Some examples include ShuffleNet [30], MobileNet [9], and CondenseNet [10], which have proven to be effective even when operating on small devices where computational resources are limited.

In our work, we combine semantic data selection for data steering, adversarial synthesis for training space expansion, and CondenseNet compression to sparsify

the built Re-ID classifier representation. Our solution operates on single images during inference, able to perform the Re-ID step in a one-shot paradigm<sup>1</sup>.

### 3 Methodology and Framework Overview

Figure 1 illustrates our methodology pipeline, which follows a generative - discriminative paradigm: (a) training data sets  $\{X_j\}$  of image patches are produced by a person detector, where each image patch is either associated to a known person identity label  $j \in \{1, \dots, N\}$ , or an ‘unknown’ identity label  $j = 0$ . (b) An image augmentation component then expands on this dataset. This component consists of (c) a facial filter network  $F$  based on multi-view bootstrapping and OpenPose [25]; and (d) DCGAN [21] processes, whose discriminator networks  $D_j$  are constrained by the semantic selector  $F$  to control domain specificity. The set of DCGANs, namely network pairs  $(D_j, G_j)$ , are employed to train generator networks  $G_j$  that synthesise unseen samples  $x$  associated with labels  $j \in \{0, \dots, N\}$ . These generators  $G_j$  are then used to produce large sets of samples. We focus on two types of scenarios: (1) a setup where we synthesize content for each identity class  $j$  individually, and (2) one where only a single ‘unlabeled person’ generator  $G$  is produced using all classes  $\{X_j\}$  as input, with the aim to generate generic identity content, rather than individual-specific imagery. Sampled output from generators is (e) unified with the original frame sets and labels, forming the input data for (f) training a Re-ID CondenseNet  $R$  that learns to map sample image patches  $x_j$  to ID score vectors  $s_j \in \mathbb{R}_+^{(N+1)}$  over all identity classes. This yields the sparse inference network  $R$  built implicitly compressed in order to support lightweight inference and deployment via a single network.

#### 3.1 Adversarial Synthesis of Training Information

**Adversarial Network Setup.** We utilise the generic adversarial training process of DCGANs [21] and its suggested network design in order to construct a de-convolutional, generative function  $G_j$  per synthesised label class  $j \in \{1, \dots, N\}$  that *after training* can produce new images  $x$  by sampling from a sparse latent space  $Z$ . Instead, a single ‘generic person’ network  $G$  is built in some experiments utilising all  $\{X_j\}$ . As in all adversarial setups, generative networks  $G$  or  $\{G_j\}$  are paired with discriminative networks  $D$  or  $\{D_j\}$ , respectively. The latter map from images  $x$  to an ‘is synthetic’ score  $v = D(x) > 0$ , reflecting network support for  $x \notin \{X_j\}$ . Essentially, the discriminative networks then learn to differentiate generator-produced patches ( $v \gg$ ) from original patches ( $v \ll$ ). However, we add to this classic dual network setup [16], a third externally trained classifier  $F$  that filters and thereby controls/selects the input to  $D_j$  - in our case one that restricts input to those samples where the presence of faces can be established<sup>2</sup>.

<sup>1</sup> Whilst results are competitive in this setting, discovering and matching segments during inference [14, 15, 20, 28, 34] is not used and could potentially further improve performance.

<sup>2</sup> We also modify the initial layer of the DCGAN to deal with a temporal gap of the specified number of frames. <https://github.com/vponcelo/DCGAN-tensorflow>.

**Facial Filtering.** We use the face keypoint detector from OpenPose [25] as the filter network  $F$  to semantically constrain the input to  $D_j$  and  $D$ . If at least one such keypoint can be established then face detection is defined as successful, where formally  $F(x_j \in X_j) \in [0, 1]$  is assigned to reflect either the absence (0) or presence (1) of a face.

**Training Process.** All networks then engage in an adversarial training process utilising Adam [13] to optimise the networks  $D$ ,  $\{D_j\}$ , and  $G$ ,  $\{G_j\}$ , respectively, according to the discussion in [21], whilst enforcing the domain semantics via  $F$ . The following detailed process describes this training regime: **(1)** each  $D$  or  $D_j$  is optimised towards minimising the negative log-likelihood  $-\log(D(x))$  based on the relevant inputs from  $\{X_j\}$  iff  $F(x_j) = 1$ , *i.e.* on original samples that are found to contain faces. **(2)** Network optimisation then switches to back-propagating errors into the entire networks  $D(G(z))$  or  $D_j(G_j(z))$ , respectively, where  $z$  is sampled from a randomly initialised Gaussian to generate synthetic content. Consider that whilst the generator weights are adjusted to minimise the negative log-likelihood  $-\log(D(G(z)))$ , encouraging  $v$  to get lower scores, the discriminator weights are adjusted to maximise it, prompting  $v$  to get higher scores. DCGAN training then proceeds by alternating between **(1)** and **(2)** until acceptable convergence.

### 3.2 Re-ID Network Training and Compression

Once the synthesis networks  $G$  and  $\{G_j\}$  are trained, we sample their output and combine it with all original training images (withholding 15% per class for testing) to train  $R$  as a CondenseNet [10], optimised via standard stochastic gradient descent with Nesterov momentum. Structurally,  $R$  maps from  $256 \times 256$ -sized RGB-tensors to a score vector over all identity classes. We perform 120 epochs of training on all layers, where layer-internal grouping is applied to the dense layers in order to actively structure network pathways by means of clustering [10]. This principle has been proven effective in DenseNets [11], ShuffleNets [30], and MobileNets [9]. However, CondenseNets extend this approach by introducing a compression mechanism to remove low-impact connections by discarding unused weights. As a consequence, the approach produces an ID inference network<sup>3</sup> which is implicitly compressed and supports lightweight deployment.

## 4 Datasets

**DukeMTMC-reID.** First we confirm the viability of a GAN-driven CondenseNet application in a traditional Re-ID setting (*e.g.* larger cardinality of identities, outdoor scenes) via the DukeMTMC-reID [22] dataset, which is a subset of a multi-target, multi-camera pedestrian data corpus. It contains eight

<sup>3</sup> <https://github.com/vponcelo/CondenseNet/>.



**Fig. 2.** DCGAN synthesis examples. Samples generated by  $G(z)$  with (b) or without (a) semantic controller. (c)  $1^{st}$  row: examples of generated images from  $G_0$  and  $G_j$  without semantic controller;  $2^{nd}$  row: with semantic controller;  $3^{rd}$  row: original samples from  $X_0$  and  $\{X_j\}$ . Columns in (c) are, from left to right, ‘unknown’ identity 0 and identities  $j \in \{1, \dots, N\}$ , respectively.

85-min high-res videos with pedestrian bounding boxes. It covers 1,812 identities, where 1,404 identities appear in more than two cameras and 408 identities (distractor IDs) appear in only one<sup>4</sup>.

**Market1501.** We also use a large-scale person Re-ID dataset called Market1501 [31] collected from 6 cameras covering 1,501 different identities across 19,732 images for testing and 12,936 images for training generated by a deformable part model [4].

**LIMA.** The **L**ong term **I**ntity aware **M**ulti-target multi-camer**A** tracking dataset [14], provides us with our main testbed for the approach. In contrast to previous datasets, image resolution is high enough in this dataset to effectively apply face detection as a semantic steer. LIMA contains a large set of 188,427 images of identity-tagged bounding boxes gathered over 13 independent sessions, where bounding boxes are estimated based on OpenNI NiTE operating on RGB-D and are grouped into time-stamped, local tracklets. The dataset covers a small set of 6 individuals filmed in various indoor environments, plus an additional ‘unknown’ class containing either background noise or multiple people in the same bounding box. Note that the LIMA dataset is acquired over a significant time period capturing actual people present in a home (*e.g.* residents and ‘guests’). This makes the dataset interesting as a test bed for long-term analysis, where people’s appearance varies significantly, including changes in clothing. In our experiments, we use a train-test ratio of 12:1 implementing a leave-one-session-out approach for cross-validation in order to probe how well performance generalises to different acquisition days.

## 5 Experiments and Results

We perform an extensive system analysis by applying the proposed pipeline mainly to the LIMA dataset. We define as the LIMA baseline the best so-far

<sup>4</sup> Evaluation protocol located at: [https://github.com/layumi/DukeMTMC-reID\\_evaluation](https://github.com/layumi/DukeMTMC-reID_evaluation).

reported micro precision metric on the dataset achieved by a hybrid M2&ME approach given in [14] - that is via tracking by recognition-enhanced constrained clustering with multiple enrolment. This approach assigns identities to frames where the accuracy of picking the correct identity as the top-ranking estimate is reported. Against this, we evaluate performance metrics for our approach judging either the performance over all ground truth labels  $j$ , including the ‘unknown content’ class (**ALL**), that is  $j \in \{0, \dots, N\}$ , or only for known identity ground-truth (**p-ID**), that is  $j \in \{1, \dots, N\}$ . We use two metrics: **prec@1** as the rank-one precision, *i.e.* the accuracy of selecting the correct identity for test frames according to the highest class score produced by the final Re-ID CondenseNet  $R$ , and **mAP** as mean Average Precision over all considered classes. Table 1 provides an overview of the results.

**Deep CondenseNet without Augmentation ( $R$  only).** The baseline (Table 1, row 1) is first compared to results obtained when training CondenseNet ( $R$ ) on original data only (Table 1, row 2). This deep compressed network outperforms the baseline **ALL prec@1** by 2.88%, in particular generalising better for cases of significant appearance change such as wearing different clothes over the session (*e.g.* without jacket and wearing a jacket afterwards. The **p-ID mAP** results (*i.e.* discarding the ‘unknown’ class) at 96.28% show that removing distracting content, *i.e.* manual semantic control during the test procedure, can produce scenarios of enhanced performance over filtered test subsets. We will now investigate how semantic control can be encoded via externally trained networks applied during training.

**Direct Semantic Control ( $FR$ ).** Simply introducing a semantic controller  $F$  to face-filter the input of  $R$  is, however, counter-productive and reduces performance significantly across all metrics (Table 1, row 5). Restricting  $R$  to train on only 39% of the input this way withholds critical identity information.

**Augmentation via DCGANs ( $G$ ).** Instead of restricting training input to the Re-ID network  $R$ , we therefore analyse how Re-ID performance is affected when semantic control is applied to generic DCGAN-synthesis via  $G$  of a cross-identity person class as suggested in [33]. Figure 2 shows examples of generated images and how the semantic controller affects the synthesis appearance. Augmentation of training data with 24k synthesised samples without semantic control (Table 1, row 3) improves performance slightly across all metrics, confirming benefits discussed in more detail in [33]. Table 2 confirms that applying such DCGAN synthesis together with CondenseNet compression to the DukeMTMC-reID dataset produce results comparable to [31]. Note that whilst the large deep ResNet50+LSRO [33] approach outperforms our compressed network significantly (Table 2, row 6), this comes at a cost of increasing the parameter cardinality by about an order of magnitude<sup>5</sup>. Moreover, non-controlled synthesis is generally limited. Indeed, on LIMA no further improvements can be made

<sup>5</sup> Require approximately  $8 \times$  fewer parameters and operations to achieve comparable accuracy *w.r.t.* other dense nets (*i.e.* 600 million less operations to perform inference on a single image) [10].

**Table 1. Results for LIMA** - top rank precision (**prec@1**) and mean Average Precision (**mAP**) for baseline (row 1), non-semantically controlled deep CondenseNet approaches (rows 2–4), and various forms of semantic control (rows 5–7). Note improvements across all metrics when utilising: compressed deep learning (row 2), augmentation (row 3), and semantically selective filtering (rows 6–7).

	ALL prec@1	p-ID prec@1	ALL mAP	p-ID mAP
No semantic control				
1: Baseline (M2&ME) [14]	89.1	-	-	-
2: No augmentation ( $R$ )	91.98	93.49	90.90	96.28
3: Augmentation $24kG \rightarrow R$	<i>92.43</i>	<i>94.27</i>	<i>91</i>	<i>96.95</i>
4: Augmentation $48kG \rightarrow R$	91.74	93.48	90.61	96.54
Semantic control via $F$				
5: No augmentation ( $FR$ )	82.02	92.14	72.90	95.48
6: Augment. $F322kG \rightarrow R$	<b>92.58</b>	<b>94.57</b>	<b>91.14</b>	97.02
7: $(24kG_0 + F24kG_j) \rightarrow R$	92.44	94.37	90.96	<b>97.04</b>

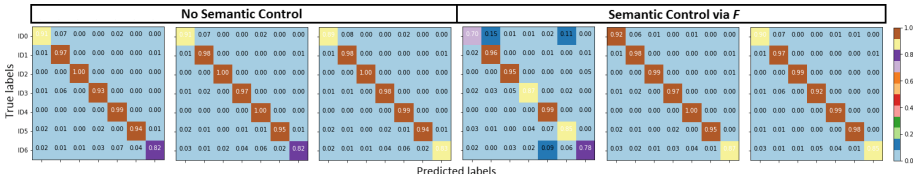
**Table 2. Results for DukeMTMC-reID** - top rank precision (**prec@1**) for classification and Single-Query (S-Q) performance with **No Semantic Control\***. Our results outperform [31] when using augmentation (row 4), or using Market1501 as synthesis input (row 5). However, the performance of the  $8\times$  larger ResNet50+LSRO [33] cannot be achieved in our setting of compression for lightweight deployment.

Method/NSC*	prec@1	prec@5	mAP	CMC@1 S-Q	mAP S-Q
1: Baseline BoW+KISSME [31]	-	-	-	25.13	12.17
2: Baseline LOMO+XQDA [31]	-	-	-	30.75	17.04
3: No augmentation ( $R$ )	87.70	95.54	87.79	<i>29.04</i>	<i>15.99</i>
4: Augmentation $24kG \rightarrow R$	88.08	95.73	88.26	<b>36.45</b>	<b>21.11</b>
5: Transfer 24k (Market) $G \rightarrow R$	<b>88.84</b>	<b>95.82</b>	<b>88.64</b>	<i>35.95</i>	<i>20.6</i>
6: ResNet50+LSRO [33] ( $\gg 8x$ )	-	-	-	67.68	47.13

by scaling up synthesis beyond 24k, whereby performance drops slightly across all metrics and overfitting to the synthesised data can be observed (Table 1, row 4). We now introduce semantic control to the input of augmentation and observe that the scaling-up limit can be lifted. Diminishing returns take over at levels above 300k though (*i.e.* 54% of synthesis *w.r.t.* original training data). We report results when synthesising 322k of imagery via  $G$ , improving results for all metrics (Table 1, row 6). We note that these improvements are achieved by synthesising distractors rather than individual-specific augmentations.

**Individual-Specific Augmentation ( $G_0 + G_j$ ).** To explore class-specific augmentation we train an entire set of DCGANs, *i.e.* produce generators  $G_j$  and  $G_0$ , respectively as specific identity and non-identity synthesis networks, and apply semantic control  $F$  to the identity classes  $j \in \{1, \dots, N\}$ . We observe that when balancing the synthesis of training imagery across all classes equally only slightly improves on **p-ID mAP**, whilst other measures cannot be advanced (Table 1,





**Fig. 3.** Some results as confusion matrices. Columns from left to right correspond to the experimental settings grouped by the presence of semantic selection, according to Table 1 rows 2–4 and 5–7, respectively.

row 7). Figure 3 provides further result visualisations. The limited improvements of this approach compared to non-identity-specific training (despite synthesis of overall more training data) suggest that, for the LIMA setup at least, person individuality can indeed be encoded by augmentation-supported modelling of a large, generic ‘person’ class against a more limited, non-augmented representation of individuals. Furthermore, experiments on the most challenging LIMA sessions demonstrate that the pre-trained generator  $G$  can generalize at re-training individual-specific generators  $G_0$  and  $G_j$  so as to reduce training cost of DCGAN individual-specific augmentation.

## 6 Conclusion

We introduced a deep person Re-ID approach that brought together semantically selective data augmentation with clustering-based network compression to produce light and fast inference networks. In particular, we showed that augmentation via sampling from a DCGAN, whose discriminator is constrained by a semantic face detector, can outperform the state-of-the-art on the LIMA dataset for long-term monitoring in indoor living environments. To explore the applicability of our framework without face detection in outdoor scenarios, we also considered well-known datasets for person Re-ID aimed at people matching, achieving competitive performance on the DukeMTMC-reID dataset.

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

## References

1. Barbosa, I.B., Cristani, M., Caputo, B., Rognhaugen, A., Theoharis, T.: Looking beyond appearances: synthetic training data for deep CNNs in re-identification. *Comput. Vis. Image Underst.* **167**, 50–62 (2018). <https://doi.org/10.1016/j.cviu.2017.12.002>
2. Bondi, E., Pala, P., Seidenari, L., Berretti, S., Del Bimbo, A.: Long term person re-identification from depth cameras using facial and skeleton data. In: Wannous, H., Pala, P., Daoudi, M., Flórez-Revuelta, F. (eds.) UHA3DS 2016. LNCS, vol. 10188, pp. 29–41. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91863-1\\_3](https://doi.org/10.1007/978-3-319-91863-1_3)

3. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: IEEE International Conference on Computer Vision Workshops, pp. 2590–2600 (2017). <https://doi.org/10.1109/ICCVW.2017.304>
4. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9), 1627–1645 (2010). <https://doi.org/10.1109/TPAMI.2009.167>
5. Filković, I., Kalafatić, Z., Hrkać, T.: Deep metric learning for person re-identification and de-identification. In: 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics, pp. 1360–1364 (2016). <https://doi.org/10.1109/MIPRO.2016.7522351>
6. Goodfellow, I., et al.: Generative Adversarial Nets. In: *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680. Curran Associates, Inc. (2014)
7. Haghghat, M., Abdel-Mottaleb, M.: Low resolution face recognition in surveillance systems using discriminant correlation analysis. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition, pp. 912–917 (2017). <https://doi.org/10.1109/FG.2017.130>
8. Hasan, M., Babaguchi, N.: Long-term people reidentification using anthropometric signature. In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems, pp. 1–6 (2016). <https://doi.org/10.1109/BTAS.2016.7791184>
9. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. *CoRR* abs/1704.04861 (2017)
10. Huang, G., Liu, S., van der Maaten, L., Weinberger, K.Q.: CondenseNet: an efficient densenet using learned group convolutions. preprint [arXiv:1711.09224](https://arxiv.org/abs/1711.09224) (2017)
11. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017)
12. Khan, F.M., Brèmond, F.: Multi-shot person re-identification using part appearance mixture. In: 2017 IEEE Winter Conference on Applications of Computer Vision, pp. 605–614 (2017). <https://doi.org/10.1109/WACV.2017.73>
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. *CoRR* abs/1412.6980 (2014)
14. Layne, R., et al.: A dataset for persistent multi-target multi-camera tracking in RGB-D. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1462–1470 (2017). <https://doi.org/10.1109/CVPRW.2017.189>
15. Liu, X., Ma, X., Wang, J., Wang, H.: M3l: multi-modality mining for metric learning in person re-identification. *Pattern Recognit.* **76**, 650–661 (2018). <https://doi.org/10.1016/j.patcog.2017.09.041>
16. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: Guyon, I., et al. (eds.) *Advances in Neural Information Processing Systems*, vol. 30, pp. 406–416. Curran Associates, Inc., New York (2017)
17. McConville, R., Byrne, D., Craddock, I., Piechocki, R., Pope, J., Santos-Rodriguez, R.: Understanding the quality of calibrations for indoor localisation. In: IEEE 4th World Forum on Internet of Things (2018). <https://doi.org/10.1109/WF-IoT.2018.8355159>
18. Perez, L., Wang, J.: The effectiveness of data augmentation in image classification using deep learning. *CoRR* abs/1712.04621 (2017)
19. Pham, T.T.T., Le, T.L., Dao, T.K.: Improvement of person tracking accuracy in camera network by fusing WiFi and visual information. *Informatica* **41**, 133–148 (2017)

20. Ponce-López, V., Escalante, H.J., Escalera, S., Baró, X.: Gesture and action recognition by evolved dynamic subgestures. In: Proceedings of the British Machine Vision Conference, pp. 129.1–129.13 (2015). <https://dx.doi.org/10.5244/C.29.129>
21. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: Proceedings of the International Conference on Learning Representations (2015)
22. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 17–35. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-48881-3\\_2](https://doi.org/10.1007/978-3-319-48881-3_2)
23. Sadri, F.: Ambient intelligence: a survey. *ACM Comput. Surv.* **43**(4), 36:1–36:66 (2011). <https://doi.org/10.1145/1978802.1978815>
24. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 2107–2116 (2017)
25. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2017)
26. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSN Trans. Comput. Vis. Appl.* **10**(1), 4 (2018). <https://doi.org/10.1186/s41074-018-0039-6>
27. Twomey, N., et al.: The SPHERE challenge: activity recognition with multimodal sensor data. preprint [arXiv:1603.00797](https://arxiv.org/abs/1603.00797) (2016)
28. Wu, L., Wang, Y., Li, X., Gao, J.: What-and-where to match: deep spatially multiplicative integration networks for person re-identification. *Pattern Recognit.* **76**, 727–738 (2018). <https://doi.org/10.1016/j.patcog.2017.10.004>
29. Yu, S.I., Meng, D., Zuo, W., Hauptmann, A.: The solution path algorithm for identity-aware multi-object tracking. In: The IEEE Conference on Computer Vision and Pattern Recognition (2016). <https://doi.org/10.1109/CVPR.2016.420>
30. Zhang, X., Zhou, X., Lin, M., Sun, J.: ShuffleNet: an extremely efficient convolutional neural network for mobile devices. *CoRR abs/1707.01083* (2017). <https://arxiv.org/abs/1707.01083>
31. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: 2015 IEEE International Conference on Computer Vision, pp. 1116–1124 (2015). <https://doi.org/10.1109/ICCV.2015.133>
32. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: past, present and future. *arXiv preprint arXiv:1610.02984* (2016)
33. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3754–3762 (2017). <https://doi.org/10.1109/ICCV.2017.405>
34. Zhou, S., et al.: Deep self-paced learning for person re-identification. *Pattern Recognit.* **76**, 739–751 (2018). <https://doi.org/10.1016/j.patcog.2017.10.005>