On the Effectiveness of Synthetic Data Sets for Training Person Re-identification Models

Rita Delussu, Lorenzo Putzu and Giorgio Fumera
Department of Electrical and Electronic Engineering, University of Cagliari
Email: {rita.delussu,lorenzo.putzu,fumera}@unica.it

***Abstract*—Person re-identification is a prominent topic in computer vision due to its security-related applications, and to the fact that issues such as variations in illumination, background, pedestrian pose and clothing appearance make it a very challenging task in real-world scenarios. State-of-the-art supervised methods require a huge manual annotation effort for training data and exhibit limited generalisation capability to unknown target domains. Synthetic data sets have recently been proposed as one possible solution to mitigate these problems, aimed at improving generalisation capability by encompassing a larger amount of variations in the above mentioned visual factors, with no need for manual annotation. However, existing synthetic data sets differ in many aspects, including the number of images, identities and cameras, and in their degree of photorealism, and there is not yet a clear understanding of how all such factors affect person re-identification performance. This work makes a first step towards filling this gap through an in-depth empirical investigation, where we use existing synthetic data sets for model training and real benchmark ones for performance evaluation. Our results provide interesting insights towards developing effective synthetic data sets for person re-identification.**

I. Introduction

Person re-identification (Re-Id) consists in matching images of a person of interest across different non-overlapping cameras. It has become a relevant research topic because of its widespread applications, chiefly in the fields of video surveillance and public security. At the same time, it is also a challenging computer vision task due to the presence of occlusions, both static and dynamic, and variations in visual factors such as illumination changes, shadows, differences in background, different camera type and perspective (e.g., orientation and scale), as well as different pedestrian poses, sizes, clothing appearance and attributes. Significant advancement has been achieved recently with the advent of Convolutional Neural Networks (CNNs) [1]. The most effective state-of-the-art methods employ supervised learning techniques and require a training set of cross-camera pedestrian images manually annotated with their identities. Although such methods have achieved excellent within-data set performance [2], [1], they suffer from over-fitting to a different extent and exhibit a limited generalisation capability to unseen target domains, which is highly desirable in real-world applications [3], [4], [5]. This is mainly due to the limitations of existing data set, e.g., relatively limited number of samples, limited variability and distribution of visual factors, as well as annotation errors [6]. In addition, emerging privacy regulations impose restrictions on data collection, also affecting Re-Id. Such issues hinders the development of large-scale and diverse source training data necessary for building effective CNN models. The interest in *synthetic* images has significantly grown in the last years in several computer vision applications, as possible solution to mitigate the above-mentioned issues. Synthetic images allow to generate very large data sets that can be *automatically* annotated without errors; they also allow (in principle) to control the data distribution and balancing of visual factors such as the ones mentioned above [1]. Existing synthetic data sets for Re-Id differ in size, number of identities and camera views, as well as in several visual factors such as illumination, pedestrian clothing appearance, background, and camera perspective. In particular, some of them focus on a single visual factor (e.g., clothing appearance [7]) or on realistic environment simulation (e.g., scene lighting [8]). Other data sets contain variations of different visual factors, such as viewpoint, pose, illumination and background [9], or pedestrian clothing, race, attributes and environments [10]; however, their empirical analysis was limited to one single factor, e.g., the influence of pedestrian rotation angles [9]. As another example, the large-scale RandPerson synthetic data set (almost 2M images) [10] was shown to improve the generalisation capability of Re-Id models with respect to the one attained by training on real-world benchmark data sets two orders of magnitude smaller, including Market-1501 and DukeMTMC-reID, but not on MSMT17, whose training set is just twice that of DukeMTMC-reID; this seems to suggest that just increasing data set size is not sufficient to improve generalisation capability, but the influence of the other factors is not clear. We also point out that in some works on synthetic data sets, the training data also included real images, which does not allow to understand the contribution of synthetic data alone [8]; moreover, only a few synthetic data sets have been directly compared [10]. A comprehensive and thorough analysis on the effectiveness of synthetic data sets for training Re-Id models is therefore still lacking; in particular, an evaluation of how the different features involved are beneficial to generalise to real images. This work makes a first step towards filling this gap through an in-depth quantitative evaluation of different factors that characterise existing synthetic data sets. To this aim, we train a widely used Re-Id model on publicly available synthetic data sets and test it on benchmark data sets of real images. In particular, we carry out a focused evaluation of the contribution of three factors on which it was possible to

operate to make the available synthetic data sets more directly comparable: the number of camera views, of identities, and of images per identity. Our analysis provides interesting insights on the role of the above factors and their interaction with data set size, overall quality of simulated scenes (especially human models and background), and the distribution and balance of the number of identities per camera and the number of images per identity in each camera. We believe that our results are a useful starting point towards the definition of guidelines on the development of effective synthetic data sets for Re-Id.

This work is organised as follows. Synthetic data sets in computer vision and in Re-Id are described in Sects. II and III; Sect. IV and V report our experimental setting and results; a discussion if finally given in Sect. VI.

II. Synthetic data sets in computer vision

Synthetic data have become very common for different real-world computer vision applications, such as medical image analysis [11], face recognition [12], crowd counting [13], tracking [14], object (e.g., vehicle) and pedestrian detection [15]. In recent years they are being increasingly used in place of real images for training deep learning models, since the generated images present high-quality and precise, automatic annotations, avoiding the effort of manual annotation and numerical limitations [16]. Existing approaches to generate synthetic images can be grouped into three main categories: image composition, adversarial networks (AN), and computer graphics engines (CGE). The simplest approach is image composition: it consists in superimposing a real or synthetic object model (e.g., a pedestrian) on a real image which is used as the background scene [17], [18], [19], [20]. The AN-based approach consists in generating synthetic images through an adversarial learning strategy [21]. It is mainly used for image-to-image translation, to reduce the gap between a source and a *specific* target domain, or for training data augmentation. The most known architecture is the Generative AN (GAN) [22], which has then been extended to CycleGAN [23], StarGAN [24] and Deep Convolutional GAN (DCGAN) [25]. CGEs allow producing fully synthetic images, instead. The most used CGEs are two well-known game design software, Unity [26] and Unreal [27]; the video game Grand Theft Auto V [28] (GTA for short); and the computer graphics tools MakeHuman [29], Adobe Fuse CC (Adobe for short) [30] and Blender [31].

Focusing on computer vision tasks related to video surveillance, the most widely used approach is the one based on CGE. It allows controlling every detail in the scene and, thus introducing variability of visual features, including human models. Since different CGEs present different features (e.g., MakeHuman and Adobe are more effective in generating 3D human models, whereas the other CGEs are better at generating 3D scenes), sometimes they have been used in combination to generate a synthetic data set, e.g., MakeHuman and Blender [32], [33], [9]. GANs have also been used to make synthetic images generated by CGEs look more photo-realistic [34].

		published data			downloaded data			
data set	year	#IDs	#images	#cam	#IDs	#images	#cam	access date
SOMAset [7]	2017	50	100,000	250	50	100,000	250	2021/12/28
SyRI [8]	2018	100	1,680,000	280	100	56,000	280	2021/10/12
PersonX [9]	2019	1266	273,456	6	1266	273,456	6	2021/09/22
RandPerson [10]	2020	8000	1,801,816	19	8000	132,145	4	2021/10/12
VC-Clothes [32]	2020	512	19,060	4	512	19060	4	2021/12/07
GPR [34]	2020	754	443,352	12	-	-	-	N.A
GPR+ [35]	2021	808	475,104	12	-	-	-	N.A
FineGPR [36]	2021	1150	2,028,600	36	1150	4600	4	2021/12/10
Synthetic18K [37]	2021	18,306	1,408,600	4	11,673	909908	-	2021/12/17
UnrealPerson [33]	2021	3000	120,000	34	2800	1,255,297	28	2021/12/09

TABLE I

Synthetic data sets details (number of identities #ID, images #images and cameras #cam) extracted from the corresponding papers (left) and the downloaded versions (right, together with the access date). Download links (if available) are in the corresponding references.

To the best of our knowledge, all of the publicly available synthetic image data sets for Re-Id have been generated using CGEs exclusively (see Sect. III). The reason is that CGEs allow to generate complex 3D human models, and in particular to control and to introduce variations to all the fine-grained details of pedestrian appearance, e.g., clothing, attributes and accessories [32], [33], scene details such as illumination and shadows [33], [9], complex background and camera perspective [9], and even occlusions with static objects or between pedestrians [33].

III. Synthetic data sets for person re-id

To our knowledge, ten different synthetic data sets have been developed so far for Re-Id. Their main characteristics are reported in Table I. Two of them, GPR and GPR+, turned out to be not available online. For most of the other data sets, the statistics reported in the respective papers (Table I, left) are different from the ones of the versions available for download (Table I, right). All ten data sets are described in the following. Fig. 1 reports two example images from each of them.

SOMAset was generated using MakeHuman and Blender [7]. It focuses on application scenarios where Re-Id has to be also performed under different clothing (clothing-independent or long-term Re-Id). To this aim, it contains 8 different types of clothes for each identity; each of the $50 \times 8 = 400$ subject-clothing combinations is rendered from 250 different camera orientations, with a different pose for each orientation. **SyRI** was generated using Unreal and Adobe [8]. Unreal was used to generate the virtual environment, composed of several scenes (e.g., shopping mall, museum, classroom) acquired under 140 illumination conditions. Adobe was used to generate human models with different body shapes and clothing. **PersonX** was generated using Unity [9], under different illuminations and viewpoints (i.e., front, right, back and left). Human models present different ages, body shapes, clothing, etc., and in some cases they carry accessories (e.g., backpack, hat). For each of the six cameras, the same pedestrian models (same identity, clothes, poses and viewpoint) have been used, and each pedestrian image has been rendered four times using three different realistic backgrounds and one background with a uniform colour. **RandPerson** was generated using Unity and MakeHuman [10]. It contains

images of several scenes (e.g., streets, cities, gyms) under different illuminations, Different pedestrian clothes have been generated from 10,000 texture maps, obtained by combining more than 625 colours and 16 patterns (e.g., stripes, spots). **VC-Clothes** was generated using GTA [32]. It contains four scenes (street, gate, parking lot and a natural scene) under different illumination conditions. Similarly to SOMAset, it focuses on clothing-independent Re-Id: human models present different clothes for each identity, as well as different ages, body shapes, etc. **GPR** was generated using GTA, and then the resulting images were modified through a GAN to make them look more photo-realistic [34]. Different weather and illumination conditions were simulated from 26 different scenes, e.g., mountain, street, mall. Human models present different body shapes, clothing, etc. Two further versions of GPR were generated using the same CGE: GPR+ [35], which contains a higher number of identities and of images than GPR, and **FineGPR** [36], which contains even more identities, cameras and images. Moreover, nine scenes (e.g., school, park, street) were created under seven weather (e.g., sunny, clouds) and illumination conditions (both day- and night-time). Human models present different attributes in terms of upper- and lower-body clothing colours, hats, bags, etc. **Synthetic18K** [37] was generated using Unity, and contains one indoor and three outdoor scenes (e.g., town square, subway station) under different weather and illumination conditions. Human models present different body shapes and clothing, and in some cases, different accessories (e.g., bags). **UnrealPerson** was generated using MakeHuman and Unreal [33], and contains four scenes (three urban outdoor and one indoor) under different illumination conditions; human models present more than 200 types of clothes and in some cases different accessories (e.g., masks, glasses, hats).

As mentioned above, most data sets' released version differ from those described in the respective papers. In particular, for RandPerson, SyRI and FineGPR, a much lower number of images were actually available, respectively 7.3%, 3.3% and 0.22% of the amount stated in the papers. In addition, the released version of RandPerson contains images from four cameras instead of nineteen, and in the experiments reported in the corresponding paper [10] only images from such four cameras have been used, to reduce redundancy and training time. However, it is not clear which visual factors introduced redundancy. On the contrary, the number of images of the downloaded version of UnrealPerson is about ten times the one reported in the corresponding paper [33], although the number of cameras is lower (28 instead of 34). In general, the main criterion adopted to generate the above data sets is to introduce a certain amount of variability in pedestrian appearance, including different body shapes, clothes, attributes and accessories, to a different extent depending on the data set; different ages and races have also been considered in VC-Clothes and SOMAset. Variability in the virtual scenes has also been introduced: all data sets except for SOMAset present different scenes, cameras and illuminations, even if with many differences among them (e.g., SyRI presents the highest number of illumination conditions, 140). Interestingly, only RandPerson and UnrealPerson focused also on generating photo-realistic scenes: they were generated under a set-up similar to real video surveillance systems, with multiple people moving at the same time and partial occlusions by static objects or by other people. The degree of photo-realism can be observed from the examples in Fig. 1. Focusing on human model appearance, in our opinion SyRI presents better visual details, whereas RandPerson and PersonX look less photo-realistic than the other data sets in terms of body shape and textures. Focusing on the background scene, RandPerson and GPR exhibit the highest degree of photo-realism.

It is evident that existing synthetic data sets for Re-Id exhibit many differences among them and since a comprehensive and thorough analysis of this issue is still missing, it is pertinent to ask how and to what extent the performance of Re-Id models trained on synthetic data is affected by the different features involved.

IV. Experimental setting

The main goal of our experiments is to give a first evaluation of the effectiveness of synthetic training data, focusing on different features that can be observed in existing data sets related to data set dimensions and visual factors. Besides GPR and GPR+, which were not available online, we excluded from our experiments Synthetic18K since the available documentation did not allow us to recover the camera ID of each image. On the other hand, we included SOMAset and VC-Clothes, although they focus on the more challenging clothing-independent setting and thus, for each identity, they contain images with different clothes. To perform our evaluation, we considered a single ResNet50 CNN model [38] pre-trained on ImageNet, which is widely used in Re-Id, usually as a backbone for more specific architectures. Although it does not achieve state-of-the-art performance on benchmark data sets, its choice is suitable to our goal, which is not to attain competitive performance but to compare synthetic data sets on common ground. To directly evaluate the effectiveness of synthetic data, we considered a cross-domain setting: we used, in turn, each synthetic data set as the training set, without adding to it any real image, and a data set of real images as the testing set. For this reason, in the case of VC-Clothes and PersonX, which were originally subdivided into a training set and a testing set (query and gallery sets), we used for training also all the testing images. For testing we used three well known benchmark data sets: Market-1501 (Market for short) [39], DukeMTMC-reID (Duke) [40] and MSMT17 (MSMT) [41], which are summarised in Table IV.

As performance measures we used both the cumulative matching curve (CMC) at ranks $k = 1, 5, 10, 20$, and mean Average Precision (mAP). During training, we used horizontal flip and random crop with a probability of 0.5 to reduce over-fitting. For optimisation, Stochastic Gradient Descent was used with momentum 0.9 and weight decay 5×10^{-4}; the learning rate was set to 0.00035.

Fig. 1. Synthetic data sets samples (left to right): SOMAset, SyRI, PersonX, GPR, RandPerson, VC-Clothes, FineGPR, Synthetic18k and UnrealPerson.

	#IDs / #images			#Cameras
Data set	Train	Query	Gallery	-
Market	751 / 12936	750 / 3368	751 / 15913	6
Duke	702 / 16522	702 / 2228	1110 / 17661	8
MSMT	1041 / 30248	3060 / 11659	3060 / 82161	15

TABLE II
REAL DATA SETS DETAILS: NUMBER OF IDENTITIES AND OF IMAGES (# IDS / # IMAGES) IN EACH SUB-SET, AND NUMBER OF CAMERAS.

Based on the above setting, we carried out the following experiments. We first evaluated the performance achieved using the whole synthetic data sets for training. For PersonX we also separately considered images with a realistic background and with a uniform background. However, since these data sets present significant differences in their dimensions and visual factors (see above), the above experiment does not allow to understand the contribution of each such aspect. Therefore we identified three features, that are the overall number of images, cameras and identities, on which it was possible to operate and compare synthetic data sets under more similar conditions. Accordingly, we progressively introduced several constraints on the above features, leading to five different experimental settings involving subsets of the considered data sets (see Table IV). In Experiment 1 (**Exp-1**) we reduced the data sets to a common number of cameras, which was set to 4, corresponding to the lowest number of cameras among the considered data sets (RandPerson and VC-Clothes). Whenever possible we chose four cameras from different scenes; for PersonX, we selected three cameras with a realistic background and one with a uniform background. In Experiment 2 (**Exp-2**) we reduced the data sets to a common number of both cameras (4, the same as Exp-1) and of images per identity (16, corresponding to the minimum value across the data sets, except for VC-Clothes and RandPerson that do not contain an equal number of images per identity, but present less than 16 images for some identities). For each identity, we selected the first 16 images. In Experiment 3 (**Exp-3**), we considered instead a common number of cameras (4, the same as in Exp-1 and Exp-2) and of identities (512, which is the minimum value among the considered data sets, corresponding to VC-Clothes); we selected the first 512 identities of each data set. In Experiment (**Exp-4**), we reduced the data sets to a common number of identities (512, the same as in Exp-3) and of images per identity (16, the same as in Exp-2). Finally, in Experiment 5 (**Exp-5**), we reduced the data sets to the same number of cameras (4, the same as in Exp-1, Exp-2 and Exp-3), identities (512, the same as in Exp-3 and Exp-4), and images per identity (16, the same as in Exp-2 and and Exp-4). In each of the above experiments, we also tried to understand how the various visual factors (e.g., clothing appearance) contributed to model performance.

V. EXPERIMENTAL RESULTS

The results of our experiments are presented and discussed in the following.

Training on the whole synthetic data sets. The results are reported in Table III. We observed that UnrealPerson, which is the largest data set, attained the highest performance on all target domains, and only RandPerson, which contains the highest number of identities (and thus exhibits in principle the largest variability in pedestrians' appearance) attained comparable performances, limited to Market. Instead, SOMAset and VC-Clothes were among the data sets that led to the worst performances: this was expected since they present different clothing appearances for each identity (see Sect. III). Also SyRI attained a rather poor performance: this may be due to the fact that it presents the lowest number of images per identity in each camera. Two other data sets, FineGPR and PersonX w/o background, attained a relatively poor performance. This was expected for FineGPR, due to its very small number of images; however, it is interesting that it outperformed SyRI, despite the latter contains about twelve times as many images: this seems to indicate that the quality and diversity of synthetic images (see Sect. III) can be more relevant than their bare amount. Another interesting result is that PersonX w/ background attained a much better performance than PersonX w/o background (we remind the reader that the same pedestrian images are present in both data sets): this clearly indicates that CNN models benefit from a realistic simulation of the background; this is confirmed by the fact that the performance of the whole PersonX (the union of PersonX w/o and w/ BG) is slightly worse than PersonX w/ BG, except for CMC at rank 20.

In table III we also report results obtained by training on real data sets under a cross-data set setting, using the same model.

We first observe that training on UnrealPerson led to better performances than training on each real data set, in all target domains. This might be due to UnrealPerson's high variability and level of realism. Despite RandPerson presents a lower number of cameras with respect to real data sets, it exhibited comparable performances in most cases. This suggest that also RandPerson presents a significant variability and a good level of realism. Training on the other synthetic data sets exhibited instead worse performances than those obtained by training on real ones, and in some cases, the gap was remarkable. This gap might be due to, e.g., the lower number of images (e.g., FineGPR), of identities (e.g., SOMAset), or on the absence of a background scene (e.g., PersonX).

Exp-1: common number of cameras. Table IV shows that reducing the data sets to a common number of cameras lead to a considerable reduction of the number of images for UnrealPerson, and to a lower extent also for PersonX. UnrealPerson attained the best performance also in this experiment, except for CMC at rank 20 on Market, where RandPerson slightly outperformed it. It is worth noting that, despite a drop in size to 11% with respect to the previous experiment (see Tables III and IV), the performance of UnrealPerson did not worsen that much. This seems to agree with the observation made above about FineGPR and SyRI, that the quality and diversity of synthetic images can be more relevant than their amount. Further confirmation comes from the fact that PersonX is outperformed by RandPerson and UnrealPerson, despite being larger than them: this can be explained by three different reasons: the human models of PersonX exhibit a lower quality than the ones of RandPerson and UnrealPerson(see Sect. III); some images of PersonX present a uniform, non-realistic background; each pedestrian image in PersonX has been replicated with all the different (uniform or realistic) backgrounds, which reduces diversity.

Exp-2: common number of cameras and images. The additional constraint on the number of images per identity lead to a considerable reduction in the size of PersonX and UnrealPerson, and to a lower extent also of VC-Clothes. Even in this case, the best performance in all target domains is attained by UnrealPerson (except for ranks 10 and 20 when Market is the target domain), although in this experiment RandPerson has three times more images. Surprisingly, PersonX and VC-Clothes attained a slightly better performance than in Exp-1. In particular, since VC-Clothes contains different clothes for each identity, it may have benefited from the increase of discriminant capability of clothing appearance cues due to the reduction of the number of images per identity.

Exp-3: common number of cameras and identities. Using the same number of cameras and of identities should allow the reciprocal influence between the number of images per identity and the diversity in clothing appearance between different identities to better emerge. We first observe that the additional constraint on the number of identities with respect to Exp-1 lead to a drastic reduction of the size of RandPerson (which also incurred the highest reduction in the number of identities) and to a lower extent for UnrealPerson and PersonX. UnrealPerson remained the most effective data set, and exhibited only a slight decrease in performance with respect to Exp-1. Taking into account that also RandPerson outperforms PersonX, despite containing in Exp-3 one order of magnitude less images, this seems to confirm that a realistic background and a higher amount of diversity in pedestrian appearance, as well as a better quality of the human models, are beneficial to the effectiveness of synthetic data sets, probably to a larger extent than the overall number of images.

Exp-4: common number of identities and images per identity. These two constraints also lead to a very similar number of images among all data sets (see Table IV), making them all more comparable. It is also worth noting that the size of UnrealPerson was further reduced by one order of magnitude with respect to Exp-2 and Exp-3: this time, this led to a drastic decrease in performance, probably due also to the lower number of identities per camera, and of images per identity in each camera, with respect to Exp-3 (note that in Exp-4 the constraint on the number of cameras was released, and UnrealPerson has much more cameras than the other data sets). In particular, less than one image per individual was present in each camera on average, i.e., some identities were not present in some cameras. Nevertheless, UnrealPerson exhibited the best performances on all target domains.

Exp-5: common number of cameras, identities, and images per identity. This last experiment combines all the constraints considered in the previous ones, which makes the considered data sets most similar in the corresponding factors. Note that, with respect to Exp-4, only UnrealPerson and PersonX were modified, by *reducing* the number of cameras, which resulted in an *increase* of the number of images per camera and the number of images per individual in each camera. As can be seen from Table IV), this led to an increase in performance of PersonX. Note also that UnrealPerson becomes the best data set again, as in previous experiments, although its performance is now lower. This is an interesting result that again confirms our previous observations about the relatively less relevant role of the bare data set size regarding factors such as the diversity in pedestrian appearance.

VI. Discussion and Conclusion

The results of our experimental analysis of synthetic data sets for training Re-Id models can be summarised as follows. (i) The visual quality of synthetic images, particularly the quality of human models and the presence of a realistic background, as well as the diversity of their visual features, seem to be more relevant than the bare amount of images. In other words, simply increasing the number of synthetic images is not effective if too simple human models or backgrounds are used, and if no care is taken to introduce diversity, especially in pedestrians' clothing appearance. (ii) Another relevant factor is the distribution and balance of the number of identities per camera and the number of images per identity in each camera. In particular, several images per identity should be present for each camera. In general, having a balanced distribution of such factors is a desired feature in synthetic

		Target: Market					Target: Duke					Target: MSMT				
	Training	mAP	rk-1	rk-5	rk-10	rk-20	mAP	rk-1	rk-5	rk-10	rk-20	mAP	rk-1	rk-5	rk-10	rk-20
Synthetic	SOMAset	0.9	1.8	5.6	8.6	13.1	0.6	1.4	4.0	6.4	9.2	0.2	0.6	1.9	2.8	4.4
	SyRI	4.8	12.4	25.3	32.3	40.6	3.0	8.9	17.7	21.8	27.3	1.1	5.1	10.1	13.3	17.3
	VC-Clothes	4.8	15.9	28.7	35.7	43.2	3.9	12.2	21.6	26.5	31.8	1.1	5.0	10.3	13.3	17.5
	FineGPR	11.5	31.7	46.7	53.1	59.7	11.5	27.6	40.1	45.4	50.4	2.7	11.3	19.4	24.2	29.3
	PersonX w/o BG	13.7	32.5	51.1	59.1	66.3	7.7	17.1	27.9	32.3	38.3	1.5	4.8	9.5	12.6	16.5
	PersonX	21.8	45.4	63.3	70.3	77.1	17.2	33.5	46.2	52.2	58.3	2.7	8.8	15.2	19.1	23.7
	PersonX w/ BG	22.2	48.2	64.3	70.4	76.6	17.3	33.2	47.3	54.2	60.8	2.8	9.4	15.9	19.9	24.6
	RandPerson	36.4	64.8	80.8	85.9	90.3	29.9	50.7	65.5	70.8	76.5	7.2	22.0	32.9	38.4	44.1
	UnrealPerson	41.7	70.8	82.8	87.4	90.7	41.6	64.6	76.5	81.0	84.4	11.3	30.9	42.8	48.8	54.7
Real	Market	-	-	-	-	-	30.88	53.41	67.06	72.49	77.24	6.44	19.08	29.14	34.36	40.22
	Duke	30.44	62.2	78.24	83.31	87.65	-	-	-	-	-	8.09	24.7	36.44	42.23	48.15
	MSMT	38.33	69.77	83.94	88.18	91.66	28.49	47.17	67.01	74.33	79.34	-	-	-	-	-

TABLE III

PERFORMANCES OBTAINED USING ALL AVAILABLE IMAGES OF THE CONSIDERED SYNTHETIC DATA SETS ON THREE REAL BENCHMARK ONES.

		Training data set				Target: Market					Target: Duke					Target: MSMT				
	Constraints	Name	#IDs	#images	#cam.	mAP	rk-1	rk-5	rk-10	rk-20	mAP	rk-1	rk-5	rk-10	rk-20	mAP	rk-1	rk-5	rk-10	rk-20
Exp-1	4 cameras	UnrealPerson	2800	138275	4	**39.1**	**69.4**	**82.0**	**86.0**	89.4	**38.4**	**62.0**	**74.6**	**78.9**	**82.7**	**10.3**	**29.5**	**41.2**	**46.3**	**52.4**
		RandPerson	8000	132145	4	36.4	64.8	80.8	85.9	**90.3**	29.9	50.7	65.5	70.8	76.5	7.2	22.0	32.9	38.4	44.1
		PersonX	1266	182304	4	22.4	47.9	65.1	72.2	78 .4	17.2	33.3	46.5	53.0	59.1	2.7	8.7	15.2	18.8	23.5
		VC-Clothes	512	19060	4	4.8	15.9	28.7	35.7	43.2	3.9	12.2	21.6	26.5	31.8	1.1	5.0	10.3	13.3	17.5
Exp-2	4 cameras 16 im. per ID	UnrealPerson	2799	39739	4	**37.1**	**68.4**	**80.7**	84.6	88.4	**36.5**	**61.5**	**73.8**	**77.9**	**82.2**	**10.8**	**32.0**	**43.4**	**48.9**	**54.2**
		RandPerson	8000	125326	4	35.3	63.6	80.3	**85.2**	**89.8**	30.4	52.6	66.8	71.8	76.9	7.3	21.9	33.4	38.5	44.2
		PersonX	1266	20256	4	24.5	52.3	67.2	72.6	77.9	20.4	38.6	52.4	58.3	64.0	3.8	12.3	20.0	24.7	29.9
		VC-Clothes	512	8176	4	9.3	26.9	41.7	47.7	55.0	9.4	25.4	36.5	42.8	47.9	2.8	12.2	20.1	24.5	29.5
Exp-3	4 cameras 512 IDs	UnrealPerson	512	37161	4	**34.8**	**65.6**	**79.2**	**83.3**	**87.5**	**34.1**	**58.7**	**71.3**	**76.3**	**80.7**	**9.4**	**28.4**	**39.6**	**44.9**	**50.8**
		RandPerson	512	8390	4	26.2	53.2	71.0	78.4	83.9	20.8	40.6	55.0	60.8	68.1	4.5	15.8	25.8	30.5	36.5
		PersonX	512	73728	4	19.1	42.5	59.1	66.1	73.0	16.1	31.3	44.6	51.0	57.3	2.2	7.3	13.0	16.3	20.7
		VC-Clothes	512	19060	4	4.8	15.9	28.7	35.7	43.2	3.9	12.2	21.6	26.5	31.8	1.1	5.0	10.3	13.3	17.5
Exp-4	512 IDs, 16 im. per ID	UnrealPerson	512	8192	28	**29.9**	**60.6**	**76.8**	**81.8**	**86.1**	**29.6**	**55.4**	**69.3**	**73.2**	**77.4**	**10.2**	**32.1**	**43.9**	**49.4**	**54.8**
		RandPerson	512	7996	4	25.5	53.6	71.0	77.7	84.0	21.6	41.8	57.3	63.0	68.8	4.2	15.1	24.8	29.8	35.3
		PersonX	512	8192	6	14.2	33.6	50.6	57.9	65.5	10.9	22.3	34.9	40.4	47.4	1.6	5.4	10.2	13.3	17.4
		VC-Clothes	512	8176	4	9.3	26.9	41.7	47.7	55.0	9.4	25.4	36.5	42.8	47.9	2.8	12.2	20.1	24.5	29.5
Exp-5	4 cameras 512 IDs 16 im. per ID	UnrealPerson	512	8192	4	**27.6**	**58.3**	**73.7**	**79.8**	**84.7**	**27.6**	**52.8**	**65.4**	**70.7**	**75.5**	**8.5**	**27.9**	**39.1**	**44.1**	**50.1**
		RandPerson	512	7996	4	25.5	53.6	71.0	77.7	84.0	21.6	41.8	57.3	63.0	68.8	4.2	15.1	24.8	29.8	35.3
		PersonX	512	8192	4	18.9	42.1	59.5	66.5	73.9	16.0	32.3	44.8	51.1	57.4	2.4	8.0	14.2	17.7	22.1
		VC-Clothes	512	8176	4	9.3	26.9	41.7	47.7	55.0	9.4	25.4	36.5	42.8	47.9	2.8	12.2	20.1	24.5	29.5

TABLE IV

PERFORMANCE ATTAINED ON THE THREE TARGET DOMAINS USING FIVE DIFFERENT VERSIONS OF THE SYNTHETIC DATA SETS (SEE THE TEXT FOR MORE DETAILS). BEST RESULTS FOR EACH EXPERIMENT ARE HIGHLIGHTED IN BOLD.

data sets, and certainly among the main ones that prompted their use; however, it turned out to be not achieved to the same degree in all existing data sets, and in particular in RandPerson, that presents a significantly unbalanced number of images over its four cameras (6,577 to 35,572). (iii) A further general observation that emerges from our results is that the performance attained on the MSMT data set under *all* the considered experimental settings, was always notably worse than the one on Market and Duke (see Tables III and IV). This could be expected to some extent since MSMT is the largest and more challenging benchmark data set and contains a significantly higher amount of cameras, identities and images than Market and Duke (see Table IV). On the one hand these results show that there may still be a significant gap to be filled by synthetic data sets before they can be effectively used in real-world application scenarios, Indeed, few models trained on synthetic images exhibited better or comparable performances than those trained on real ones. On the other hand, these results confirm the potential of synthetic images, capable of bridging this gap by leveraging different features, not necessarily related to image quality.

A possible limitation of our results is that they refer to a single CNN model: extending our experimental investigation to other models is an obvious and interesting follow-up. A more focused analysis of the trade-off between the total number of images and the number of images per identity in each camera would also be interesting. For instance, among the existing synthetic data sets, UnrealPerson and Synthetic18K (if camera annotations are made available) seem to be the most suitable for this investigation since they present more than 16 images per identity in each camera. Finally, the effectiveness of synthetic data sets may be improved by using, e.g., synthetic-to-real domain adaptation techniques [13], [34], even in scenarios where the target domain is unknown (in this case, benchmark data sets of real images could be used). In this same scenario, collaborative training between multiple synthetic or even real data sources could be used in order to improve generalisation capability, by exploiting and combining the features of multiple data sets. It would therefore be interesting to investigate whether and to what extent such techniques can alleviate the need for realistic human models and background pointed out above.

ACKNOWLEDGMENT

This work was supported by the projects "Law Enforcement agencies human factor methods and Toolkit for the Security and protection of CROWDs in mass gatherings" (LETSCROWD), EU Horizon 2020 programme, grant agreement No. 740466, and "IMaging MAnagement Guidelines and Informatics Network for law enforcement Agencies" (IMMAGINA), European Space Agency, ARTES Integrated Applications Promotion Programme, contract No. 4000133110/20/NL/AF.

REFERENCES

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *CoRR*, vol. abs/2001.04193, 2020.

[2] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *CoRR*, vol. abs/1610.02984, 2016.

[3] Y. Zou, X. Yang, Z. Yu, B. V. K. V. Kumar, and J. Kautz, "Joint disentangling and adaptation for cross-domain person re-identification," in *European Conference on Computer Vision ECCV*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12347. Springer, 2020, pp. 87–104.

[4] S. Zhou, Y. Wang, F. Zhang, and J. Wu, "Cross-view similarity exploration for unsupervised cross-domain person re-identification," *Neural Computing and Applications*, pp. 1–11, 2021.

[5] R. Delussu, L. Putzu, G. Fumera, and F. Roli, "Online domain adaptation for person re-identification with a human in the loop," in *25th International Conference on Pattern Recognition, ICPR*. IEEE, 2020, pp. 3829–3836.

[6] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. I. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 523–536, 2019.

[7] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis, "Looking beyond appearances: Synthetic training data for deep cnns in re-identification," *Comput. Vis. Image Underst.*, vol. 167, pp. 50–62, 2018. [Online]. Available: https://www.kaggle.com/vicolab/somaset

[8] S. Bak, P. Carr, and J. Lalonde, "Domain adaptation through synthesis for unsupervised person re-identification," in *Computer Vision - ECCV 2018 - 15th European Conference Proceedings, Part XIII*, ser. Lecture Notes in Computer Science, vol. 11217, 2018, pp. 193–209. [Online]. Available: https://github.com/swbak/SyRI

[9] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 608–617. [Online]. Available: https://github.com/sxzrt/Instructions-of-the-PersonX-dataset

[10] Y. Wang, S. Liao, and L. Shao, "Surpassing real-world source training data: Random 3d characters for generalizable person re-identification," in *The 28th ACM International Conference on Multimedia*. ACM, 2020, pp. 3422–3430. [Online]. Available: https://github.com/VideoObjectSearch/RandPerson

[11] M. Meharban, M. Sabu, and S. Krishnan, "Introduction to medical image synthesis using deep learning:a review," in *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, 2021, pp. 414–419.

[12] M. Abdolahnejad and P. Liu, "Deep learning for face image synthesis and semantic manipulations: a review and future perspectives," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5847–5880, 2020.

[13] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2019, pp. 8198–8207. [Online]. Available: https://gjy3035.github.io/GCC-CL/

[14] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," *CoRR*, vol. abs/2001.10773, 2020. [Online]. Available: https://europe.naverlabs.com/research/computer-vision/proxy-virtual-worlds-vkitti-2/

[15] N. Aranjuelo, S. García, E. Loyo, L. Unzueta, and O. Otaegui, "Key strategies for synthetic data generation for training intelligent systems based on people detection from omnidirectional cameras," *Comput. Electr. Eng.*, vol. 92, p. 107105, 2021. [Online]. Available: https://datasets.vicomtech.org/v4-osd/OSD_download.zip

[16] S. Nikolenko, *Synthetic data for deep learning*. Springer, 2021, vol. 174.

[17] R. Delussu, L. Putzu, and G. Fumerar, "Scene-specific crowd counting using synthetic training images," *Pattern Recognition*, accepted with major revision. [Online]. Available: https://github.com/lputzu/Synthetic_Crowd_Counting_Images

[18] H. K. Ekbatani, O. Pujol, and S. Seguí, "Synthetic data generation for deep learning in counting pedestrians," in *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, ICPRAM*, M. D. Marsico, G. S. di Baja, and A. L. N. Fred, Eds., 2017, pp. 318–323.

[19] C. Li, S. Ge, D. Zhang, and J. Li, "Look through masks: Towards masked face recognition with de-occlusion distillation," in *MM '20: The 28th ACM International Conference on Multimedia*, C. W. Chen, R. Cucchiara, X. Hua, G. Qi, E. Ricci, Z. Zhang, and R. Zimmermann, Eds., 2020, pp. 3016–3024.

[20] E. Yaghoubi, D. Borza, S. V. A. Kumar, and H. Proença, "Person re-identification: Implicitly defining the receptive fields of deep learning classification frameworks," *Pattern Recognit. Lett.*, vol. 145, pp. 23–29, 2021.

[21] P. Shamsolmoali, M. Zareapoor, E. Granger, H. Zhou, R. Wang, M. E. Celebi, and J. Yang, "Image synthesis with adversarial networks: A comprehensive survey and case studies," *Inf. Fusion*, vol. 72, pp. 126–146, 2021.

[22] I. J. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," *CoRR*, vol. abs/1701.00160, 2017.

[23] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision, ICCV*. IEEE Computer Society, 2017, pp. 2242–2251.

[24] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 8789–8797.

[25] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *4th International Conference on Learning Representations, ICLR*, Y. Bengio and Y. LeCun, Eds., 2016.

[26] Unity Technologies, "Unity," https://unity.com/.

[27] Epic Games, "Unreal engine," https://www.unrealengine.com/en-US/.

[28] "Grand Theft Auto V, Script Hook V," http://www.dev-c.com/gtav/scripthookv/.

[29] M. Community, "MakeHuman: Open Source Tool for Making 3D Characters," http://www.makehumancommunity.org, 2020.

[30] Mixiamo, https://www.adobe.com/it/wam/fuse.html.

[31] "Blender," https://www.blender.org/.

[32] F. Wan, Y. Wu, X. Qian, Y. Chen, and Y. Fu, "When person re-identification meets changing clothes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops*. Computer Vision Foundation / IEEE, 2020, pp. 3620–3628. [Online]. Available: https://wanfb.github.io/dataset.html

[33] T. Zhang, L. Xie, L. Wei, Z. Zhuang, Y. Zhang, B. Li, and Q. Tian, "Unrealperson: An adaptive pipeline towards costless person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation / IEEE, 2021, pp. 11 506–11 515. [Online]. Available: https://github.com/FlyHighest/UnrealPerson

[34] S. Xiang, Y. Fu, G. You, and T. Liu, "Unsupervised domain adaptation through synthesis for person re-identification," in *IEEE International Conference on Multimedia and Expo, ICME*. IEEE, 2020, pp. 1–6.

[35] ——, "Taking a closer look at synthesis: Fine-grained attribute analysis for person re-identification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3765–3769.

[36] S. Xiang, G. You, M. Guan, H. Chen, F. Wang, T. Liu, and Y. Fu, "Less is more: Learning from synthetic data with fine-grained attributes for person re-identification," *CoRR*, vol. abs/2109.10498, 2021. [Online]. Available: https://github.com/JeremyXSC/FineGPR

[37] O. C. Uner, C. Aslan, B. Ercan, T. Ates, U. Celikcan, A. Erdem, and E. Erdem, "Synthetic18k: Learning better representations for person re-id and attribute recognition from 1.4 million synthetic images," *Signal Process. Image Commun.*, vol. 97, p. 116335, 2021. [Online]. Available: https://hucvl.github.io/synthetic18k/

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *International Conference on Computer Vision and Pattern Recognition,*, 2016, pp. 770–778.

[39] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[40] Z. Zhang, J. Wu, X. Zhang, and C. Zhang, "Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project," *CoRR*, vol. abs/1712.09531, 2017.

[41] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *CVPR*, 2018, pp. 79–88.