Image and Vision Computing 163 (2025) 105753

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Synthetic data sets for person Re-Identification: A critical analysis☆

Rita Delussu [a],*, Lorenzo Putzu [b], Fadi Boutros [c], Carmen Bisogni [d], Naser Damer [c], Giorgio Fumera [b]

[a] Department of Engineering, University of Sassari, Via Vienna 2, Sassari, Italy
[b] Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi, Cagliari, Italy
[c] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[d] Department of Computer Science, University of Salerno, Salerno, Italy

ARTICLE INFO

Keywords:
Person Re-Identification
Generalisation capability
Synthetic training data
Visual variations
Photorealism

ABSTRACT

Supervised methods for person Re-Identification (Re-Id) need extensive manual annotation, limiting data set size and the resulting generalisation capability to unseen target data. Unsupervised methods avoid manual annotation but typically attain a lower performance. Synthetic training data can mitigate these issues, as they allow generating large data sets encompassing more representative variations in visual factors such as background scenes and pedestrian appearance without requiring manual annotation and without privacy issues arising from recent regulations. Existing synthetic data sets vary in size, diversity of human models, camera views, backgrounds, as well as photorealism. It is, however, not yet clear how all such factors affect Re-Id performance. We conduct a comprehensive and systematic analysis and experimental evaluation of existing synthetic data sets, to understand how the main factors characterising them affect the generalisation capability to real data. Our results provide useful guidelines towards developing effective synthetic data sets for Re-Id.

1. Introduction

Collecting data sets for Computer Vision (CV) tasks is often costly and labour-intensive, particularly for manual annotation, which can take months and require significant funding [1]. Consequently, real benchmark data sets tend to be limited in both size and representativeness.

Synthetic data has emerged as a promising solution to these challenges and is increasingly used in CV and other fields [2]. The role of synthetic data has evolved significantly, especially with the rise of Deep Learning (DL), and they now primarily serve as training data—either alone or alongside real images [1,2]. This approach aims to expand the training data set, enhancing performance and reducing the risk of over-fitting.

In particular, synthetic data offer several advantages over real data: (i) they allow for generating large-scale data sets, encompassing more representative and controlled variations in factors such as illumination, weather conditions and viewpoints [3], and enabling the simulation of application-specific scenarios and potentially enhancing generalisation and mitigating overfitting [4]; (ii) they come with automatically generated annotations, which is particularly beneficial for tasks requiring detailed supervision, such as pixel-wise or region-level annotations [3]; (iii) they raise fewer privacy concerns, as all samples are artificially created [1]. This is especially relevant in light of increasingly strict privacy regulations (e.g., GDPR) [5–7], which impose limitations on collecting identifiable personal data [7].

Thanks to these advantages, synthetic data have been widely adopted in diverse domains such as remote sensing [8], facial recognition [9], person re-identification [10], crowd counting [3], medical imaging [11], handwriting recognition [12], defect detection or quality inspection [13], and autonomous driving [14]. Among these domains, Person Re-Identification (Re-Id) stands out as the one with the highest number of synthetic data sets. This is likely due to its relevance in security-related applications, the persistent challenges faced in real-world deployments, and the strict privacy constraints that limit the collection of real identity data [2,6]. Furthermore, prior studies have demonstrated that certain synthetic data sets can exhibit stronger generalisation capabilities than real data sets in cross-data set evaluation settings [15], highlighting their potential value in Re-Id research.

While most synthetic data sets for Re-Id are generated using computer graphics engines, a wide range of approaches have been used to create virtual environments and human models, resulting in data sets that differ significantly in terms of photorealism, as well as in

☆ This article is part of a Special issue entitled: 'XAISynData' published in Image and Vision Computing.

* Corresponding author.
E-mail address: rdelussu@uniss.it (R. Delussu).

https://doi.org/10.1016/j.imavis.2025.105753
Received 13 March 2025; Received in revised form 22 September 2025; Accepted 26 September 2025
Available online 17 October 2025

the number of images, identities, cameras, and scenes [2]. Despite the potential influence of these factors on Re-Id model performance, there is not yet a clear understanding of their actual impact. Indeed, existing empirical evaluations are typically limited to the original papers that introduce each data set, where the focus is primarily on demonstrating the effectiveness of a *single* proposed data set, without a broader comparison across other synthetic alternatives. Moreover, such studies usually included the investigation of at most a *single* factor among the above-mentioned ones, such as the number of identities [16], cameras [16] or scenes [17], or the impact of viewpoint variations [18,19], without jointly considering their interaction or cumulative impact on generalisation.

As a first step toward addressing the aforementioned research gap, this article makes the following contributions:

- a comprehensive analysis and empirical evaluation of existing synthetic data sets for Re-Id is conducted, focusing on the traditional RGB image setting;
- a systematic investigation of the impact of key factors on the generalisation performance of Re-Id models trained on synthetic data and tested on real benchmarks is carried out, focusing in particular on training set size, number of training identities, cameras, and scenes, as well as their combinations;
- actionable insights to guide the design and informed selection of synthetic data sets for Re-Id are provided.

This study extends our preliminary work [15] by including additional data sets and conducting a broader set of experiments on the effects of the number of images, cameras, identities, images per camera, and images per identity per camera.

This work is organised as follows. Section 2 provides an overview of Re-Id challenges, synthetic image generation techniques for Re-Id, and analyses of synthetic data sets performed by other authors. In Section 3, we outline existing synthetic data sets, compare them to benchmark data sets of real images, and critically examine their factor distributions and alignment with real-world scenarios. Section 4 presents our empirical evaluation, the experimental setup and our results, while a related discussion is provided in Section 5. Conclusions and future directions are presented in Section 6. All the essential code needed to analyse the data sets and recreate the subsets used in our experiments is available at the following link.

2. Related work

In this section, we first provide background on Re-Id, outlining the main tasks, challenges, and commonly adopted solutions. We then review the literature on synthetic image generation, with a focus on the techniques employed across various CV tasks and, in particular, in the context of Re-Id. Finally, we discuss prior analyses conducted by other authors that examine the quality, advantages, and limitations of existing synthetic data sets.

2.1. Background on person Re-Identification

Re-Id is inherently a *cross-scene* task since it aims at matching the individual in the query image, typically selected manually by the user from one camera view, with individuals acquired by *different* and non-overlapping camera views. Accordingly, benchmark data sets of real images are made up of bounding boxes of pedestrians, mostly manually drawn, with at least two images of each identity from two or more different cameras [20]. Moreover, images of each identity need to be manually annotated with a consistent ID label [20], a process that demands substantial manual effort.

In recent years, traditional Re-Id based on RGB images has been complemented by even more challenging settings, the main ones being *clothing-independent* and multi-modal Re-Id. Clothing-independent Re-Id [21], also referred to as cloth-changing or long-term Re-Id, typically assumes longer time spans than traditional Re-Id, during which clothing changes may occur. Multi-modal Re-ID, on the other hand, leverages multiple sensing modalities, such as combining RGB and infrared (IR) images, to enable person matching under varying lighting conditions, particularly at night [22]. Both settings introduce new challenges that stem from the need to learn features invariant to either clothing changes or sensor modalities.

Across all three Re-Id settings (traditional, long-term, and multi-modal Re-Id), the most effective state-of-the-art approaches are based on supervised learning. While these methods often achieve strong performance within the same data set [20], they typically suffer from limited generalisation capability to unseen target domains, with significant performance drops when evaluated on different data sets [15,20]. This problem is a specific manifestation of the well-known *domain shift* (DS) phenomenon, which arises when a model trained on a given *source* domain is applied to a different, albeit related, *target* domain [23].

Several approaches have been proposed to address DS, including supervised and unsupervised domain adaptation or domain generalisation methods. The latter aims at improving the generalisation capability of the model on any target domain using several source domains for training, for example, by merging multiple real data [24] or by including synthetic data [25]. Although these strategies are applied to different tasks, both studies point out aspects relevant to this work. Specifically, they note that increasing the number of cameras [24] or the number of synthetic training samples [25] does not necessarily correspond to performance improvements. The effectiveness of domain generalisation methods mainly depends on how much the source domains represent the target one. In this context, synthetic data have emerged as a possible alternative to real data sets, in different CV tasks, even demonstrating better generalisation performance than real data sets [15].

2.2. Synthetic data generation for computer vision tasks

The techniques employed to generate synthetic images differ widely, depending not only on the target application domain but also on the specific vision task, such as classification, detection, or segmentation. Broadly speaking, existing image generation methods can be categorised into three main approaches, each suited to different use cases: generative models, computer graphics engines, and image composition techniques.

Generative models can be categorised into Generative Adversarial Networks (GANs) [26] and Diffusion Models (DMs) [27]. The GAN-based approach employs a generator-discriminator framework trained adversarially to produce synthetic images. Key types include CycleGANs [28], StarGANs [29], and Deep Convolutional GANs (DCGANs) [30], widely used to generate synthetic data, alongside basic data augmentation, across a variety of CV tasks [4,31]. In contrast, DMs generate data by gradually adding noise to input images and then learning to reverse this process [32]. Although more recent than GANs, DMs have already proven effectiveness in several CV tasks, including image classification, image-to-image translation, and text-to-image translation [33].

Computer graphics engines are currently capable of generating images and videos with a high level of photorealism. In addition to commonly used rendering tools such as Blender,[1] Adobe Fuse CC[2] (Adobe), and MakeHuman,[3] recent works have also employed game-oriented platforms like Unity[4] and Unreal Engine[5] (Unreal for short) as

[1] https://www.blender.org/
[2] https://www.adobe.com/it/wam/fuse.html
[3] http://www.makehumancommunity.org
[4] https://unity.com/
[5] https://www.unrealengine.com/en-US/

well as the Script Hook V library, which allows users to employ (offline) the Grand Theft Auto V video game (GTAV for short).[6] Computer graphics engines have been used for different purposes, to generate, e.g.: bounding boxes of pedestrians from different scenes for Re-Id [6,19,21,34]; videos of normal and abnormal events for behaviour analysis; scenes for object detection and tracking, as well as for crowd counting [2]; benchmark data sets for behaviour analysis, tracking, Re-Id and face recognition; scene-specific images or more challenging views for object and pedestrian detection; more representative training images for crowd counting (e.g., including different weather and lighting conditions) [2].

Image composition methods generate synthetic data by combining elements from one or more images. Techniques range from simple image pasting — overlaying new items to a real or synthetic background — to more advanced image fusion techniques that merge specific features from multiple *real* images (e.g., pedestrian or face images). Simple pasting has been applied to improve the robustness in tasks like object detection, pedestrian detection, and face recognition by adding varied overlaps (e.g., objects, people, face accessories) or by creating diverse, scene-specific data sets for tasks like crowd counting [2]. Image fusion is primarily used in Re-Id [35] and face recognition [36]. For Re-Id, real images are fused by interpolating features or mixing elements like body parts and backgrounds [35]. In face recognition, synthetic faces from 3D models are combined with real backgrounds to simulate realistic pose and lighting [36].

2.3. Synthetic data generation for person Re-Identification

In all three above-mentioned Re-Id settings (traditional, long-term, and multi-modal Re-Id), the use of artificially generated images has been explored as a viable data source. However, to our knowledge, no publicly available synthetic data sets currently exist for multi-modal Re-Id. In contrast, for both traditional and long-term Re-Id, several synthetic data sets have been released. Most of them have been created using computer graphics engines, since they allow users to generate a wide variety of human models with diverse attributes such as skin, eye and hair colour, gender, height, weight, and clothing colour and style.

While some computer graphics engines, like MakeHuman, are particularly effective for modelling detailed human models, they lack support for building complete scenes or realistic backgrounds. Conversely, platforms like Unity and Unreal Engine allow users to design rich virtual environments, including urban streets, natural landscapes, subway stations, and parks. As a result, some synthetic Re-Id data sets have been developed using only Unity or Unreal, whereas others have combined them with tools like MakeHuman to enrich both character diversity and environmental realism. Video game platforms such as GTAV have also been used [21,34]. Further details on the existing synthetic data set are given in Section 3.2.

2.4. Previous analyses of synthetic data sets for person Re-Identification

In the following, we summarise the analyses conducted in previous works that investigated the impact of various factors on Re-Id performance. As noted in Section 1, existing Re-Id data sets have typically been analysed only within the original studies that introduced them, often with minimal comparisons to alternative data sets. These analyses have primarily focused on two key factors: the number of identities [10,16–18,37] and viewpoint variations inside the training set [18,19,38]. Among them, the analysis of the number of identities emerges as the primary focus in several works. Most of the reported results indicate a positive correlation between the number of identities in the training set and the performance attained on real data [10,16]. On the other hand, others highlighted a saturation point beyond which performance no longer improved [10,37], and in some cases, even degraded, potentially due to overfitting [17].

Viewpoint variations have been investigated in two works only [18,19], with contrasting findings. In [19], a training set covering all viewpoints was found to be the most beneficial, with side views (left/right) proving more effective than front or back views. In contrast, [18] reported that the four principal views (0°, 90°, 180°, 270°) lead to better performances than the other ones [18]. Additionally, [39] highlighted the importance of viewpoint also at test time, showing that side views yield higher retrieval accuracy than front or back views, consistent with the findings of [19].

Other authors investigated additional aspects such as the number of images per identity and the number of cameras [16], and the number of scenes [17]. Results indicate that increasing either the number of cameras or images per identity tends to improve model generalisation. However, an excessive number of cameras (e.g., over 40) was found to cause a performance drop [16]. Similarly, increasing the number of scenes was found to be beneficial up to a certain point, beyond which a saturation effect was observed [17].

Finally, some authors argued that higher photorealism might reduce the synthetic-to-real gap, thus improving the performance of Re-Id models trained on synthetic data [10,16]; the influence of the photorealism is, however, still an open question.

In contrast to prior work, which typically investigated individual factors in isolation and within the context of a single synthetic Re-Id data set, our study conducts a comprehensive, systematic evaluation across all currently available synthetic data sets for Re-Id. We assess key factors, including the number of identities, images per identity, images per camera, cameras per scene, scene diversity, and degree of photorealism, within a unified framework. Importantly, unlike some studies that combine real and synthetic data for training (e.g., [40]), we exclusively rely on synthetic data during training. This design choice allows us to isolate and quantify the individual and combined effects of synthetic data characteristics on model performance without interference from real data.

3. Real and synthetic data sets for person Re-Identification

In this section, we first introduce the main real benchmark data sets for Re-ID. We then describe existing synthetic data sets and provide a comparison and a critical analysis of them, discussing their main features and limitations.

3.1. Real benchmark data sets

Collecting a Re-Id data set from real images or video frames involves a considerable manual annotation effort, including the extraction of pedestrian bounding boxes and the consistent labelling of instances belonging to the same identity [20]. Consequently, existing data sets suffer from two primary shortcomings: (i) their size is relatively limited with respect to real-world application scenarios [20] and (ii) they exhibit a limited variability in factors such as weather conditions, lighting, scale, viewpoints, and scene backgrounds [41]. For instance, many data sets primarily consist of daytime images captured by RGB cameras, overlooking scenarios that require coverage during adverse weather or lighting conditions [22]. The most commonly used large-scale data sets made up of RGB images for Re-Id are Market-1501 [42], DukeMTMC [43] and MSMT17 [23], hereafter referred to as Market, Duke and MSMT. We intentionally exclude smaller data sets as they are no longer deemed challenging and are seldom utilised as benchmarks. Nonetheless, we include the OccludedReID data set [44], hereafter referred to as OccReId, in our evaluation due to its realistic and challenging nature, despite its limited size. It contains images of pedestrians with varying degrees of occlusion, which is a common and critical issue in real-world Re-Id scenarios.

[6] http://www.dev-c.com/gtav/scripthookv/

Table 1
Statistics of synthetic and real data sets: number of identities (#IDs), images, cameras and scenes, as reported in the respective papers (published data), and evaluated on the downloaded versions (including the min, average and maximum number of images per identity #imgsXID and the min, average and maximum number of identities per camera #IDsXcam) in the access date reported in the right-most column. The symbol '*' indicates that camera information is not available. The occlusions/whole body (1-5) information is considered as cameras.

	Data set	Published data					Downloaded data						
		#IDs	#images	#cam	#scenes	Year	#IDs	#images	#cam	#scenes	#imgsXID	#IDsXcam	access date
Synthetic	UnrealPerson [10]	3,000	120,000	34	4	2021	2,800	1,255,297	28	4	48/448/911	2,009/2,636/2,799	2021/12
	PersonX [19]	1,266	273,456	6	6	2019	1,266	273,456	6	6	216/216/216	1,266/1,266/1,266	2021/09
	ClonedPerson [45]	5,621	887,766	24	6	2022	4,826	763,953	24	6	59/198/1,222	682/3,160/4,826	2022/10
	SyRI [40]	100	1,680,000	–	–	2018	100	56,000	280	–	560/560/560	100/100/100	2021/10
	RandPerson [17]	8,000	1,801,816	19	11	2020	8,000	132,145	19	11	11/17/19	6,577/7,599/8,000	2021/10
	FineGPR [18]	1,150	2,028,600	36	9	2021	1,150	2,028,600	324	9	1,764/1,764/1,764	870/894/916	2022/12
	WePerson [16]	1,500	4,000,000	560	14	2021	900	833,458	880	22	153/926/1,763	21/263/359	2023/01
Real	Market1501 [42]	1,501	32,668	6	–	2015	1,501	32,668	6	–	–	–	2017/12
	DukeMTMC [43]	2,834	–	8	–	2016	1,404	36,411	8	–	–	–	2017/12
	MSMT17 [23]	4,101	126,441	15	–	2018	4,101	124,068	15	–	–	–	2019/04
	OccludedReId [44]	200	2,000	–	–	2018	200	2,000	5*	–	–	–	2025/05

Market [42] is a collection of 32,668 bounding boxes of 1,501 identities acquired from 6 cameras placed in front of a supermarket. They are divided into 751 identities for training (12,936 images) and 750 identities for testing, corresponding to the remaining images, which are further subdivided into a gallery set (15,913 images) and a query set (3,368 images). **Duke** [43] contains 36,411 bounding boxes of 1,404 identities captured from 8 cameras placed on a university campus. They are subdivided into 702 identities for training (16,522 images) and 702 identities for testing, corresponding to the remaining images, which are further subdivided into a gallery set (17,661 images) and a query set (2,228 images). **MSMT** [23] is made up of 126,441 bounding boxes of 4,101 identities captured from 15 cameras. They are split into 1,401 identities for training (32,621 images) and 2,700 identities for testing, further subdivided into a gallery set (82,161 images) and a query set (11,659 images). **OccReId** [44] contains 2000 images of 200 identities. The data set is composed of occluded images (5 images per identity) and whole body images (5 images per identity). The occluded images are used as query set, whereas the whole body images are used as gallery set [44]. The main statistics of these four data sets are summarised in the last rows of Table 1.

3.2. Synthetic data sets

Several synthetic data sets have been proposed so far for Re-Id task: Synthetic18k [6], PersonX [19],RandPerson [17], Virtually Changing-Clothes (VC-Clothes) [21], GTA Person Re-Id (GPR) [34], GPR+ [38], FineGPR [18], Synthetic person Re-Id (SyRI) [40],SOMAset [37], UnrealPerson [10], WePerson [16], ClonedPerson [45]. They have been mostly generated using computer graphics engines such as Unity, Unreal, and Blender, often in combination with human model generators like MakeHuman.

Synthetic18K and PersonX have been generated using Unity.

Synthetic18K [6] contains four scenes (three outdoor and one indoor) under different weather conditions and times of day and 18,306 human models (identities), for a total of 1,408,600 bounding boxes of pedestrians.

PersonX [19] contains three different scenes. Each pedestrian image is rendered in all of them, and additionally, also on three uniform backgrounds with different colours. This data set focuses on viewpoint changes, i.e., each pedestrian is captured from different viewpoints. To increase diversity in pedestrian appearance, the 1,266 identities of this data set (547 females and 719 males) present different ages, body shapes, skin colours, etc. In total, 273,456 bounding boxes of pedestrians are present.

VC-Clothes, GPR and WePerson have been generated using GTAV.

VC-Clothes [21] contains four scenes (street, gate, parking lot and a natural scene) under different illumination conditions, for a total of 19,060 bounding boxes of 512 different identities, representative of different ages, body shapes, etc. Contrary to most of the other synthetic data sets, VC-Clothes focuses on *clothing-independent* Re-Id, and therefore, images of each identity differ in clothing appearance and attributes.

GPR [34] is composed of 443,352 images of 754 identities generated using 12 cameras, under 12 weather conditions (e.g., cloudy and foggy), 8 different illuminations (e.g., afternoon and midnight), and 26 scenes, such as beach, street, school and mall. Two extensions of GPR containing a larger number of identities and of images were also generated, GPR+ [38] and FineGPR. In particular, **FineGPR** [18] differs from GPR and GPR+ in that its pedestrian images contain more fine-grained details, mostly attributes of human models such as upper- and lower-body clothing colours and several accessories such as hats, bags, etc. It contains over 2 million images and 1,150 identities acquired from 36 cameras. Moreover, nine scenes (e.g., park and street) were generated under several illumination conditions (e.g., sunny and cloudy).

WePerson [16] is composed of 4 million bounding boxes of 1,500 identities generated in 14 scenes (10 outdoor and 4 indoor), under 40 different viewpoints (cameras) per scene, for a total of 560 cameras. The images have been generated using seven weather conditions (e.g., cloudy and snowy) and seven illumination conditions (e.g., afternoon and night). Moreover, several occlusions among different pedestrians or objects have been simulated.

Other synthetic data sets have been generated using different software combinations.

SyRI [40] was generated using two computer graphics software, Unreal and Adobe. It contains 100 identities and 1,680,000 images generated in 140 scenes under more than 100 different illumination conditions, and 2 cameras per scene (see Section 3.3 for more details).

The computer graphics software MakeHuman was used to generate RandPerson, ClonedPerson, SOMAset and UnrealPerson, together with Unity, Blender and Unreal, respectively.

RandPerson [17] comprises about 1,8 million bounding boxes of 8,000 identities. The corresponding human models were generated using MakeHuman. To create different pedestrian clothes, a combination of more than 600 colours and 16 patterns was used, which brought about 10,000 texture maps. Pedestrians are rendered in 11 scenes (8 outdoor and 3 indoor, e.g., gym and urban), which were generated using Unity under different illumination conditions. More than one camera view was considered for some scenes, resulting in 19 camera views.

ClonedPerson [45] consists of 887,766 bounding boxes of 5,621 individuals rendered in 6 scenes generated using Unity, from 4 camera views per scene, for a total of 24 cameras. The human models are generated using MakeHuman. Their clothes were generated by cloning the outfit of real images of persons extracted from the DeepFashion2 [46] data set, which contains different images of popular clothing categories from both commercial shopping stores and consumers.

SOMAset [37] focuses on clothing-independent or long-term Re-Id, analogously to VC-clothes. It contains 50 identities, each one rendered

Fig. 1. Two sample images depicting two identities on different backgrounds from each synthetic data set for Re-Id. From left to right, top to bottom: SOMAset [37], SyRI [40], PersonX [19], GPR [34], RandPerson [17], VC-Clothes [21], FineGPR [18], Synthetic18k [6], UnrealPerson [10], WePerson [16], ClonedPerson [45].

with 8 different types of clothes; each of the resulting 400 subject-clothing combinations is rendered from 250 different cameras, with a different pose for each orientation.

UnrealPerson [10] contains a total of 120,000 bounding boxes generated in 4 scenes (three urban outdoor scenes and one indoor), from 34 cameras. Each scene was generated using Unreal, under different illumination conditions. The bounding boxes depict 3,000 human models generated using MakeHuman, with considerable variability in the types of clothes (more than 200) and, in some cases, different accessories (e.g., masks, glasses and hats).

Examples of images from the above synthetic data sets are shown in Fig. 1, while the main statistics are summarised in the first rows of Table 1.

3.3. Critical analysis

In this section, we analyse and discuss the strengths and weaknesses of synthetic data sets for Re-Id, with particular attention to their structure and suitability for training effective models.

It is worth knowing that we excluded the following synthetic data sets from our analysis (and consequently from our experiments) due to specific limitations:

- **VC-Clothes** and **SOMAset**, since they are designed for clothing-invariant or long-term Re-Id, which diverges from the focus of this study;
- **GPR** and **GPR+**, since they were unavailable for download at the time of writing [34,38];
- **Synthetic18k**, since the image filenames include only identity labels, lacking camera metadata essential for our experimental design[7];
- **SyRI**, since it presents too few identities (100) to support the training of generalisable Re-Id models [15]. Furthermore, discrepancies exist between the paper's description [40] and the actual data set structure, which presents only 2 images per identity per camera, making it unsuitable for our purposes.

During our analysis, we observed several inconsistencies between the versions of the data sets described in the original papers (hereafter *published* versions) and those actually available for download (hereafter *downloaded* versions). These discrepancies concern various aspects, such as the number of images, identities, cameras, or scenes. To account for these differences and ensure transparency in our evaluation, Table 1 reports statistics for both the published and downloaded versions of each data set. For the latter versions, we also reported more detailed statistics related to the number of images per identity and the number of identities per camera, useful for our experimental evaluation.

With the only exception of PersonX [15], the downloaded versions present a different number of identities, images, cameras or scenes with respect to published versions. UnrealPerson is the only data set whose downloaded version contains a larger number of images than the published one (about 10 times higher, 1.25M vs 120K). The other downloaded versions contain a smaller number of images. In particular, the downloaded versions of RandPerson and SyRI are small subsets of the published versions, containing about 7% and 3% of their images, respectively. Interestingly, these subsets turn out to be the same subsets used for the experiments in the respective papers [17,40]. We point out that using only a small fraction of a large synthetic data set may seem counterintuitive. For RandPerson, this choice was motivated by the aim of reducing redundancy and training time, although the source of redundancy was not specified by the authors. These observations further motivate the analysis we will present in this paper, aimed at better understanding, among other issues, how data set size interacts with other factors in determining the performance of Re-Id models trained on synthetic data.

While synthetic datasets often appear large in terms of image count, several issues undermine their effectiveness:

- **low number of identities**, in some cases lower than real-world counterparts, in datasets such as SyRI, FineGPR, PersonX;
- **sparse coverage**, with very few average images per identity per camera, in datasets such as WePerson and FineGPR (almost one in the case of WePerson);
- **incomplete cross-camera representation**, since not all identities are visible in every camera, in datasets such as WePerson, UnrealPerson, ClonedPerson, and FineGPR.
- **unrealistic viewpoints**, since they employ camera perspectives almost always perpendicular to the ground or unusually close to the subjects, which differ significantly from standard CCTV angles in real-world surveillance.

A further issue is that in many cases there is an unclear difference between viewpoints, cameras and scenes, which translates into statistics that differ in terms of the number of cameras or the number of identities in each camera. Indeed, we refer to "viewpoint" as the position or perspective from which a camera captures a given scene. Accordingly, a fixed camera corresponds to a single viewpoint, which means that the number of viewpoints is equal to the number of fixed

[7] No documentation nor the authors clarified this issue.

cameras. Of course, as in real data sets, there can be several fixed cameras capturing the same scene from different positions, resulting in the same number of different viewpoints. However, for some synthetic data sets, the presence of a camera with *different viewpoints* is mentioned. While this could be interpreted as simulating PTZ (Pan-Tilt-Zoom) cameras, it becomes unrealistic when the number of viewpoints is as high as 36 and 40 for FineGPR and WePerson, respectively, and when such viewpoints correspond to capturing the same individual from 360 degrees. Such data sets can be useful for making targeted investigations related to viewpoints, but they become too dispersive for training effective Re-Id models, as demonstrated in [15].

4. Experimental evaluation

This section describes the objectives of our experimental evaluation, the experimental settings and the main results.

4.1. Objectives of our evaluation

Our aim is to systematically evaluate how different characteristics of existing synthetic data sets influence a Re-Id model's ability to generalise to real-world data. Building on our critical analysis in Section 3.3, we have identified the following key factors for investigation:

- **Data set size.** One of the potential advantages of data synthesis is that it allows for generating a much larger number of images than those affordable for real data sets. The statistics in Table 1 confirm that most synthetic Re-Id data sets exceed one of the largest benchmark data sets (MSMT), up to one order of magnitude. It is, however, interesting to investigate whether and to what extent this substantial increase in data set size leads to better performance or if smaller, well-curated data sets might be equally effective.
- **Number of identities.** Unlike data set size, synthetic data sets for Re-Id do not consistently feature more identities than real-world benchmarks (see Table 1). This raises an interesting question about the role of identity count: it is unclear whether it has been considered less critical or if it may be due to the resource cost of generating diverse 3D human models. Investigating this factor could reveal its true impact on model performance and generalisation.
- **Number of images per identity.** Table 1 shows that the number of images per identity varies greatly across synthetic data sets, which, instead, is another factor that could significantly affect how well a model learns appearance variations of the same individual.
- **Scene and camera variability.** Since Re-Id is inherently a cross-scene task, one may think that the generalisation capability of a Re-Id model can benefit from an increasing number of scenes (e.g., image background) as well as from their variability (e.g., lighting conditions and different camera views of the same scene). To investigate this complex factor, we carried out experiments by varying the **number of scenes**, the **number of cameras per scene** (i.e., the camera angles and perspectives), and the **number of images per identity and per camera**.

Ideally, to assess the influence of these factors, we would isolate each factor and train a Re-Id model on modified versions of a given synthetic data set, where specific values are assigned to each factor (such as data set size or number of identities). However, this approach is complicated since many of these factors are interdependent. For instance, reducing the data set size may imply decreasing either the number of identities or the number of images per identity or both. This necessitates choices on how to control each factor to keep the analysis manageable. Additional challenges include the differences among synthetic data sets in each of the considered factors and the unbalanced distribution of certain factors within synthetic data sets, shown in Table 1. For example, specific identities may appear only within certain camera subsets, and the number of cameras can vary across different scenes. These variations make it difficult to establish direct comparisons across data sets.

Taking into account these challenges, we opted to analyse each synthetic data set individually, but applying consistent criteria across all of them for the factors under investigation. Specifically, we followed two main criteria: (i) using balanced values of factors such as the number of images per identity and the number of cameras per scene; (ii) including only identities that appear in all the selected cameras. Following these criteria, we designed our analysis as follows:

1. we retained all the available **scenes**;
2. we selected the maximum and balanced subset of **cameras per scene**;
3. from such a subset of cameras per scene, we iteratively selected 1 to 6 cameras per scene (i.e., 4, 8, ..., 24 cameras in total);
4. for each subset of cameras, we discarded the identities that do not appear in each camera;
5. for each of the remaining **identities** we iteratively selected an increasing number of **images per camera** starting from 1 image to 6 (the greatest common value among the considered data sets).

Applying these criteria indirectly constrains the values of the other key factors like data set size, number of identities, and images per identity, ensuring consistency without additional adjustments. For benchmarking, we also trained models on the full, unconstrained versions of each synthetic data sets. Finally, we carried out a comparison among all the available synthetic data sets after carefully selecting a subset from each of them to make the comparison as fair as possible (see Section 5).

4.2. Experimental settings

As explained in Section 3.3, the synthetic data sets considered in our experiments are FineGPR [18], UnrealPerson [10], ClonedPerson [45], PersonX [19], WePerson [16] and RandPerson [17]. Unfortunately, RandPerson and WePerson were not fully compatible with our experimental criteria defined above. Specifically, neither data set contains images of every identity across all cameras, and WePerson also lacks sufficiently large subsets needed for our analysis (i.e., any possible subset contains less than 100 identities). Consequently, these data sets were only included in the final comparison with the other synthetic data sets (see Section 5), rather than being part of the full experimental setup.

We carried out cross-data set and cross-domain (synthetic-to-real) experiments, i.e., we used subsets of each synthetic data set (see Section 4.1) as the training set, and each of the benchmark, real data sets mentioned in Section 3, namely Market [42], Duke, MSMT [23] and OccReId [44], as the testing set.

We would like to point out that we did not mix real and synthetic data sets in our training, nor did we retrain or fine-tune the Re-Id models on any real images. Our goal was to isolate and rigorously assess the generalisation capabilities provided by synthetic data sets alone. Including real images during training, while common in other works, would have made it difficult to disentangle the impact of the synthetic data from that of the real samples, thereby undermining the clarity of the analysis.

Unless otherwise specified, in all the experiments we used a ResNet-50 [47] model, which is one of the most common for the Re-Id task, as a feature extractor. The model was trained in a classification setting using a weighted combination of cross-entropy and triplet loss, to foster the separation between features of different identities. For optimising the loss function, we used Stochastic Gradient Descent, with a momentum of 0.9, a learning rate of 0.00035, and a weight decay of 5×10^{-4}.

To demonstrate the generalisability of our findings and provide further support for the robustness and applicability of our conclusions across various architectural paradigms, we extended the final comparison presented in Section 5 by including results obtained using a Vision Transformer (ViT) [48] as a feature extractor. For efficiency, we adopted a small-size variant of ViT (ViT-small) to reduce training time. The model was trained in a classification setting using a weighted combination of triplet loss and ID loss, optimised with SGD with a momentum of 0.9, a weight decay of 1×10^{-4} and a learning rate of 0.005.

To reduce overfitting, during training of both models, we employed several image augmentation techniques, including horizontal flipping, random cropping, random erasing, and padding.

During inference, for ResNet-50, the output of the global average pooling layer (preceding the fully connected layer) was extracted as the feature vector representation for each input image, while for ViT-small, the output of the CLS token was used. In both cases, the dissimilarity between query and gallery images was then measured using the Euclidean distance between their corresponding feature vectors.

To evaluate performance, we considered two common metrics, i.e., mean Average Precision (mAP):

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q} \mathrm{AveP}(q)$$

where Q is the total number of queries, and $\mathrm{AveP}(q)$ is the average precision for a given query q, and Cumulative Matching Curve (CMC) at rank 1 (R1), rank 5 (R5) and rank 10 (R10):

$$CMC(k) = \sum_{R=1}^{k} P(R)$$

where $P(R)$ is the fraction of queries for which the gallery image of the correct identity (the highest-ranked image, if multiple exist) is found at rank R, for k equal to 1, 5 and 10, respectively. It is worth highlighting that the mAP measure gives a more comprehensive assessment of a Re-Id system's performance, as it considers the precision across all ranks, whereas the CMC curve only accounts for the top-ranked image of the query identity.

Furthermore, to validate the statistical significance of the observed performance differences, we applied the non-parametric Friedman test to each result table, considering the mAP values across all data sets and training configurations. In this context, each row of the table represents a distinct training configuration (e.g., Table 2 presents 43 configurations), defined by a specific combination of number of cameras, number of identities, number of images per identity and per camera, and consequently a different total number of training images. For each of these configurations, the corresponding mAP value was collected across all four datasets and used for the Friedman test to compare the relative performance between configurations. In all cases, the test yielded p-values below 0.05 (ranging between $p = 0.013$ and $p = 4 \cdot 10^{-8}$), which confirms the statistical robustness of the findings. Therefore, the performance differences reported in the tables can be considered statistically significant.

4.3. Results

In this section, we separately present and briefly discuss the results attained using each synthetic data set on the ResNet-50 architecture in distinct tables. A more thorough discussion and a comparison of the results attained by the different synthetic data sets on the ResNet-50 and the ViT-small architectures are given in Section 5. Each table in this section reports the results attained using a given synthetic data set for training on the four real data sets used for testing (target) and is composed of several sub-tables corresponding to the different values of the **number of cameras** (#cam). The rows in each sub-table correspond to a different **number of images per identity and per camera** (#imIdCam, ranging from 1 to 6, with only one exception for ClonedPerson, see below); the corresponding overall number of identities (#IDs) and data set size (overall number of images, #im) are also reported. To enable comparison with the complete downloaded versions, the corresponding results are presented in the first row of each table. For reference, we have also included in the first row of each sub-table the results obtained using all images for all identities captured by the selected cameras. Here, the label 'unb' indicates that the number of images per identity and per camera is unbalanced. We remind that the essential code needed to generate all experimental subsets is available at the following link.

4.3.1. UnrealPerson

As shown in Table 1, the downloaded version of this data set contains an unbalanced number of cameras across scenes. To address this, and according to the criteria defined in Section 4.1, we first balanced the number of cameras per scene by discarding 4 cameras from the last scene. We then varied the number of cameras per scene (from 1 to 6) and the number of images per identity per camera (from 1 to 6). Results are reported in Table 2.

As can be observed, for each sub-table, the balanced versions provided, in most cases, comparable or better performances with respect to the corresponding unbalanced version, despite a significantly lower data set size. For instance, on the MSMT target data set, the balanced versions outperformed the corresponding unbalanced ones for each number of cameras, except for the case of 4 cameras. Moreover, better results are obtained using a lower number of images per identity and per camera, i.e., for $\#\mathrm{imIdCam} \in [1, 3]$, especially when a higher number of cameras are used. However, for the smallest number of cameras (4 and 8), a larger number of images per identity is necessary to obtain comparable results. Similarly, when a lower number of identities was used, a higher number of images per identity was necessary to obtain comparable results.

4.3.2. ClonedPerson

Based on the data set statistics shown in Table 1 and following the criteria defined in Section 4.1, we selected configurations with 4 and 8 cameras. For the 4-camera setting, we sampled between 1 and 6 images per identity per camera, while for the 8-camera setting the range was limited to 1 to 5 images, as no identity had 6 images available in all 8 cameras. We did not consider configurations with more than 8 cameras, as the number of qualifying identities would become too small compared to real benchmark data sets. The corresponding results are reported in Table 3.

Note first that similar results are attained on all target data sets using both unbalanced versions of ClonedPerson, despite one of them containing twice the number of cameras, and thus images, as the other. This seems to indicate that just increasing the number of images (and of cameras) does not necessarily improve the generalisation capability. On the other hand, some of the balanced versions provided comparable or better performances than the corresponding unbalanced version, despite containing a lower number of images. For instance, the balanced version of ClonedPerson with 8 cameras and 1 image per identity and per camera outperformed the unbalanced counterpart on the MSMT target data set in all the considered metrics, as well as on Duke (except for mAP), despite a notable reduction of about 94% in the number of images.

4.3.3. FineGPR

According to the data set statistics reported in Table 1 and following our selection criteria (Section 4.1), we selected 4 scenes and sampled between 1 and 7 cameras per scene (one more than in previous settings), and 1 to 6 images per identity per camera. The number of qualifying identities remained constant across all configurations (549). Results are presented in Table 4.

Table 2
Results attained using UnrealPerson [10] as training set. For the meaning of the columns and of the sub-tables see Section 4.3. Best results in each sub-table are highlighted in bold.

Source: Unreal				Target: Market				Target: Duke				Target: MSMT				Target: OccludedReId			
#cam	#IDs	#im	#imIdCam	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
28	2,800	1,255,297	-	41.7	70.8	82.8	87.4	41.6	64.6	76.5	81.0	11.3	30.9	42.8	48.8	73.7	78.6	91.5	94.6
	2,800	1,155,076	unb	42.1	71.4	82.3	86.5	41.2	64.0	76.0	80.5	11.4	31.1	43.4	49.2	73.6	79.0	91.4	94.4
	1,789	42,936	1	40.4	**70.9**	83.1	87.0	**40.5**	**65.5**	**77.0**	**80.3**	**12.0**	**34.0**	**45.9**	**51.1**	72.2	79.1	91.4	94.2
	1,517	72,816	2	**41.7**	**70.9**	**83.5**	**87.1**	40.4	64.6	76.3	**80.3**	11.6	32.6	44.5	49.8	**73.8**	**81.0**	**91.7**	**94.9**
24	1,266	91,152	3	40.6	70.2	82.7	86.7	40.0	63.8	75.3	80.0	11.4	32.0	44.0	49.5	72.8	80.7	91.6	94.0
	1,008	96,768	4	40.5	69.0	82.4	86.4	40.0	64.7	75.0	79.7	11.1	31.3	43.3	48.9	71.8	78.1	90.3	93.8
	768	92,160	5	39.9	69.7	82.5	86.4	40.4	64.5	76.4	80.2	11.2	31.3	43.6	49.1	73.2	80.6	90.7	94.4
	522	75,168	6	39.2	68.1	81.7	85.7	38.4	62.3	75.0	79.6	10.9	30.7	43.5	48.8	70.4	76.7	89.6	94.1
	2,800	939,573	unb	34.2	63.7	77.8	82.5	41.4	64.4	76.6	80.8	11.5	31.5	43.5	48.9	74.4	81.3	91.5	94.9
	1,789	35,780	1	40.2	71.4	**83.8**	**87.4**	**40.3**	**65.8**	**76.9**	**81.0**	**12.6**	**35.7**	**47.8**	**52.8**	71.2	77.1	90.3	94.1
	1,517	60,680	2	**41.1**	**71.5**	83.6	87.3	40.2	64.8	75.4	79.5	12.2	34.0	45.9	51.2	71.8	77.8	90.5	93.9
20	1,266	75,960	3	**41.1**	71.1	83.4	87.0	40.2	64.3	75.7	80.1	11.8	33.2	45.1	50.8	73.7	80.7	**92.4**	**94.7**
	1,008	80,640	4	40.8	70.4	83.0	87.3	40.1	63.9	76.2	80.0	11.4	32.0	44.0	49.5	**73.9**	**81.7**	90.6	94.4
	768	76,800	5	39.7	69.3	82.7	86.4	40.1	63.4	75.8	79.9	11.3	31.6	44.0	49.6	71.7	77.8	90.6	93.0
	522	62,640	6	38.4	68.6	81.3	86.2	38.3	62.7	73.8	77.8	11.2	31.6	44.6	50.2	70.3	77.6	88.7	92.6
	2,800	691,841	unb	40.8	69.2	82.3	86.5	41.9	64.8	76.8	81.5	11.0	30.1	42.1	48.0	73.1	77.9	91.0	94.6
	1,789	28,624	1	39.3	70.3	83.1	86.7	37.9	63.8	74.9	79.4	**12.3**	**35.5**	**47.0**	**52.2**	68.8	76.0	86.7	92.4
	1,517	60,680	2	40.8	**71.6**	**83.4**	**87.5**	40.1	**65.3**	75.8	79.5	11.8	33.2	45.1	50.4	73.0	79.8	90.6	94.2
16	1,266	60,768	3	**41.3**	72.0	83.2	86.9	39.7	63.8	74.9	79.4	11.6	32.9	44.1	49.7	**73.3**	80.0	90.4	93.4
	1,008	64,512	4	40.4	70.6	82.9	86.7	39.8	64.3	75.3	**79.8**	11.3	32.0	44.0	49.4	72.3	79.2	89.5	93.2
	768	61,440	5	40.0	69.1	82.2	86.0	**40.7**	64.5	**76.2**	79.7	11.2	31.8	43.7	49.4	71.9	**80.5**	**90.8**	**93.6**
	522	50,112	6	39.0	68.7	82.1	85.6	38.1	63.0	74.7	78.4	11.0	31.4	43.2	48.6	70.7	77.6	89.5	**93.6**
	2,800	503,698	unb	41.4	70.6	82.7	86.8	40.7	64.1	75.6	80.0	10.9	30.6	42.3	47.4	72.6	77.9	90.8	94.6
	1,789	21,468	1	37.2	68.9	82.5	86.8	34.7	60.7	72.8	77.3	11.5	**34.5**	**46.0**	**51.1**	65.3	71.0	86.0	90.5
	1,517	54,612	2	40.2	**71.6**	83.3	87.3	38.7	63.7	75.7	79.4	**11.7**	33.5	45.2	50.3	70.7	77.2	89.6	93.1
12	1,266	45,576	3	40.2	70.3	**83.5**	87.1	39.7	64.3	76.1	80.0	11.6	32.7	44.6	49.8	71.8	78.7	90.4	93.8
	1,008	48,384	4	**40.3**	70.0	**83.5**	**87.4**	**40.4**	**65.4**	**76.9**	**80.6**	10.9	31.2	43.0	48.5	**72.1**	**79.0**	**91.0**	**94.7**
	768	46,080	5	39.6	69.4	82.4	86.3	39.3	63.3	75.5	79.4	11.1	31.3	42.7	48.3	69.6	75.2	88.7	93.1
	522	37,584	6	39.0	69.3	82.1	85.6	37.3	62.8	74.6	78.3	10.4	30.1	41.9	47.6	69.8	76.7	88.9	93.1
	2,800	336,513	unb	40.9	70.8	82.0	85.9	39.7	63.5	76.1	80.0	10.3	29.0	40.7	46.3	71.3	77.6	89.4	93.5
	1,789	14,312	1	31.4	63.7	77.9	82.7	30.0	56.4	69.0	74.0	10.1	32.3	43.8	49.1	59.7	67.3	83.0	87.6
	1,517	24,272	2	37.2	68.3	81.5	85.1	36.3	62.1	73.8	78.1	11.3	**33.4**	**44.8**	**50.6**	66.5	72.9	87.1	91.5
8	1,266	30,384	3	39.5	69.7	**82.7**	86.3	37.1	62.0	74.3	78.4	**11.6**	**33.4**	**44.8**	50.4	**68.9**	76.1	**89.9**	92.6
	1,008	32,256	4	39.5	69.5	82.6	**86.8**	37.8	63.1	74.4	**79.3**	11.1	32.4	43.8	49.2	67.9	72.9	88.2	**93.8**
	768	30,720	5	**40.0**	**70.5**	82.6	86.3	**38.5**	**63.6**	**75.6**	**79.3**	11.0	31.9	43.2	48.4	**68.9**	**76.3**	88.8	93.5
	522	25,056	6	38.1	68.6	81.4	85.5	36.7	61.4	73.4	78.2	10.8	30.7	42.8	48.1	67.7	74.6	87.4	92.5
	2,800	138,275	unb	38.4	68.6	81.2	85.3	38.3	61.3	74.2	78.8	10.3	29.5	41.2	46.3	68.7	74.4	87.4	92.0
	1,789	7,156	1	21.4	51.5	68.3	74.1	20.4	44.3	58.7	63.3	6.9	25.6	36.6	41.6	46.7	54.0	71.9	78.9
	1,517	12,136	2	29.8	62.2	76.8	81.7	28.0	54.1	66.4	71.0	9.4	30.8	41.8	46.9	57.7	64.3	81.6	88.5
4	1,266	15,192	3	34.1	66.9	79.9	83.8	31.8	56.9	70.0	74.4	9.8	30.5	41.5	46.4	63.5	70.7	**85.7**	**90.7**
	1,008	16,128	4	35.7	67.6	80.9	**85.2**	34.4	60.1	71.0	75.1	**10.1**	**31.0**	**42.3**	**47.5**	63.6	**71.1**	85.3	90.2
	768	15,320	5	**36.2**	**68.6**	**81.3**	84.8	**35.1**	**61.4**	**72.5**	**77.1**	9.8	29.8	40.7	45.9	**64.2**	70.5	84.6	90.6
	522	12,528	6	34.6	65.9	79.4	83.5	33.3	58.9	70.9	74.9	9.7	29.2	40.6	46.1	62.3	69.1	84.6	89.9

Table 3
Results attained using ClonedPerson [45] as training set. For the meaning of the columns and of the sub-tables see Section 4.3. Best results in each sub-table are highlighted in bold.

Source: ClonedPerson				Target: Market				Target: Duke				Target: MSMT				Target: OccludedReId			
#cam	#IDs	#im	#imIdCam	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
24	4,826	763,953	-	47.6	77.1	89.9	93.3	33.8	57.8	71.9	76.9	8.6	26.4	39.4	45.3	60.8	67.0	83.6	88.7
	4,826	404,701	unb	41.7	73.3	86.8	91.4	28.4	51.3	65.9	71.2	5.9	20.2	30.6	36.4	59.7	66.7	82.3	87.3
	2,035	16,280	1	37.4	70.5	83.9	88.6	29.3	**53.7**	**69.0**	**74.3**	**7.2**	**24.6**	**35.7**	**40.9**	56.8	**64.3**	80.0	**86.2**
8	1,673	26,768	2	39.3	72.7	85.9	89.9	**29.5**	**53.7**	68.1	72.9	6.4	21.8	33.0	38.5	56.4	63.9	79.7	86.0
	1,126	27,024	3	**40.3**	**73.2**	85.8	90.6	28.9	52.2	67.0	72.6	6.2	21.3	32.8	38.0	**57.2**	63.8	**80.9**	85.6
	843	26,976	4	39.3	72.3	**86.0**	**90.7**	28.0	52.3	66.3	71.5	6.1	20.7	31.6	37.0	55.4	63.1	79.0	85.3
	601	24,040	5	37.5	70.6	85.0	89.0	26.2	50.0	64.3	69.4	6.1	21.0	32.0	37.2	54.4	61.4	76.1	83.4
	4,826	228,144	unb	41.0	73.0	86.6	90.4	28.6	52.8	68.3	73.9	5.4	18.2	29.1	34.9	58.2	65.1	82.6	88.2
	3,928	15,712	1	29.5	63.4	78.3	83.4	24.3	49.7	64.2	69.8	**6.7**	**24.5**	**36.4**	**41.7**	50.7	58.6	75.6	81.6
	3,580	28,640	2	34.6	69.1	83.1	87.6	24.7	48.9	63.6	68.3	6.4	23.0	34.0	39.3	52.8	59.6	78.8	85.1
4	3,479	41,748	3	35.6	69.7	84.4	88.7	24.6	49.7	62.8	69.3	5.7	21.0	32.2	37.6	51.0	56.6	76.8	84.7
	3,283	52,528	4	36.5	70.6	84.2	88.7	26.4	50.6	65.7	**71.0**	5.3	19.5	30.0	35.6	55.1	60.7	79.8	86.0
	3,013	60,260	5	37.5	**71.1**	84.6	**89.1**	**26.8**	**51.3**	**66.1**	**71.0**	5.6	19.8	30.6	35.9	**55.2**	**61.2**	**80.4**	**86.3**
	2,161	51,864	6	**37.8**	**71.1**	**85.0**	**89.1**	26.5	50.4	64.5	70.8	5.5	20.0	30.2	35.8	54.0	60.9	78.6	85.3

Note that for each number of cameras (i.e., for each sub-table), there is at least one balanced version of FineGPR that provided a comparable or better performance with respect to the corresponding unbalanced version, despite using half of the identities and a much lower overall number of images. Moreover, the performance of all the considered balanced and unbalanced versions is often comparable to, and in some cases better than, the ones attained by the whole data set, despite the latter containing about 12 to 80 times as large a number of cameras. This seems to suggest that using a very large number of cameras (say, hundreds) does not necessarily improve the generalisation capability. On the other hand, similarly to the trend observed for the UnrealPerson and ClonedPerson data sets, when the number of cameras is very low, a higher number of images per identity is often necessary to attain a good performance.

Table 4
Results attained using FineGPR [18] as training set. For the meaning of the columns and of the sub-tables, see Section 4.3. Best results in each sub-table are highlighted in bold.

Source:FineGPR				Target:Market				Target:Duke				Target:MSMT				Target: OccludedReId			
#cam	#IDs	#im	#imIdCam	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
324	1150	2,228,600	-	9.9	27.6	45.0	52.0	11.0	26.8	39.2	45.3	4.4	19.0	29.6	34.9	42.5	48.3	63.6	71.2
28	1150	176,890	unb	12.1	34.0	49.9	57.0	9.6	25.9	37.1	42.1	4.2	19.0	29.1	34.3	51.1	56.3	74.0	80.9
	549	15,372	1	**9.8**	**31.4**	**46.5**	**52.9**	9.5	26.3	36.5	**41.7**	**3.8**	**19.2**	**28.8**	**33.3**	**49.8**	**57.1**	**73.4**	**78.6**
		30,744	2	8.6	29.2	43.4	48.8	**9.6**	**27.5**	**37.3**	41.4	3.4	17.9	27.0	31.4	47.0	53.2	68.6	75.3
		46,116	3	8.8	28.8	43.5	46.9	7.5	23.3	32.9	38.0	3.4	17.9	27.1	32.1	44.2	49.5	67.0	73.5
		61,488	4	7.0	24.3	37.4	43.4	8.5	25.5	35.0	39.3	3.7	18.9	28.1	32.9	43.8	51.1	67.9	72.9
		76,860	5	8.9	29.4	43.0	48.6	8.9	25.9	35.5	40.4	3.6	18.4	27.8	32.2	36.9	39.7	57.6	64.7
		92,232	6	8.3	27.6	42.3	48.4	8.1	24.7	34.8	39.2	3.7	18.8	28.4	33.1	31.9	36.7	51.2	56.6
24	1150	151,620	unb	9.9	28.9	44.2	51.1	10.6	27.4	39.1	43.8	3.5	16.9	26.5	31.3	46.6	50.0	66.0	74.1
	549	13,176	1	10.0	31.1	45.2	50.5	7.0	22.7	32.6	36.8	3.8	19.1	28.8	33.6	**47.6**	**53.9**	**69.6**	**75.8**
		26,352	2	10.0	31.5	45.6	52.2	**9.6**	**27.1**	35.8	40.6	3.9	19.5	29.3	33.9	44.0	49.4	64.2	71.2
		39,528	3	**10.3**	**32.4**	**49.4**	**53.9**	7.8	23.9	33.8	38.0	**4.1**	**20.3**	**29.8**	**34.6**	44.4	50.0	67.4	72.9
		52,704	4	8.1	26.9	41.2	47.1	8.9	26.3	**36.2**	**40.8**	3.6	18.6	27.7	32.3	36.6	41.8	56.4	62.7
		65,880	5	9.1	30.3	45.0	51.3	7.8	23.9	33.3	37.9	3.4	18.6	27.3	31.7	36.4	41.7	55.7	63.3
		76,056	6	8.9	29.2	43.9	50.4	7.2	21.4	32.6	37.8	3.7	19.3	28.2	32.8	39.1	44.5	58.2	64.1
20	1150	126,360	unb	9.1	28.0	42.6	49.5	10.0	26.2	38.1	43.1	3.9	18.8	29.0	34.0	45.2	51.3	68.0	74.9
	549	10,980	1	9.1	28.7	**44.8**	50.8	**8.4**	23.7	33.8	38.4	3.7	19.0	28.3	32.9	**46.3**	**54.5**	**68.8**	**76.1**
		21,960	2	**9.3**	**30.8**	**44.8**	**51.7**	7.9	24.0	33.7	38.1	**3.9**	19.9	**29.5**	**34.2**	45.6	53.7	67.1	73.2
		32,940	3	8.1	27.8	41.9	47.6	6.2	21.0	30.3	33.9	3.7	**19.3**	28.2	33.0	38.6	46.3	58.6	65.8
		43,920	4	8.9	29.5	42.7	48.8	8.2	23.6	34.0	39.3	3.2	17.5	26.3	30.8	36.0	42.5	56.0	62.4
		54,900	5	8.3	28.5	41.7	48.2	9.1	**25.5**	**36.0**	**40.2**	3.4	18.4	27.6	31.9	34.1	40.3	56.3	63.0
		65,880	6	7.6	26.1	40.4	46.6	7.3	22.6	31.9	36.2	3.6	18.0	27.1	31.8	35.7	42.0	57.3	63.2
16	1150	101,080	unb	10.0	30.1	45.7	53.0	7.7	22.3	32.1	36.8	3.2	16.1	25.5	29.9	46.8	51.8	69.1	77.0
	549	8,784	1	9.0	29.7	43.5	50.3	**7.9**	**23.7**	**33.3**	**38.4**	3.1	17.4	25.9	30.4	**45.1**	**53.5**	**67.7**	73.3
		17,568	2	**9.3**	**30.2**	**45.3**	**51.4**	6.7	21.2	30.3	34.4	3.2	17.5	**27.3**	**31.9**	44.2	52.6	**67.7**	**73.6**
		26,352	3	8.8	29.6	42.8	49.3	6.8	21.1	30.7	35.2	3.2	17.7	26.0	30.7	39.1	45.7	59.6	65.7
		35,136	4	8.6	29.2	42.8	48.5	7.1	21.5	31.2	35.5	3.1	16.7	25.7	30.1	36.8	43.0	57.9	63.9
		43,920	5	8.6	29.0	42.3	48.6	6.8	21.9	31.4	35.4	**3.3**	**17.9**	27.1	31.5	41.5	48.1	65.2	71.4
		52,704	6	6.0	23.0	36.0	42.3	7.6	23.0	32.4	35.7	3.2	17.5	26.2	30.8	34.2	38.7	55.1	62.0
12	1150	75,810	unb	9.4	29.4	43.6	50.8	7.4	21.3	30.8	36.0	2.7	14.4	22.9	27.8	46.0	52.6	70.0	77.4
	549	6,588	1	**9.1**	**29.8**	**43.9**	**50.2**	6.2	20.2	28.2	32.9	3.3	17.7	26.8	31.1	37.8	44.4	58.6	66.4
		13,176	2	7.5	26.4	39.5	46.1	6.5	20.5	29.8	34.6	**3.6**	**18.6**	**27.6**	**31.8**	35.5	42.3	57.6	63.4
		19,764	3	8.0	27.2	41.5	47.4	**7.2**	**21.7**	**31.0**	**35.4**	3.1	16.8	25.4	29.6	**40.0**	**46.3**	**64.0**	**69.6**
		26,352	4	7.5	26.0	39.0	45.3	6.6	20.8	30.3	33.6	3.2	17.4	26.2	30.8	35.7	43.1	58.1	64.7
		32,940	5	7.3	26.1	38.5	45.5	6.3	21.0	30.2	34.7	3.0	17.1	25.4	29.7	33.7	40.2	56.0	62.5
		39,528	6	7.4	25.1	39.0	45.0	8.0	24.2	32.5	37.5	3.2	17.4	26.2	30.3	33.9	40.0	55.1	61.8
8	1150	50,540	unb	9.2	27.6	43.5	50.3	6.2	19.3	29.1	33.3	2.8	14.9	23.6	27.8	41.9	46.8	65.2	72.4
	549	4,392	1	7.7	25.7	39.2	45.5	6.3	18.7	28.3	32.4	2.9	15.6	23.9	28.3	**39.7**	**47.2**	**62.8**	**69.1**
		8,784	2	8.0	27.3	40.9	46.9	6.5	**21.2**	28.0	32.4	**3.3**	16.9	25.9	30.0	37.0	43.7	60.2	66.4
		13,176	3	7.3	25.1	37.9	44.7	6.6	21.1	28.9	32.8	3.2	17.0	25.9	30.4	37.7	45.3	60.6	67.3
		17,568	4	7.7	26.2	40.2	46.7	6.5	20.2	29.4	33.8	3.2	17.2	**26.5**	30.7	37.0	43.9	59.6	65.3
		21,960	5	**8.2**	**28.1**	**41.8**	**47.7**	**6.7**	20.9	**30.3**	**34.4**	**3.3**	**17.5**	26.4	**31.0**	39.2	44.9	61.0	67.1
		26,352	6	7.9	26.7	40.9	47.4	6.5	20.2	29.5	34.0	2.9	15.9	24.5	28.9	31.6	36.5	50.2	58.6
4	1150	25,270	unb	7.7	25.5	40.1	46.9	6.1	18.3	27.9	33.1	2.6	13.7	22.2	26.8	35.2	39.9	55.4	65.6
	549	2,196	1	**8.5**	**27.0**	**42.7**	**49.4**	5.2	15.8	23.9	27.7	2.8	15.0	23.3	27.6	34.8	39.0	57.0	67.3
		4,392	2	7.7	26.1	39.8	46.1	6.0	18.8	27.1	31.7	2.9	15.5	24.2	28.7	34.4	39.9	57.9	66.4
		6,588	3	7.2	24.0	37.3	44.6	**6.2**	**19.6**	**28.5**	**32.9**	**3.1**	16.2	24.7	29.0	**39.3**	**45.6**	**63.2**	**72.3**
		8,784	4	6.1	22.2	35.3	41.4	5.7	18.0	26.9	31.3	3.0	**16.3**	**24.8**	**29.2**	36.9	42.2	60.0	67.8
		10,980	5	7.2	24.2	37.7	43.7	6.0	19.3	**28.5**	32.3	2.9	16.0	23.9	28.4	37.7	43.6	60.0	67.2
		13,176	6	7.1	24.5	38.3	43.9	5.1	16.6	24.9	28.7	2.9	15.8	23.8	28.1	35.0	40.0	56.0	64.7

4.3.4. PersonX

As shown in Table 1, PersonX includes 6 scenes rendered with a single camera each: 3 with uniform backgrounds and 3 urban environments. To reduce redundancy, we selected one uniform-background scene (blue) and the three urban scenes. For each selected scene, we sampled between 1 and 6 images per identity. The corresponding results are reported in Table 5.

Notably, all the considered reduced versions of PersonX outperformed the original version on all three target data sets and for almost all the performance metrics; in particular, the best results were attained by the *smallest* version, which was made up of about 5k images vs more than 270k of the original version and a single image per identity and per camera. This is further evidence that just increasing the data set size does not guarantee a better performance. In the specific case of PersonX, the reason for this behaviour could be partly due to the redundancy of its original version, which contains the same pedestrian images rendered on three uniform backgrounds of different colours.

5. Discussion

In this section we provide a more thorough discussion of the results summarised above, including a comparison, whenever possible, among the different synthetic data sets.

5.1. Quantitative comparison

To facilitate the comparison among synthetic data sets, we reported in Figs. 2 and 3 the performance of their original downloaded (D) versions, including WePerson and RandPerson, as well as the best-performing reduced versions (S) of the four data sets considered in our experiments. To highlight the consistency of the obtained results, reported in Section 4.3, we included the results obtained on both the ResNet-50 and the ViT-small architectures. In most cases (except for ClonedPerson), the reduced (S) versions of the data sets outperformed the downloaded (D) versions, highlighting that a simple increase in

Table 5
Results attained using PersonX [19] as training set. For the meaning of the columns and of the sub-tables, see Section 4.3. Best results in each sub-table are highlighted in bold.

Source: PersonX				Target: Market				Target: Duke				Target: MSMT				Target: OccludedReId			
#cam	#IDs	#im	#imIdCam	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10
6	1,266	273,456	-	21.8	45.4	63.3	70.3	17.2	33.5	46.2	52.2	2.7	8.8	15.2	19.1	43.8	47.5	65.0	75.1
		182,304	unb	22.4	47.9	65.1	72.2	17.2	33.3	46.5	53.0	2.7	8.7	15.2	18.8	40.0	41.2	61.9	70.5
		5,064	1	**25.9**	**53.6**	59.1	**74.7**	**21.7**	**40.9**	**54.8**	**60.3**	**4.4**	**14.7**	**23.4**	**27.9**	**48.5**	**53.7**	**70.1**	**78.6**
		10,128	2	24.7	52.1	67.8	73.6	20.7	39.1	52.5	59.7	3.9	12.8	20.9	25.5	45.1	48.7	69.2	75.6
4	1,266	15,192	3	25.2	52.0	68.1	74.4	**21.7**	40.7	54.2	59.5	3.8	12.4	20.6	25.1	46.4	50.8	68.2	76.4
		20,256	4	24.9	51.8	67.2	73.4	20.7	38.8	53.1	59.3	3.9	12.3	20.4	24.9	44.1	47.7	67.2	74.3
		25,320	5	25.0	51.9	**68.3**	73.6	20.9	38.1	52.7	59.1	3.7	11.7	19.4	23.7	44.9	48.8	66.8	75.5
		30,384	6	24.3	51.5	67.6	73.9	20.5	37.8	52.6	58.8	3.8	12.0	20.0	24.3	44.7	47.8	67.7	75.3

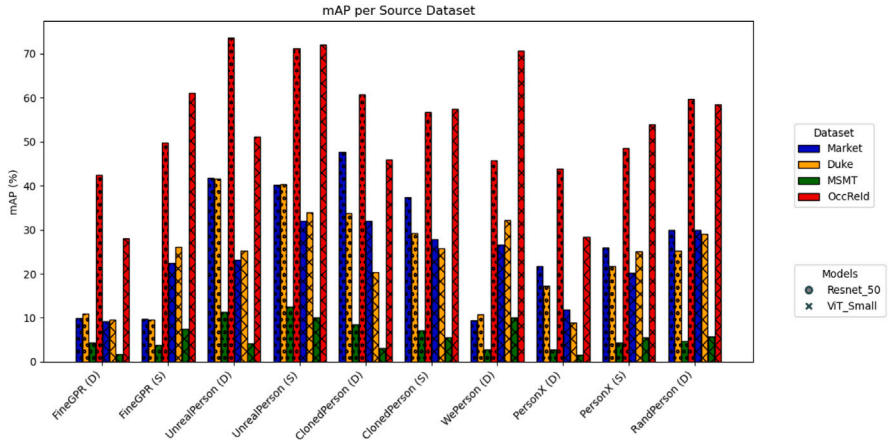

Fig. 2. Results in terms of mAP attained using the downloaded version of the synthetic data sets (D), compared with the best results attained by the reduced versions (S) of the four data sets considered in Section 4.3.

training set size is not always beneficial to generalisation capability. This trend was consistent across both architectures, confirming that the findings presented earlier are applicable to different model types. This outcome is particularly striking when considering the scale of the reductions. For instance:

- **FineGPR**: S version has 549 IDs, 28 cameras, and 15,572 images; D version has 1,150 IDs, 324 cameras, and over 2M images.
- **UnrealPerson**: S version has 1,789 IDs, 20 cameras, and 35,780 images; D version has 2,800 IDs, 28 cameras, and over 1M images.
- **ClonedPerson**: S version has 2,035 IDs, 8 cameras, and 16,280 images; D version has 4,826 IDs, 24 cameras, and 763,953 images.
- **PersonX**: S version has 1,266 IDs, 4 cameras, and 5,064 images; D version has 1,266 IDs, 6 cameras, and 273,456 images.

These findings underline that a compact but carefully curated data set can yield superior results, reducing unnecessary data complexity while preserving or improving performance. It is worth highlighting that any synthetic data set achieves good performance when the target data set is OccReId. It indicates that the considered synthetic data sets are also effective in the case of several occlusions.

A further observation is that the best results on Duke and MSMT were attained using UnrealPerson (both the D and S versions). On the other hand, the best performance on Market was attained, instead, using the D version of ClonedPerson data set. This indicates that despite its lack of balance across all factors and not having the highest number of images in its original form, ClonedPerson seems to be more effective than the other data sets. Likely reasons are its superior visual quality, as its pedestrian images have been derived from real ones, as well as variability, as it contains the highest number of identities among the considered data sets (Table 1).

Lastly, Figs. 2 and 3 also show that the weakest performances, across both architectures, were typically associated with FineGPR and WePerson. The underlying reason is not immediately clear, but a notable distinction is their much higher number of cameras (by one to two orders of magnitude compared to others).

We hypothesise that, beyond a certain point, increasing the number of cameras may introduce too much variation in the feature space, diluting the identity-specific information and thereby hindering model convergence. We will investigate this hypothesis further in the next section.

5.2. Further analysis on the number of cameras

We conducted additional experiments to investigate the effect of increasing the number of cameras. As previously mentioned, WePerson lacks a sufficient number of images per identity, making it unsuitable for this analysis. Instead, we focused on the FineGPR and UnrealPerson data sets, which offer a higher number of cameras and a better balance across key factors.

We reused the ResNet-50 Re-Id models trained in the experiments described in Section 4.3, and evaluated the distributions of Euclidean distances in the feature space between identity embeddings. Specifically, we examined how these distributions evolve as the number of cameras increases, ranging from 4 to 24 in steps of 4. The results are summarised in the box plots shown in Fig. 4. For each camera configuration, we computed intra-class distances (between all distinct pairs of images of the same identity) and inter-class distances (between images of different identities). This analysis provides insights into how camera variability affects the feature space structure, and, in particular, the compactness and separability of identity embeddings.

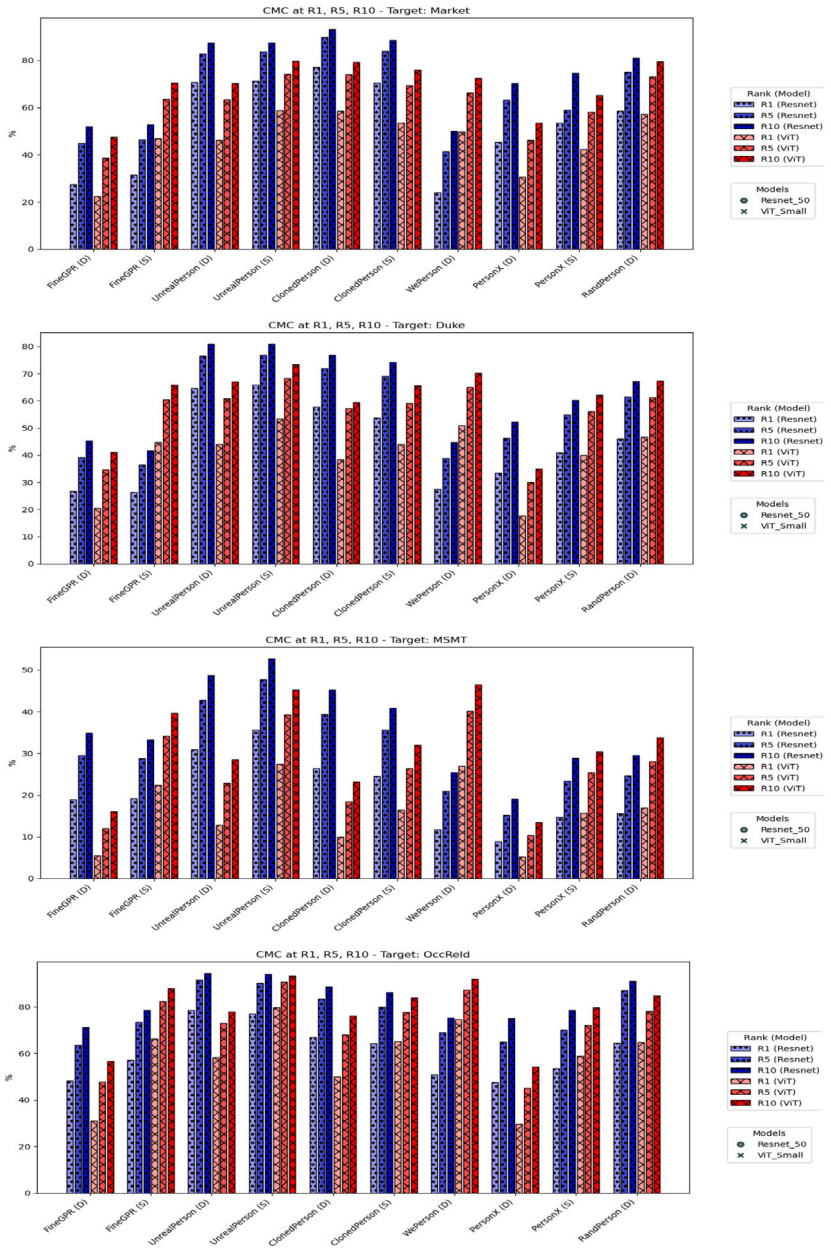

Fig. 3. Results in terms of CMC (R1, R5 and R10) attained using the downloaded version of the synthetic data sets (D), compared with the best results attained by the reduced versions (S) of the four data sets considered in Section 4.3. From top to bottom: Market, Duke, MSMT and OccReId.

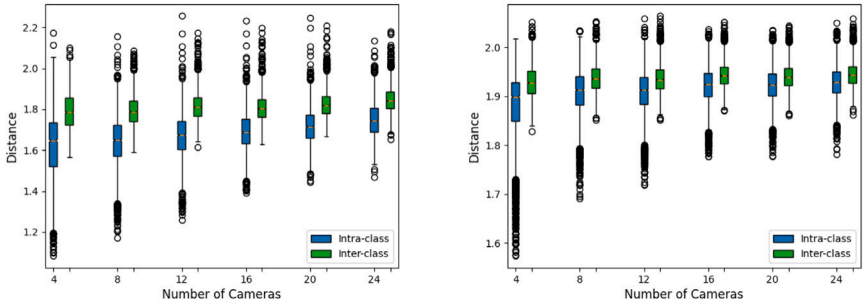

Fig. 4. Distribution of the distances in feature space computed between all pairs of images of the same identity (intra-class) and of different identities (inter-class) for UnrealPerson [10] (left plot) and FineGPR [18] (right plot), as a function of the number of cameras.

Table 6
FID scores computed among each pair of synthetic and real data sets (bottom rows), and among real data sets (top rows).

	Training set	Test Sets		
		Market	Duke	MSMT
Real	Market [42]	6.809	62.362	82.711
	Duke [43]	58.703	6.726	61.849
	MSMT [23]	78.593	62.437	1.381
Synthetic	UnrealPerson [10]	42.292	61.911	61.971
	ClonedPerson [45]	73.962	93.210	100.478
	RandPerson [17]	104.017	109.175	99.112
	PersonX [19]	124.222	135.008	115.965
	FineGPR [18]	146.460	120.928	96.740
	WePerson [16]	185.454	156.698	143.565

As shown in Fig. 4, increasing the number of cameras leads to a noticeable reduction in the dispersion of the distance distributions. In particular, the range between minimum and maximum distances narrows significantly. This trend is especially evident in the UnrealPerson data set. Moreover, the dispersion appears to stabilise when the number of cameras reaches 16, suggesting a point of diminishing returns. Beyond this threshold, the gap between intra-class and inter-class distances continues to shrink, which may negatively impact Re-Id performance by reducing the separability of identity features in the embedding space. These observations indicate that while increasing the number of cameras initially enhances feature diversity and generalisation, an excessive number may introduce redundancy and reduce the discriminative power of the learned representations, ultimately hindering Re-Id accuracy.

5.3. Photorealism of synthetic data sets

As discussed in Section 5.1, one key factor that may significantly affect the generalisation capability of Re-Id models trained on synthetic images is the degree of photorealism. Higher photorealism is likely to reduce the synthetic-to-real domain gap, thereby improving cross-domain performance. It is worth noting that domain shift issues are not exclusive to synthetic data. They also affect models trained on real-world data, as demonstrated by cross-dataset evaluations where models trained on one real data set perform poorly on another [15,23].

While our initial observation in Section 5.1, highlighting ClonedPerson as the most photorealistic among the synthetic data sets, was based on a subjective visual assessment, we aim to support these claims with a quantitative metric. The Fréchet Inception Distance (FID) [49] is commonly used to evaluate the realism of generated images, by measuring the distance between the feature distributions of synthetic and real data. Lower FID scores indicate a closer match between distributions, and prior work has shown a correlation between lower FID and improved performance in Re-Id and face recognition [50,51].

Accordingly, we used the FID metric to measure the distance between the available synthetic data sets and the three real-world benchmarks considered in our experiments, namely Market, Duke and MSMT17. We intentionally excluded OccReId, as it was designed for evaluating occlusion-specific scenarios and does not reflect standard Re-Id conditions. In particular, the FID metric was computed between the whole synthetic data sets (considered as training sets) and the testing sets of real data sets. For reference, we also computed the FID metric among the training and the testing set of each pair of real data sets, providing a baseline for natural inter-dataset variation. The results are summarised in Table 6.

The results show that UnrealPerson and ClonedPerson consistently yield the lowest FID scores across all three real data sets, which aligns well with their superior performance in our experimental evaluations. This provides further evidence that the degree of photorealism correlates with better generalisation to real-world scenarios. Overall, the trends observed in Table 6 are consistent with the performance rankings of the downloaded (D) versions shown in Figs. 2 and 3, providing further evidence that photorealism is a critical factor for effective synthetic training data in Re-Id.

5.4. Final remarks

We conclude this paper by discussing the role of the factors listed in Section 4 on the performance of Re-Id models trained on synthetic images, in light of our experimental results. Some factors, like the number of identities and images per identity per camera, allow for a clearer assessment of their impact on Re-Id performance. For others, only indirect insights are possible, since the characteristics of the available synthetic data sets — mainly their imbalance — limit direct investigation.

Total number of images. Our results indicate that total image count is not the primary factor driving the generalisation capability of a Re-Id model. In many cases, smaller subsets of a synthetic data set — such as using only 3% of UnrealPerson's images (1.2M vs 35,780) — often achieved better performance (see Figs. 2, 3), highlighting the importance of the diversity of factors within a data set over sheer size.

Number of identities. Our experiments, though not directly controlling for the number of identities (due to the above-mentioned issues), suggest that both the quantity and *diversity* of identities are crucial for Re-Id performance. ClonedPerson, which performed among the best, highlights this importance, as its diverse identities were generated from real images. This approach contrasts with synthetic data sets that rely on 3D human models from graphics engines, which often yield fewer identities (even compared to existing real data sets) due to the effort involved in model creation. Notably, our findings show that optimal performance, particularly with UnrealPerson and ClonedPerson, tends to occur with at least 1,000 to 2,000 identities, indicating that a minimum of 2,000 could be suggested for a suitable synthetic data set.

Number of cameras and scenes. Our findings align with previous studies [16,17] (which, however, focused on a single data set; more details in Section 2.4). Indeed, we found benefits in increasing the number of cameras up to a threshold — around 16 cameras; see the results in Section 5.2 — after which excessive data dispersion in feature space can occur, diminishing the representation of individual characteristics.

Although we did not conduct a focused analysis on the number of scenes, similar conclusions can be inferred: the diversity in scene backgrounds, rather than sheer quantity, is key to improving generalisation. Prior work supports this [15], showing that Re-Id models perform better when synthetic data sets feature realistic and varied backgrounds, rather than different but uniform ones. This diversity can be effectively achieved by using distinct viewpoints within a single urban setting, such as one with buildings and another with trees. Thus, placing multiple strategically positioned cameras in a single, varied scene can offer enough background diversity to improve model performance, proving more practical and efficient than creating multiple virtual environments.

Number of images per identity. Our experiments indicate that a single image per individual per camera fails to capture enough discriminative features. However, optimal or nearly optimal performance was

Table 7
Summary of empirical findings across design factors explored in the study.

Design factor	Relevance	Empirical trend	Recommended value(s)	Inconclusive aspects	Practical implications
Number of cameras	Very High	Strong positive effect. A huge number of cameras can cause data dispersion	≈16	Effect depends on viewpoint and camera diversity	Higher values allow for the generation of a smaller number of scenes
Number of identities	High	Positive effect up to saturation. Low values are not representative of real world scenarios	≥2000	May vary depending on inter-class variability	Include a large, diverse set of identities to promote generalisation could lead to a great effort in generating human models.
Images per identity per camera	High	Diminishing returns after ~10	≥2 per camera	Exact inflection point not sharply defined	Emphasise balance and identity diversity over sheer volume
Photorealism	High	Possibly beneficial, not directly measured	–	No direct metric for realism used	Improve realism when feasible, but effects not isolated
Number of total images	Low	A very low value can cause a lack in diversity factors	It depends on the previous suitable values	–	Large or huge value might not ensure better performances

frequently achieved with only 2 to 3 images per identity, suggesting this limited quantity is often sufficient for effective feature extraction.

Photorealism of synthetic images. We identified image quality, as reflected in lower FID scores, as a key factor enhancing the generalisation capability of synthetic training data, effectively reducing the gap to real data. Data sets like UnrealPerson and ClonedPerson, which achieved the lowest FID scores, consistently delivered the best performance across various experimental conditions.

Camera viewpoint or person orientation. Although we did not specifically analyse pedestrian orientation with respect to the camera, previous studies have addressed this issue [18,19], with results aligning with intuition. They found that the best performance is achieved when the training set includes a representative range of pedestrian orientations. This can be easily accomplished, as mentioned earlier, by positioning multiple cameras with different viewpoints within the same scene, as long as the backgrounds are sufficiently distinct. However, as discussed earlier, an excessive number of viewpoints could cause data dispersion in the feature space. Thus, a realistic configuration of cameras and viewpoints is preferable to avoid unrealistic scenarios where individuals are captured from every angle. In specialised cases, such as targeted investigations [18,19], data sets can be designed with specific orientations in mind, rather than being used as a substitute for real-world images for training Re-Id models with strong generalisation capabilities.

In summary, as far as our experiments are concerned, the degree of photorealism of a synthetic data set, and the number of, and diversity in, identities and camera views (or scenes), turned out to be the most critical factors affecting generalisation capability to real data. On the other hand, while data set size (i.e., the overall number of images) does play a role, its contribution seems less relevant than expected. To facilitate a clearer understanding of our empirical findings, we summarise the observed effects of each training set design factor in Table 7. The table reports the empirical trend associated with each factor, recommended value ranges when identifiable, any limitations or inconclusive outcomes, and practical implications for dataset construction.

Failure case analysis. While our study focuses on the effect of training set design on generalisation performance, we briefly analyse typical failure cases across configurations. Most commonly, errors occur in scenarios involving:

- identities with very similar appearance (e.g., clothing, body shape);
- occlusions, particularly when severe (e.g., pedestrians largely obscured by objects or other people);
- rare viewpoints not well represented in the training data (e.g., top-down or extreme side views).

These observations suggest that improving camera diversity and realism may further reduce misidentification rates, especially in edge cases. Furthermore, thanks to the inclusion of OccReId, we observed that the trained models tend to handle mild occlusions relatively well, but their performance degrades significantly in the presence of heavier occlusions. These findings confirm the importance of designing training sets that include more realistic occlusion patterns and reinforce the need for methods explicitly targeting occlusion robustness.

6. Conclusions

In recent years, synthetic data have become a promising way to address the limitations of real data in training Re-Id models, offering large data sets with diverse visual variations without manual annotation or privacy concerns. However, synthetic data sets vary widely in size, number of identities, cameras, scenes, and photorealism, and the impact of these factors on Re-Id performance remains underexplored.

Our comprehensive analysis offers key insights into this issue. We show that merely increasing data set size does not guarantee better generalisation to real images. Instead, attention should be given to both the number and diversity of identities and camera views, as well as photorealism, to reduce the synthetic-to-real domain gap. Additionally, we find that an excessive number of camera views can reduce performance by dispersing data in feature space, under-representing individual identities—especially if not balanced by sufficient images per identity.

Though our study is limited to a specific deep learning architecture, we demonstrate that our findings generalise well across architectures. We believe these insights offer valuable guidelines for developing synthetic data sets for Re-Id. Building on this, our future aim is to create a synthetic data set guided by the principles established in this work, further advancing Re-Id research.

CRediT authorship contribution statement

Rita Delussu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lorenzo Putzu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Fadi Boutros:** Writing – review & editing, Supervision. **Carmen Bisogni:** Writing – review & editing, Supervision. **Naser Damer:** Writing – review & editing, Supervision. **Giorgio Fumera:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the projects: "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. We also acknowledge the support by the Italian Ministry of Enterprises and Made in Italy (MIMIT), within the 5G technology support program, on axis 1 "House of Emerging Technologies" (CTE), Project Name "Cagliari Digital Lab" (ID: G27F22000040008). This work was partially supported by projects FAIR (PE00000013) under the NRRP MUR program funded by the EU–NGEU (CUP: J23C24000090007) and SERICS (PE00000014) under the NRRP MUR program funded by the EU-NGEU (CUP: F53C22000740007).

Data availability

The authors do not have permission to share data.

References

[1] S. Nikolenko, Synthetic Data for Deep Learning, vol. 174, Springer, 2021, http://dx.doi.org/10.1007/978-3-030-75178-4.

[2] R. Delussu, L. Putzu, G. Fumera, Synthetic data for video surveillance applications of computer vision: A review, Int. J. Comput. Vis. (2024) 1–37, http://dx.doi.org/10.1007/s11263-024-02102-x.

[3] Q. Wang, J. Gao, W. Lin, Y. Yuan, Learning from synthetic data for crowd counting in the wild, in: Int. Conf. on Computer Vision and Pattern Recognit, CVPR, 2019, pp. 8198–8207, http://dx.doi.org/10.1109/CVPR.2019.00839.

[4] P. Shamsolmoali, M. Zareapoor, E. Granger, et al., Image synthesis with adversarial networks: A comprehensive survey and case studies, Inf. Fusion 72 (2021) 126–146, http://dx.doi.org/10.1016/j.inffus.2021.02.014.

[5] G. Guo, N. Zhang, A survey on deep learning based face recognition, Comput. Vis. Image Underst. 189 (2019) 102805, http://dx.doi.org/10.1016/j.cviu.2019.102805.

[6] O.C. Uner, C. Aslan, B. Ercan, T. Ates, U. Celikcan, A. Erdem, E. Erdem, Synthetic18k: Learning better representations for person re-ID and attribute recognit. from 1.4 million synthetic images, Signal Process., Image Commun. 97 (2021) 116335, http://dx.doi.org/10.1016/j.image.2021.116335, URL https://hucvl.github.io/synthetic18k/.

[7] EiC of Pattern Recognit, Expression of concern: "what-and-where to match: Deep spatially multiplicative integration networks for person re-identification" [pattern recognition, volume 76, april 2018, pages 727-738], Pattern Recognit. 121 (2022) 108134, http://dx.doi.org/10.1016/j.patcog.2021.108134.

[8] G. Dong, G. Liao, H. Liu, G. Kuang, A review of the autoencoder and its variants: A comparative perspective from target recognition in synthetic-aperture radar images, IEEE Geosci. Remote. Sens. Mag. 6 (3) (2018) 44–68, http://dx.doi.org/10.1109/MGRS.2018.2853555.

[9] F. Boutros, M. Huber, A.T. Luu, P. Siebke, N. Damer, SFace2: Synthetic-based face recognition with w-space identity-driven sampling, IEEE Trans. Biom. Behav. Identity Sci. (2024) http://dx.doi.org/10.1109/TBIOM.2024.3371502, 1–1.

[10] T. Zhang, L. Xie, L. Wei, et al., UnrealPerson: An adaptive pipeline towards costless person re-identification, in: Int. Conf. on Computer Vision and Pattern Recognit, CVPR, 2021, pp. 11506–11515, http://dx.doi.org/10.1109/CVPR46437.2021.01134, URL https://github.com/FlyHighest/UnrealPerson.

[11] E. Lomurno, M. Matteucci, Synthetic image learning: Preserving performance and preventing membership inference attacks, Pattern Recognit. Lett. (2025) http://dx.doi.org/10.1016/J.PATREC.2025.02.003.

[12] M.A. Souibgui, A. Fornés, Y. Kessentini, B. Megyesi, Few shots are all you need: A progressive learning approach for low resource handwritten text recognition, Pattern Recognit. 160 (2022) 43–49, http://dx.doi.org/10.1016/J.PATREC.2022.06.003.

[13] V.L. Kondarattsev, S.S. Krylov, N.P. Anosova, Creating a synthetic data generator for solving industrial flaw detection problems using deep learning methods, in: Advances in Theory and Practice of Computational Mechanics, Springer Singapore, Singapore, 2022, pp. 377–390, http://dx.doi.org/10.1007/978-981-16-8926-0_25.

[14] F. Gomez-Donoso, J. Castaño-Amoros, F. Escalona, M. Cazorla, Three-dimensional reconstruction using SFM for actual pedestrian classification, Expert Syst. Appl. 213 (Part) (2023) 119006, http://dx.doi.org/10.1016/J.ESWA.2022.119006.

[15] R. Delussu, L. Putzu, G. Fumera, On the effectiveness of synthetic data sets for training person re-identification models, in: Int. Conf. on Pattern Recognit, ICPR, 2022, pp. 1208–1214, http://dx.doi.org/10.1109/ICPR56361.2022.9956461.

[16] H. Li, M. Ye, B. Du, Weperson: Learning a generalized re-identification model from all- weather virtual data, in: ACM Multimedia Conf., 2021, pp. 3115–3123, http://dx.doi.org/10.1145/3474085.3475455, URL https://github.com/lihe404/WePerson.

[17] Y. Wang, S. Liao, L. Shao, Surpassing real-world source training data: Random 3D characters for generalizable person re-identification, in: ACM Conf. on Multimedia, 2020, pp. 3422–3430, http://dx.doi.org/10.1145/3394171.3413815, URL https://github.com/VideoObjectSearch/RandPerson.

[18] S. Xiang, D. Qian, M. Guan, et al., Less is more: Learning from synthetic data with fine-grained attributes for person re-identification, ACM Trans. Multim. Comput. Commun. Appl. 19 (5s) (2023) 173:1–173:20, http://dx.doi.org/10.1145/3588441, URL https://github.com/JeremyXSC/FineGPR.

[19] X. Sun, L. Zheng, Dissecting person re-identification from the viewpoint of viewpoint, in: Int. Conf. on Computer Vision and Pattern Recognit, CVPR, 2019, pp. 608–617, http://dx.doi.org/10.1109/CVPR.2019.00070, URL https://github.com/sxzrt/Instructions-of-the-PersonX-dataset.

[20] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C.H. Hoi, Deep learning for person re-identification: A survey and outlook, IEEE Trans. Pattern Anal. Mach. Intell. 44 (6) (2022) 2872–2893, http://dx.doi.org/10.1109/TPAMI.2021.3054775.

[21] F. Wan, Y. Wu, X. Qian, Y. Chen, Y. Fu, When person re-identification meets changing clothes, in: CVPR Workshops, 2020, pp. 3620–3628, http://dx.doi.org/10.1109/CVPRW50498.2020.00423, URL https://wanfb.github.io/dataset.html.

[22] A. Zheng, Z. Chen, C. Li, J. Tang, B. Luo, Learning deep RGBT representations for robust person re-identification, Int. J. Autom. Comput. 18 (3) (2021) 443–456, http://dx.doi.org/10.1007/s11633-020-1262-z.

[23] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer GAN to bridge domain gap for person re-identification, in: Int. Conf. on Computer Vision and Pattern Recognit, CVPR, 2018, pp. 79–88, http://dx.doi.org/10.1109/CVPR.2018.00016, URL https://www.pkuvmc.com/dataset.html.

[24] H. Zhao, S. Zhang, G. Wu, J.M.F. Moura, J.P. Costeira, G.J. Gordon, Adversarial multiple source domain adaptation, in: Advances in Neural Information Processing Systems, 2018, pp. 8568–8579.

[25] A. Li, J. Wu, Y. Liu, L. Li, Bridging the synthetic-to-authentic gap: Distortion-guided unsupervised domain adaptation for blind image quality assessment, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 28422–28431, http://dx.doi.org/10.1109/CVPR52733.2024.02685.

[26] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[27] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in: Int. Conf. on Neural Information Processing Systems, 2020.

[28] J. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Int. Conf. on Computer Vision, ICCV, 2017, pp. 2242–2251, http://dx.doi.org/10.1109/ICCV.2017.244.

[29] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, J. Choo, StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation, in: Int. Conf. on Computer Vision and Pattern Recognit, CVPR, 2018, pp. 8789–8797, http://dx.doi.org/10.1109/CVPR.2018.00916.

[30] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: Int. Conf. on Learning Representations, ICLR, 2016.

[31] D.H. Pham, A.D. Nguyen, H.N. Nguyen, GAN-based data augmentation and pseudo-label refinement with holistic features for unsupervised domain adaptation person re-identification, Knowl.-Based Syst. 288 (2024) 111471, http://dx.doi.org/10.1016/j.knosys.2024.111471.

[32] F. Croitoru, V. Hondru, R.T. Ionescu, M. Shah, Diffusion models in vision: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 45 (9) (2023) 10850–10869, http://dx.doi.org/10.1109/TPAMI.2023.3261988.

[33] F. Zhan, Y. Yu, R. Wu, et al., Multimodal image synthesis and editing: The generative AI era, IEEE Trans. Pattern Anal. Mach. Intell. 45 (12) (2023) 15098–15119, http://dx.doi.org/10.1109/TPAMI.2023.3305243.

[34] S. Xiang, Y. Fu, G. You, T. Liu, Unsupervised domain adaptation through synthesis for person re-identification, in: Int. Conf. on Multimedia and Expo, ICME, 2020, pp. 1–6, http://dx.doi.org/10.1109/ICME46284.2020.9102822.

[35] E. Yaghoubi, D. Borza, S.V.A. Kumar, H. Proença, Person re-identification: Implicitly defining the receptive fields of deep learning classification frameworks, Pattern Recognit. 145 (2021) 23–29, http://dx.doi.org/10.1016/j.patrec.2021.01.035.

[36] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, T. Vetter, Analyzing and reducing the damage of dataset bias to face recognition with synthetic data, in: CVPR Workshops, 2019, pp. 2261–2268, http://dx.doi.org/10.1109/CVPRW.2019.00279.

[37] I.B. Barbosa, M. Cristani, B. Caputo, et al., Looking beyond appearances: Synthetic training data for deep CNNs in re-identification, Comput. Vis. Image Underst. 167 (2018) 50–62, http://dx.doi.org/10.1016/j.cviu.2017.12.002, URL https://www.kaggle.com/vicolab/somaset.

[38] S. Xiang, Y. Fu, G. You, T. Liu, Taking a closer look at synthesis: Fine-grained attribute analysis for person re-identification, in: Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP, 2021, pp. 3765–3769, http://dx.doi.org/10.1109/ICASSP39728.2021.9413757.

[39] R. Delussu, L. Putzu, G. Fumera, Guidelines for query and gallery image extraction in person re-identification systems, in: Computer Vision - ECCV 2024 Workshops, Proceedings, Part XIII, Vol. 15635, 2024, pp. 223–236, http://dx.doi.org/10.1007/978-3-031-91575-8_14.

[40] S. Bak, P. Carr, J. Lalonde, Domain adaptation through synthesis for unsupervised person re-identification, ECCV, in: European Conf. on Computer Vision, Vol. 11217, 2018, pp. 193–209, http://dx.doi.org/10.1007/978-3-030-01261-8_12, URL https://github.com/swbak/SyRI.

[41] A. Zahra, N. Perwaiz, M. Shahzad, M.M. Fraz, Person re-identification: A retrospective on domain specific open challenges and future trends, Pattern Recognit. 142 (2023) 109669, http://dx.doi.org/10.1016/J.PATCOG.2023.109669.

[42] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Int. Conf. on Computer Vision, ICCV, 2015, pp. 1116–1124, http://dx.doi.org/10.1109/ICCV.2015.133, URL https://www.kaggle.com/datasets/pengcw1/market-1501.

[43] E. Ristani, F. Solera, R.S. Zou, et al., Performance measures and a data set for multi-target, multi-camera tracking, in: European Conf. on Computer Vision ECCV Workshops, 2016, pp. 17–35, http://dx.doi.org/10.1007/978-3-319-48881-3_2, URL https://exposing.ai/duke_mtmc/.

[44] J. Zhuo, Z. Chen, J. Lai, G. Wang, Occluded person re-identification, in: 2018 IEEE International Conference on Multimedia and Expo, ICME, 2018, pp. 1–6, http://dx.doi.org/10.1109/ICME.2018.8486568, URL https://github.com/tinajia2012/ICME2018_Occluded-Person-Reidentification_datasets.

[45] Y. Wang, X. Liang, S. Liao, Cloning outfits from real-world images to 3D characters for generalizable person re-identification, in: Int. Conf. on Computer Vision and Pattern Recognit, CVPR, 2022, pp. 4890–4899, http://dx.doi.org/10.1109/CVPR52688.2022.00485, URL https://github.com/Yanan-Wang-cs/ClonedPerson.

[46] Y. Ge, R. Zhang, X. Wang, X. Tang, P. Luo, DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images, in: Int. Conf. on Computer Vision and Pattern Recognit, CVPR, 2019, pp. 5337–5345, http://dx.doi.org/10.1109/CVPR.2019.00548.

[47] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognit, in: Int. Conf. on Computer Vision and Pattern Recognit, CVPR, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[48] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, TransReID: Transformer-based object re-identification, in: IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 14993–15002, http://dx.doi.org/10.1109/ICCV48922.2021.01474.

[49] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local Nash equilibrium, in: Int. Conf. on Neural Information Processing Systems, 2017, pp. 6626–6637.

[50] S.H.S. Hussin, R. Yildirim, StyleGAN-LSRO method for person re-identification, IEEE Access 9 (2021) 13857–13869, http://dx.doi.org/10.1109/ACCESS.2021.3051723.

[51] M. Kim, F. Liu, A.K. Jain, X. Liu, DCFace: Synthetic face generation with dual condition diffusion model, in: Int. Conf. on Computer Vision and Pattern Recognit, CVPR, 2023, pp. 12715–12725, http://dx.doi.org/10.1109/CVPR52729.2023.01223.