International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02102-x

Synthetic Data for Video Surveillance Applications of Computer Vision: A Review

Rita Delussu[1] · Lorenzo Putzu[1] · Giorgio Fumera[1]

Received: 28 June 2023 / Accepted: 22 April 2024

Abstract

In recent years, there has been a growing interest in synthetic data for several computer vision applications, such as automotive, detection and tracking, surveillance, medical image analysis and robotics. Early use of synthetic data was aimed at performing controlled experiments under the *analysis by synthesis* approach. Currently, synthetic data are mainly used for training computer vision models, especially deep learning ones, to address well-known issues of real data, such as manual annotation effort, data imbalance and bias, and privacy-related restrictions. In this work, we survey the use of synthetic training data focusing on applications related to video surveillance, whose relevance has rapidly increased in the past few years due to their connection to security: crowd counting, object and pedestrian detection and tracking, behaviour analysis, person re-identification and face recognition. Synthetic training data are even more interesting in this kind of application, to address further, specific issues arising, e.g., from typically unconstrained image or video acquisition conditions and cross-scene application scenarios. We categorise and discuss the existing methods for creating synthetic data, analyse the synthetic data sets proposed in the literature for each of the considered applications, and provide an overview of their effectiveness as training data. We finally discuss whether and to what extent the existing synthetic data sets mitigate the issues of real data, highlight existing open issues, and suggest future research directions in this field.

Keywords Synthetic data · Computer vision · Video surveillance applications · Crowd counting · Object and pedestrian detection and tracking · Behaviour analysis · Person re-identification · Face recognition

1 Introduction

Synthetic images have become very common in the field of Computer Vision (CV) for a large variety of real-world applications (Meharban et al. 2021; Dong et al. 201; Pratt et al. 1978; Frolov et al. 2021; Nikolenko 2021). To the best of our knowledge, they firstly appeared in the CV literature in a work by Pratt et al. (1978). Synthetic data were originally used to ease the analysis and comparison of image processing methods through the generation of images from a model under strictly controlled conditions (e.g., in terms of textures, materials and light), according to the so-called *analysis by synthesis* approach (Pratt et al. 1978; Horn and Schunck 1981; Woodham et al. 1985). Since then, the interest in synthetic data has considerably increased in CV, as well as in other fields, as witnessed by the ever-increasing number of publications (see image on top of Fig. 1). Their purpose has also remarkably changed: with the rise of machine learning and especially of Convolutional Neural Networks (CNNs), synthetic images are being mainly used as training data, either alone or together with real images (Jaipuria et al. 2020; Shang et al. 2018; Nikolenko 2021), with the aim of increasing the training set size, thus improving performances and preventing over-fitting. Collecting large training sets of real data, as required by CNNs, is indeed very difficult. This is due to several issues, particularly manual annotation effort, difficulty in collecting representative examples of the target scenes or patterns of interest and, more recently, privacy-related restrictions.

Communicated by Jingdong Wang.

✉ Rita Delussu
rita.delussu@unica.it

✉ Lorenzo Putzu
lorenzo.putzu@unica.it

✉ Giorgio Fumera
fumera@unica.it

[1] Department of Electrical and Electronic Engineering, University of Cagliari, via Marengo, 09123 Cagliari, Italy

Published online: 17 May 2024

1.1 Issues of Real Data

Manual annotation effort is a well-known limitation in collecting a large amount of real training data, especially in CV applications that require fine-grained supervision, e.g., crowd counting, where the position of each pedestrian's head has to be annotated for density-based CNN models. It can also lead to annotation errors or inaccuracy, e.g., wrong identity annotations in person re-identification (Re-Id) data sets (Zheng et al. 2015; Ergys et al. 2016); missing head annotations in dense crowd images, in crowd counting data sets (Sindagi et al. 2022); misplaced or wrong track annotations in people tracking data sets due to interpolation of ground truth among adjacent frames to save manual effort (Smeulders et al. 2014). Another relevant issue in several CV applications is the difficulty of collecting real, representative examples of the target scenes or patterns of interest, e.g., the abnormal events of interest for anomalous behaviour detection or the different crowd density levels of interest for crowd counting and density estimation. This can lead to two related but different problems: data bias and imbalance. For instance, a data set for anomalous behaviour detection may be balanced in terms of the number of examples of each kind of abnormal event collected during design, which, however, may be representative only of a limited set of the possible anomalous behaviours that can occur during operation; conversely, in other applications (e.g., crowd counting) it may be possible to collect examples of the target scenes (camera views) of interest, but they may be highly unbalanced in terms of factors such as crowd size, due to, e.g., lack of dense crowd images. A related issue is domain shift, which occurs in cross-scene settings where the target scenes are different from the ones used for training, causing a decrease in the effectiveness of supervised methods. Privacy-related restrictions are instead a more recent issue of applications involving images of people. For instance, the General Data Protection Regulation (GDPR), in force in the European Union since 2016, restricts the use of images or videos acquired from video surveillance cameras to investigation activities or to prevent acts of terrorism. This makes it even more difficult to collect real data sets for applications such as Re-Id and face recognition (Boutros et al. 2022).

To address the above issues, the use of synthetic data has been proposed by many authors for tasks such as crowd counting (Sindagi and Patel 2018; Li et al. 2021; Elbishlawi et al. 2020), Re-Id (Ye et al. 2022; Wu et al. 2019), crowd behaviour analysis (Júnior et al. 2010), face recognition (Guo and Zhang 2019) and object detection (Liu et al. 2020). All these applications may benefit (or have already benefited as can be seen from the image at the bottom of Fig. 1) from synthetic data, since they allow, in principle, to automatically generate a large number of images or videos, including all the possible scenes of interest, and to automatically provide high-quality and error-free annotations (Júnior et al. 2010), with no privacy issues (Guo and Zhang 2019).

Since early work in this field, not only the purpose of synthetic data but also the techniques to generate them have considerably changed. Besides basic data augmentation, nowadays, they include computer graphics tools and image-to-image translation methods based on generative adversarial networks (GANs), also involving real images (Shamsolmoali et al. 2021; Zhu et al. 2017; Tripathi et al. 2019) and diffusion models (Wu et al. 2023; Kim et al. 2023). In some cases, it is difficult to define a clear boundary between real and synthetic images, especially when real images are used to generate new ones, since often the original image is not substantially changed but is only manipulated, e.g., by flipping, cropping, etc. In that case, the term *augmentation* or *smart-augmentation* is also used (Nikolenko 2021). In the following, we shall use the term "real" image or video to refer to a signal produced directly by a sensor (as a 2D or 3D array) and the term "synthetic" or "artificial" to refer to data generated by computer processing.

1.2 Scope and Contribution of this Survey

Among the different applications of synthetic images, in this work, we focus on CV tasks related to *video surveillance* (VS), and in particular crowd counting, object and pedestrian detection and tracking, behaviour analysis, Re-Id and face recognition. Our focus on VS applications has two main motivations. First, their relevance is rapidly increasing due to their connection to security. Second, besides exhibiting nearly all the issues mentioned above, they also exhibit specific ones, for which the use of synthetic data appears even more promising, i.e., unconstrained image or video acquisition settings; cross-scene application scenarios, where a model (e.g., for crowd counting) has to be deployed to scenes different than the ones used for training; privacy restrictions in collecting data sets for model training and evaluation (e.g., in Re-Id and face recognition); difficulty in collecting patterns representative of all the categories of interest, e.g., different kinds of anomalous behaviours. With respect to existing surveys related to synthetic data for VS applications, this work covers more recent publications, and a more comprehensive set of applications and of synthetic image generation techniques, including image synthesis (see Sect. 2). We believe a survey of this rapidly evolving field can be very useful to understand its current evolution better and to provide insights for future work.

1.3 Data Sources

We reviewed work on synthetic images for VS tasks published from January 2014 to November 2023. We collected

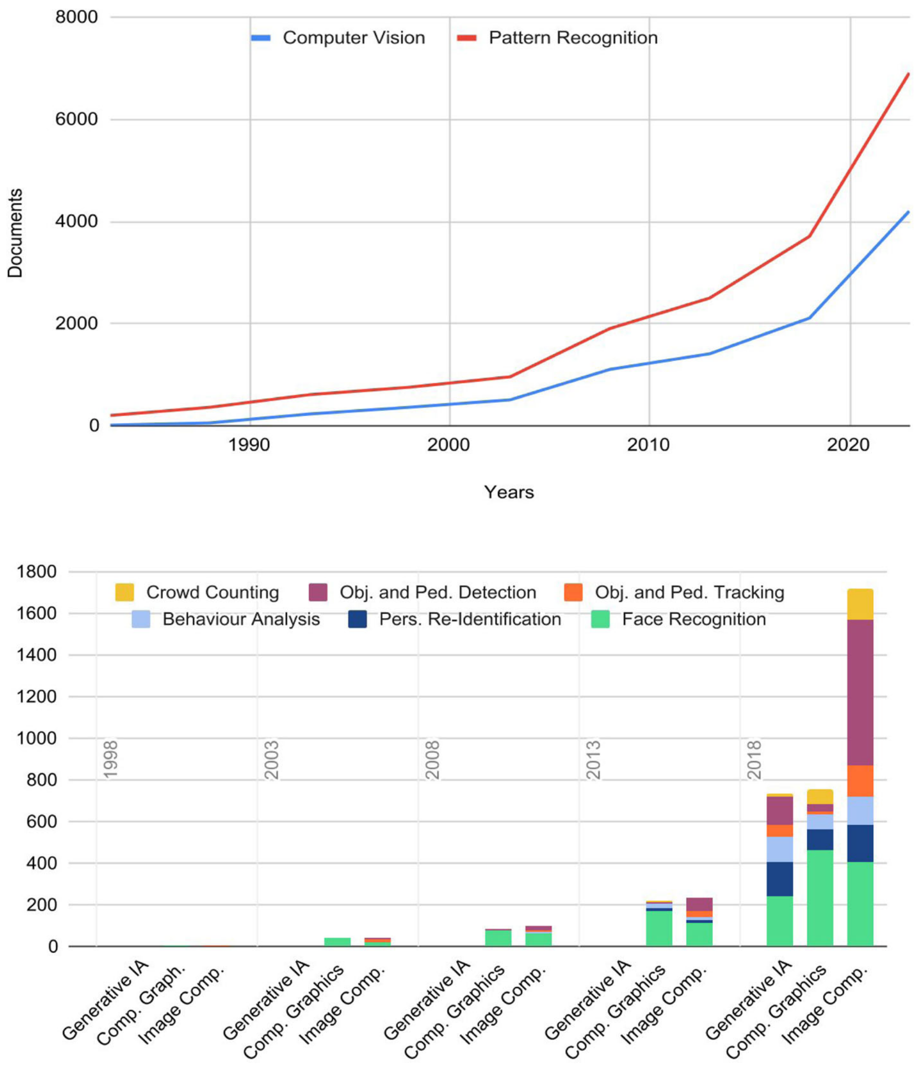

Fig. 1 Top: number of publications of the past 40 years related to synthetic images in Computer Vision (red) and in Pattern Recognition (blue). Bottom: number of publications related to synthetic image generation with Generative IA, Computer Graphics and Image composition methods for each video surveillance task over the past 25 years. We collected information from Scopus and considered any type of document (accessed on 16/11/2023) (Color figure online)

almost one hundred papers from the DBLP,[1] Scopus[2] and Web of Science[3] databases, using the following keywords: `synthetic OR synthesis AND images AND computer vision AND video surveillance`. We then selected publications from the following venues. Conference proceedings: Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), International Conference on Pattern Recognition (ICPR), International Conference on Robotics, Automation (ICRA), International Conference on Learning Representations (ICLR), International Conference on Image Processing (ICIP), International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), Winter Conference on Applications of Computer Vision (WACV), European Conference on Computer Vision (ECCV), European Signal Processing Conference (EUSIPCO), International Conference on Computer Analysis of Images and Patterns (CAIP), International Conference on Multimedia, International Joint Conference on Biometrics (IJCB), Conference on Neural Information Processing Systems (NeurIPS) and International Conference on Automatic Face and Gesture Recognition. Journals: IEEE Trans. on Pattern Analysis and Machine Intelligence, IEEE Trans. on Information Forensics and Security, IEEE Trans. on Circuits, Systems and Video Technology, Computer Vision and Image Understanding, Pattern Recognition, Neurocomputing, International Journal of Computer Vision, Neural Computing and Applications, Neural Processing Letters, Neural Network, Pattern Recognition Letters, IET Image Processing.

1.4 Paper Structure

In Sect. 3 we give a comprehensive overview of the VS applications mentioned above, focusing on issues that motivate the

[1] https://dblp.uni-trier.de/db/.

[2] https://www.scopus.com/search/form.uri?display=basic#basic.

[3] https://www.webofscience.com/wos/woscc/basic-search.

use of synthetic data. We then survey in Sect. 4 existing work on synthetic data for such applications, including a description of the techniques used to generate them. In Sect. 5, we describe and analyse all the existing synthetic data sets for the above applications and compare them with benchmark data sets of real data. In Sect. 6, we analyse the effectiveness of synthetic data in data enhancement and representation. Finally, in Sect. 7 we discuss the extent to which synthetic data solve or mitigate the existing issues of real data in the considered applications and propose possible solutions and directions for future work in this field.

2 Related Work

This survey is the first one focused on synthetic data for CV applications related to VS. Previous surveys on synthetic data focused on different aspects (Nikolenko 2021; Li 2021; Shamsolmoali et al. 2021; Zhou et al. 2021; Frolov et al. 2021; Abdolahnejad and Liu 2020; Meharban et al. 2021; Dong et al. 201; Croitoru et al. 2023). A recent book by Nikolenko (2021) provided a wide survey of synthetic data focused on their use in deep learning techniques, which was not limited to the CV field, and did not provide an in-depth analysis of VS applications, on the related synthetic data sets and on the synthetic data generation techniques proposed so far in this field, except for GANs. Other surveys on synthetic data focused on specific image synthesis approaches or on single applications: image synthesis and editing techniques based on GANs (Li 2021) and other adversarial networks (Shamsolmoali et al. 2021), text-to-image synthesis (Zhou et al. 2021; Frolov et al. 2021), diffusion models (Croitoru et al. 2023), face recognition (Boutros et al. 2023b) or face synthesys (Abdolahnejad and Liu 2020), medical images (Meharban et al. 2021) and aperture radar (Dong et al. 201).

On the other hand, the use of synthetic data has been mentioned by existing surveys on CV applications related to VS, but often in a limited way: object (car and pedestrian) detection (Liu et al. 2020) and tracking (Sun et al. 2021; Smeulders et al. 2014), Re-Id (Ye et al. 2022; Wu et al. 2019; Karanam et al. 2019; Zahra et al. 2023), face recognition (Wang and Deng 2021; Guo and Zhang 2019) and crowd counting (Li et al. 2021; Saleh et al. 2015; Leng et al. 2020; Júnior et al. 2010).

Differently from the above-mentioned work, our survey thoroughly analyses the use of synthetic images in the most relevant VS applications of CV. In particular, we first point out the specific and often shared issues of such applications that further motivate the use of synthetic data; we describe the different approaches to synthetic image generation proposed so far, and finally, we analyse the existing data sets of synthetic images and evaluate the effectiveness of synthetic training data with respect to benchmark data sets of real images.

3 Applications

In this section, we describe the VS applications of CV considered in our survey, i.e., crowd counting, object and pedestrian detection and tracking, human and crowd behaviour analysis, Re-Id and face recognition, together with the related state-of-the-art approaches. Our goal is to point out their open issues, and in particular, the ones that can be addressed (at least in principle) or have already been addressed, using synthetic data. Common issues across all the considered applications have already been discussed in Sect. 1, i.e.: manual annotation effort of real training data; limited representativeness of such data, causing, in turn, data bias and imbalance, and domain shift issues; privacy restrictions for applications involving images of people.

3.1 Crowd Counting

Crowd counting consists of estimating the number of people in a given image or video frame. This task is useful, e.g., to law enforcement agencies for monitoring crowds in public spaces, especially for large and dense crowds. State-of-the-art methods are based on CNNs (Sindagi and Patel 2017) and use either a regression or a detection approach (Sam et al. 2020). The former consists of first regressing the crowd density map and then estimating the corresponding count by simple integration over such a map (i.e., a pixel-wise sum). The latter performs firstly pedestrian detection. To take into account occlusions between people, which occur mainly in dense crowd images, and tend to increase as the camera height and its inclination relative to the ground decrease, both approaches require a training set of crowd images with manual annotation of the head position of each pedestrian; in particular, the latter approach usually performs head instead of full body detection. Therefore, the corresponding manual annotation effort is considerable, limiting the size of real data sets; it may also affect the quality of annotations. For instance, videos of the benchmark WorldExpo’10 and CityStreet data set (Zhang et al. 2016; Zhang and Chan 2019) (see Sect. 5.1) contain a manually annotated frame every 15 or 30 s, whereas the other frames have been automatically annotated by interpolating between two consecutive, manually annotated ones. This leads to low-quality annotations in terms of missing or misplaced head positions; for instance, this is particularly evident in the images of the CityStreet data set (Zhang and Chan 2019) (see Fig. 2).

Another issue, partly related to the one discussed above, is the difficulty of collecting real data sets representative of all the target scenes where a crowd counting system may be deployed. Crowd counting models are indeed affected by a mismatch between training and testing data in terms of several specific factors, such as camera viewing angle (e.g., from almost horizontal views of fixed or PTZ cameras placed at a low height to vertical views of cameras mounted on drones), scale and perspective variations, scene background (Li et al. 2021), lighting conditions (which can exhibit large variations in outdoor scenes typical of surveillance applications, e.g., from daytime to night, or under different weather conditions like rain, snow and fog), as well as crowd size and density level (Sindagi and Patel 2018; Wang et al. 2021). Since it is very difficult to collect and manually annotate crowd images representative of all possible scenes of interest in terms of the above factors (Sindagi et al. 2020), real data sets may exhibit a considerable degree of imbalance in favour of some kinds of scenes (see Fig. 2). This can lead to poor quality head localisation and thus inaccurate estimates of crowd count during inference, especially for dense crowds (Sindagi and Patel 2018). To mitigate this issue, multi-scene data sets have been proposed, made up of images from different scenes, e.g., crawled from the web (see Sect. 5.1) (Zhang et al. 2016b; Idrees et al. 2018a; Sindagi et al. 2022; Wang et al. 2021). Despite relatively large variations in terms of background, perspective and number of people, even these data sets are clearly unbalanced, e.g., street views are predominant, as well as medium-density scenes with respect to low and high-density ones (see Fig. 2). Furthermore, to achieve a certain degree of scene variation, multi-scene data sets are not collected from VS camera networks, and therefore they contain images that are not representative of realistic surveillance scenes, in terms of image perspective (e.g., group photo or photo memories are present), quality and content (e.g., some images contain trademarks) (Sindagi et al. 2022; Wang et al. 2021) (see Fig. 2).

3.2 Object and Pedestrian Detection

Detection tasks consist of locating instances of interest from pre-defined object categories inside an image, mostly by enclosing them inside a bounding box (BB) (Liu et al. 2020). Pedestrians and vehicles (e.g., car and bus) are among the categories of greater interest for surveillance applications; in particular, pedestrians belong to the category of articulated objects, which is the most challenging one for detection tasks. The detection task is useful for several surveillance applications, such as perimeter protection, traffic monitoring, and access monitoring to restricted areas; it can also be a preliminary step for other tasks, such as tracking (see Sect. 3.3) and Re-Id (Sect. 3.5).

Regardless of the detection method used, the annotation of training images consists of drawing a BB around each object instance of interest and of associating it with the corresponding label (if more than one object category is considered). In particular, BBs should be as precise (tight) as possible to limit the influence of the background.

Object detection is challenging in the presence of deformations (for non-rigid objects), scale and viewpoint variations, blur, motion, and low image resolution. Additionally, structured objects may exhibit considerable intra-class variations, e.g., in shape, colour, material and size (e.g., for vehicle detection, different car models or even different instances of the same model); pedestrians' appearance can also exhibit variations in pose and attributes (e.g., bags and hats). These issues might lead to inaccurate localisation, e.g., non-tight BBs, including background regions.

Despite the limited number of categories considered in object detection for VS applications (mainly pedestrians and vehicles), the annotation issue is still present. Indeed it can be infeasible to annotate all possible objects, and in most data sets, images or frames are only partially annotated (Liu et al. 2020).

3.3 Object and Pedestrian Tracking

Tracking consists of inferring one or more objects' motion in a sequence of frames (Smeulders et al. 2014). It is, therefore, inherently related to *video* analysis, contrary to object detection. The main tracking approaches can be categorised as single- and multi-object tracking according to the number of objects that should be tracked. This task is very useful for several surveillance applications, such as traffic monitoring and pedestrian behaviour analysis (see Sect. 3.4). As mentioned before, some tracking methods exploit instance detection as a sub-task (Sun et al. 2021): in this case, their effectiveness also depends on detection accuracy. Training videos for object tracking require manual annotation of the position (BB coordinates) of each instance of interest in every frame and an identity label that should be consistent over the whole video. The corresponding annotation effort is, therefore, considerable, especially for multi-object tracking. Tracking and detection tasks share most of the issues discussed in Sect. 3.2, related to scale and viewpoint variations and intra-class object variations. Additionally, tracking presents further challenges due to, e.g., sudden object motion, and significant scene changes in case of camera motion (e.g., when videos are acquired from drones) (Abbass et al. 2021; Sun et al. 2021).

A further issue is that, with respect to other applications, a limited number of benchmark data sets is available for object and pedestrian tracking (Smeulders et al. 2014).

3.4 Human and Crowd Behaviour Analysis

Pedestrian and crowd behaviour analysis aims at recognising predefined behaviours of interest, or at detecting abnormal events at the level of either individual, small groups or crowds

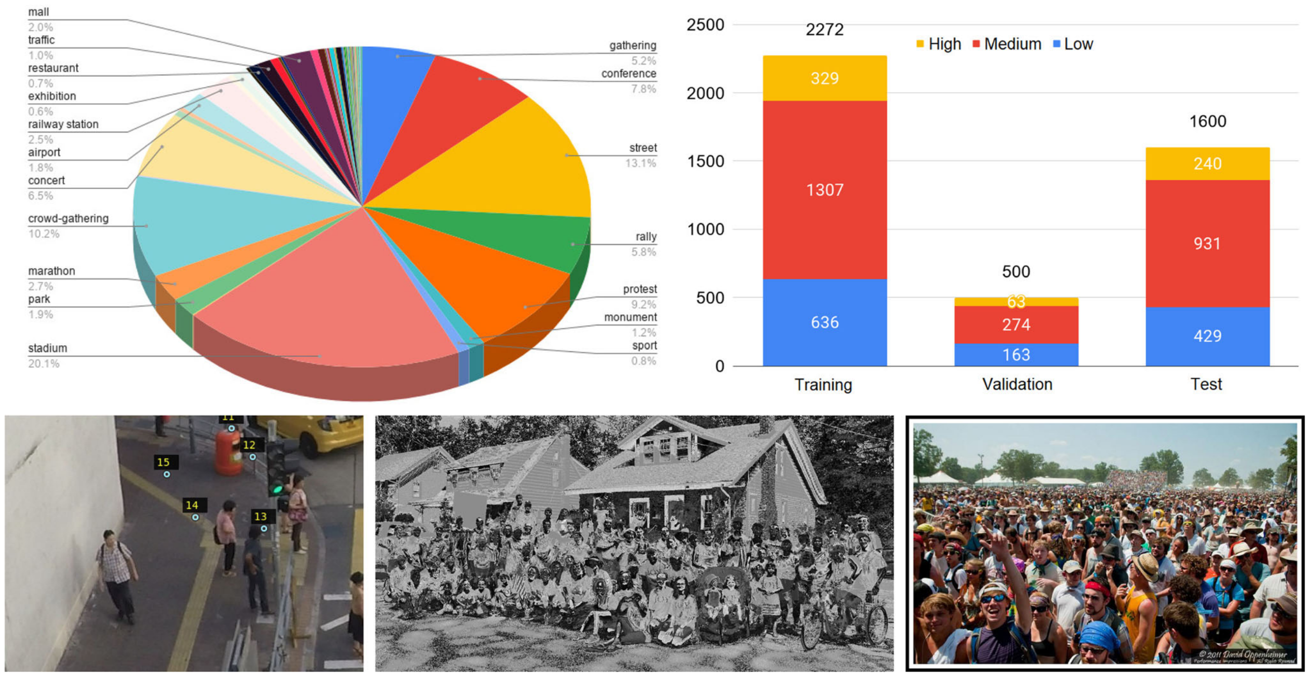

Fig. 2 Examples of the issues of real data sets for crowd counting. Top row: data imbalance in terms of scenes and crowd density [source: JHU (Sindagi et al. 2022)]. Bottom row, from left to right: low quality annotations [source: Citystreet (Zhang and Chan 2019)], unrealistic images due to perspective and content [group photo and trademarked photo; source: ShtechA (Zhang et al. 2016a)]

(Sánchez et al. 2020). This can be very useful in surveillance applications, e.g., for preventing or for quick response to violence, crimes and other dangerous or anomalous behaviours, such as anomalous patterns of crowd flow in crowded places or events (Schroder et al. 2019).

Well-known issues of anomalous behaviour detection are the lack of a universal definition of *anomaly*, which instead strongly depends on the application as well as on the context (Li et al. 2014), and the difficulty of enumerating beforehand all possible events or behaviours that can be considered as anomalous in a specific application. This directly affects the representativeness of benchmark data sets, which as a consequence contain only a limited number of labelled anomalous scenes and behaviours (Li et al. 2014; Tripathi et al. 2019). For instance, most data sets focus on fights or violent scenes and mainly consist of videos collected from the Web, often acquired during sporting events (e.g., hockey matches). Other data sets are made up of simulated scenes performed by volunteers, making them more complete but at the same time less realistic (Sánchez et al. 2020).

Additionally, anomalous behaviours could extend for indefinite and even very long frame sequences; they can also co-occur with other, possibly unrelated unusual events. In some data sets, every frame of a video is annotated (Chan et al. 2008); in other data sets, the whole video has a single annotation corresponding to the anomalous behaviour occurring in it, although it extends over a few frames only (Sultani et al. 2018).

For the above reasons, behaviour analysis, particularly anomalous behaviour detection, is probably the most challenging CV task in surveillance applications. Moreover, since it involves detection and tracking as preliminary sub-tasks (Sánchez et al. 2020), it inherits the respective issues discussed in Sects. 3.2 and 3.2.

3.5 Person Re-Identification

Person re-identification aims at matching images of people across different and non-overlapping camera views, mainly based on clothing appearance (including attributes like bags) (Karanam et al. 2019). Its main goal is to reduce the human effort in monitoring tasks and the time required to inspect recorded videos to search for individuals of interest during investigations. In the latter case, it allows searching for an individual of interest, typically using a query image, on a large gallery of pedestrian bounding boxes automatically extracted from a given set of videos.

Re-Id is inherently a *cross-scene* task since the individual in the query image has to be searched among images acquired by *different* camera views. Accordingly, benchmark data sets contain images of pedestrians from at least two different camera views. On the one hand, data set collection and manual annotation for this task require considerable effort.

Indeed, Re-Id data sets are made up of bounding boxes of pedestrians, mostly manually drawn, with at least two images of each identity from at least two different cameras. Moreover, images of each identity need to be manually annotated with the same ID label (Ye et al. 2022). On the other hand, cross-scene issues considerably affect the accuracy of Re-Id systems due to, e.g., different camera settings (perspective, resolution, and colours), lighting conditions and background, besides issues related to, e.g., changes in pedestrian pose. In addition, since during operation the gallery of pedestrian bounding boxes is automatically populated using a pedestrian detector or tracker, the accuracy of Re-Id systems is also affected by errors in these sub-systems.

Two main issues of existing data sets are that: (i) their volume is too limited with respect to real applications involving several different cameras and hours of recording, and (ii) they do not exhibit a sufficiently large variation in weather conditions, lighting, scale, viewpoint and scene background (Leng et al. 2020; Saleh et al. 2015; Liu et al. 2020). For instance, most of them only contain daytime images acquired from RGB cameras, whereas thermal cameras can also be considered to cover night-time application scenarios (Zheng et al. 2021). Furthermore, recent privacy regulations (Guo and Zhang 2019; Uner et al. 2021) impose restrictions on the acquisition of images that allow distinguishing characteristics peculiar to individuals (EiC of Pattern Recognition 2022), which hinders the collection of real data sets for this task.

3.6 Face Recognition

Face recognition aims to identify or verify the identity of an individual (Wang and Deng 2021). Verification, used, e.g., in mobile devices, consists of matching a face image against a single template of the claimed identity stored in the device. In contrast, identification consists of matching a face image against all images (different templates) in a database.

This task presents specific issues such as blur, especially in unconstrained settings, as well as variations in facial expression and pose. The annotation of training images mainly consists of the person's identity, but since most of such data sets are also used for other tasks (e.g., facial emotion recognition and facial age estimation), they also require the annotation of landmarks or keypoints for eyes, nose, mouth, etc.

Existing face recognition data sets present several kinds of biases related to gender, age, and in particular demography. Indeed, most images show Caucasian people, whereas other races are significantly less represented. Moreover, among all the surveillance-related applications, face recognition is the one most affected by privacy issues (European Union Agency for Fundamental Rights 2019). For this reason, several benchmark data sets do not contain daily life images but are made up, e.g., of photos of celebrities, which makes them not representative of typical operating conditions (Wang and Deng 2021; Li et al. 2020).

4 Approaches to Synthetic Data Generation

In the previous section, we pointed out the main issues of the considered VS tasks that can be addressed using synthetic data. This section describes the synthetic data generation approaches proposed so far for such tasks. In contrast to the previous sections that were structured according to VS tasks, this one is structured according to the data generation approaches, since each of them has been used for different tasks. In particular, we identified three main categories of approaches: generative models, computer graphics engines, and image composition. The latter consists in combining different images or image patches, including real ones. We do not consider basic data augmentation techniques (e.g., random image flipping or cropping) widely used to mitigate over-fitting, although they can be considered as a further, basic approach for producing synthetic images.

4.1 Generative Models

Generative models aim to generate artificial data (e.g., images); they can be categorised into Generative Adversarial Networks (GAN) and Diffusion Models (DM).

4.1.1 Generative Adversarial Network

The GAN-based approach aims to generate synthetic images through CNNs using an adversarial learning process. The most straightforward GAN architecture is based on a generator network ($\mathcal{G}$), aimed at generating synthetic images, and a discriminator network ($\mathcal{D}$), aimed at distinguishing real from "fake" (generated) ones (Goodfellow 2017). The adversarial learning process consists of training $\mathcal{G}$ and $\mathcal{D}$ in parallel, letting $\mathcal{G}$ generate new images until $\mathcal{D}$ cannot distinguish between real and generated ones (Shamsolmoali et al. 2021).

Different kinds of GANs have been developed; the main categories are CycleGANs (Zhu et al. 2017), StarGANs (Choi et al. 2018), and Deep Convolutional GANs (DCGANs) (Radford et al. 2016). All the above kinds of GANs have been used in the literature for generating synthetic data (besides basic data augmentation) in many of the considered CV tasks, especially for Re-Id (Zheng et al. 2021; Han et al. 2021), face recognition (Farooq et al. 2021; Li et al. 2021) and object detection (Sultana et al. 2020). In the following, we briefly describe them and their application to the considered CV tasks (Fig. 3).

CycleGANs are made up of two generators that learn two mapping functions from an image domain $\mathcal{X}$ to a different

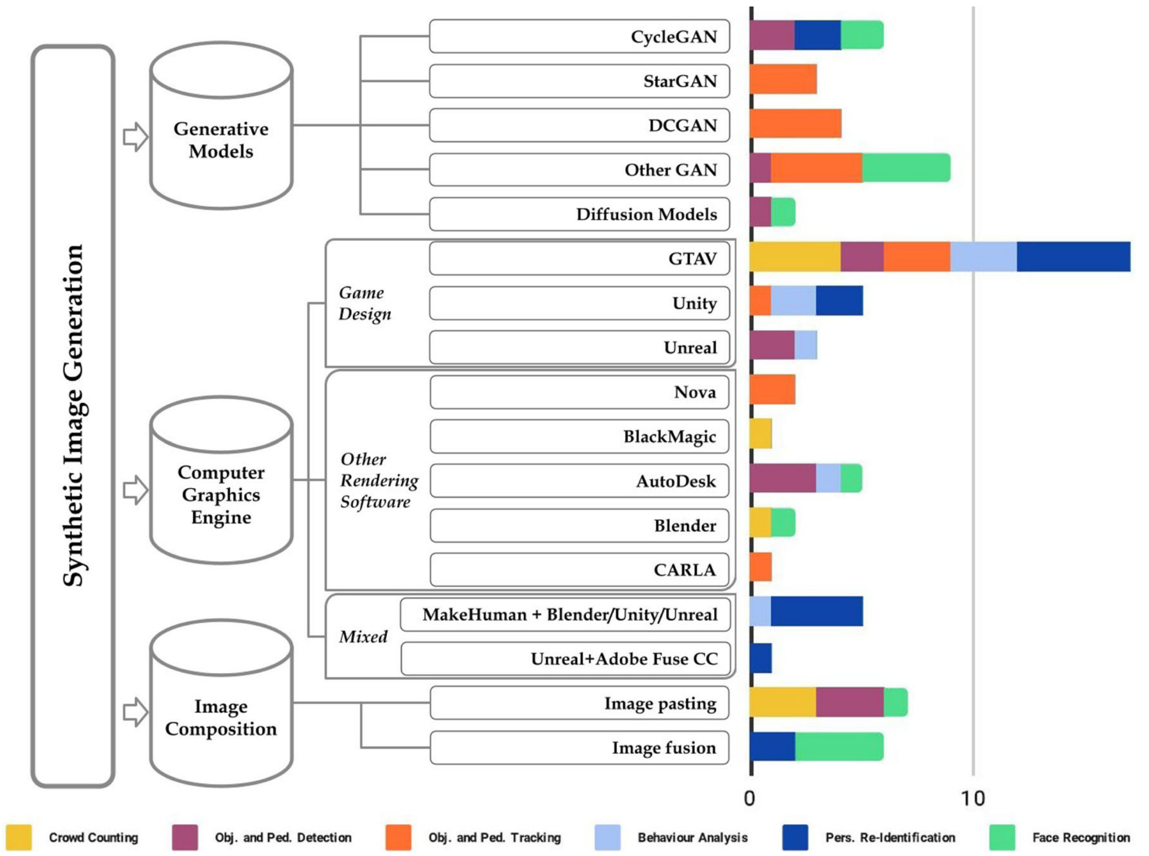

Fig. 3 Categorisation of synthetic methods for image and video generation, and chart of the respective number of papers where they have been used for the different VS tasks (denoted by colours) (Color figure online)

image domain $\mathcal{Y}$ and vice versa, and two discriminators $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$ aimed at distinguishing between real and generated images. CycleGANs have been used for Re-Id to improve accuracy under low light conditions (e.g., at night time) through the use of thermal infrared images by Zheng et al. (2021). Due to the limited number of thermal infrared data sets, a CycleGAN model was trained to translate RGB images into thermal ones. The trained GAN was then used to generate thermal images of data sets which do not contain this kind of data. Finally, a trained model can extract and combine multi-modal information to improve Re-Id accuracy. CycleGANs have been considered in a similar scenario and with an analogous objective for object detection by Guo et al. (2019). Another method devised for object detection by Sultana et al. (2020) proposed to employ CycleGAN to generate several background images of the target scene. This approach has been proposed to detect moving objects in the target scene. In face recognition, instead, CycleGANs have been used in an online setting to translate low-quality images, which are typically provided by surveillance cameras, into high-quality ones, to improve recognition accuracy (Farooq et al. 2021).

CycleGANs can modify image colour and texture but cannot perform shape transformation (e.g., of a car image into a truck one); they also exhibit limited scalability, since they can learn a mapping between only two domains simultaneously. *StarGANs* overcome the latter issue by learning the mappings between any set of domains of interest. To this aim, a single, flexible generator was used by Choi et al. (2018), which learns to translate an input image into the desired target domain; the adversarial learning process was implemented using an auxiliary classifier on top of the discriminator to manage multiple domains. StarGANs have been used in Re-Id for data augmentation with different translation strategies (Liu et al. 2019; Tian et al. 2021; Zhang and Hu 2023): Intra-data set (Liu et al. 2019), where the pedestrians' images acquired from a certain camera are translated to look similar to images acquired from a different camera in terms of background, illumination, etc.; inter-data set (Tian et al. 2021), where the images of a *source* data set are translated to look similar to images acquired from a different *target* data set. The aim is to produce images with the style of the target data set while preserving the identity of the corresponding person in the source data set. A combination of the two above strategies is employed by Zhang and Hu (2023) to bridge the gap between different domains.

A common issue of GANs is that their training process is challenging for two main reasons: the generator might collapse, i.e., it can always generate the same output for different inputs; and, often, GANs training fails to converge, meaning that the local equilibrium is far from the global one (Shamsolmoali et al. 2021). To address these issues, *DCGANs* have been proposed. They present significant differences with respect to a "classic" GAN architecture, i.e., different GAN layers are removed (e.g., hidden fully connected layers) or replaced with different ones (e.g., pooling layers are replaced by convolutional ones). Several authors have used DCGANs for data augmentation in Re-Id (Ainam et al. 2019b; Ding et al. 2019; Ainam et al. 2019a; Ding et al. 2018). In these works, a framework which includes k-means clustering algorithm and DCGAN has been proposed. Firstly k-means is applied to the features extracted from training images through an inner layer of the CNN. The obtained clusters are exploited to determine the network's most representative features and for data augmentation via DCGANs by generating new images for each cluster. Finally, generated and real images of the same domain are used to learn the model.

Several other kinds of GANs have also been used for Re-Id (Wu et al. 2021; Hussin and Yildirim 2021; Chen et al. 2019; Verma et al. 2023), face recognition (Mokhayeri et al. 2020; Qiu et al. 2021; Saez-Trigueros et al. 2021; Wood et al. 2021) and object detection (Lin et al. 2023). In the Re-Id task GANs are mainly used for data augmentation and to prevent over-fitting during training. In particular, Deep adversarial data augmentation with attribute-guided (DADAA) (Wu et al. 2021) allows attaining new pedestrian images with different colours of the upper body of pedestrians. *StyleGAN* (Hussin and Yildirim 2021) is used to generate images with different styles in terms of resolution, background, colour, illumination and poses for Re-Id. Such styles are estimated by grouping the images of the same data set. Context Rendering GAN (*CRGAN*) (Chen et al. 2019) have been proposed for a similar task, but in this case, the images are translated from a source data set to a target one. Individual-preserving and environmental-switching cyclic generation network (*IPES-GAN*) generates new images by cropping a human body and adding it to a target background image, while preserving the pedestrian's identity and pose.

Also in face recognition, GAN-based approaches are mainly used for data augmentation. In particular, in Mokhayeri et al. (2020) the augmentation process focuses on face pose. The source (real) images are transformed into a 3D image using a 3D model to simulate a specific pose. Then, these 3D images are used to train the Controlled GAN (*CGAN*) that aims to refine the realism of generated images and adapt them to the target poses. In Saez-Trigueros et al. (2021) a *conditional GAN* has been used both to increase the number of samples per identity (different facial expression and head pose) and to generate new identities. Such an approach exploits an embedding module that combines identity-related attributes and random non-identity-related attributes.

Similarly, in Qiu et al. (2021), a 3D identity mixup module is incorporated into a GAN, to generate images of virtual people with *DIS*entangled, precisely-*CO*ntrollable (DiscoFaceGAN) latent representations for the identity of non-existing people, expression, pose, and illumination. Finally, an interpolation technique between two individual images is used to increase the number of identities. Likewise, Karras et al. (2021) used a style-based GAN to obtain new face images by combining existing ones.

For object detection, a cycle-object-consistent image-to-image translation network has been proposed by Lin et al. (2023) to bridge the gap between different domains.

4.1.2 Diffusion Models

Diffusion model (DM) approaches have been recently proposed in the literature (Croitoru et al. 2023). DM has been used in several CV tasks (Azizi et al. 2023; Trabucco et al. 2023; Rombach et al. 2022) (e.g., image classification), but in a few VS tasks (Kim et al. 2023; Wu et al. 2023). DM aims to generate synthetic data by using a different approach with respect to GANs, based on the concept of "diffusion" to simulate data generation. Basically, the generation of synthetic samples consists of two phases (Ho et al. 2020; Nichol and Dhariwal 2021). The first phase, called "forward diffusion", applies a sequence of invertible transformations to "diffuse" the sample until it reaches the desired complex data distribution. It is like gradually adding noise (e.g., Gaussian) until the original data distribution is corrupted. The second phase, called "reverse diffusion", maps the obtained distribution back to the original data distribution through a sequence of inverse transformations. To this aim, the training process consists of learning the parameters of the invertible transformations, preceded by a standardisation to convert the data into a distribution with zero mean and unit variance (typically, a Gaussian distribution). In other words, a DM first corrupts the input image by progressively adding noise and then it learns to reconstruct it.

So far, works on DM models mostly focused on the evaluation of the quality of generated data (Croitoru et al. 2023); only few of them considered their use for generating training data (Kim et al. 2023; Wu et al. 2023). In particular, Kim et al. (2023) proposed an image-to-image DM approach and then used the generated images to train face recognition models. The approach consists of two stages named sampling and mixing. In the former, a face image is generated and a specific style is selected. In the latter, an image with the selected style is generated. This process is repeated several times to obtain a set of images, Another work focused on the creation of a tool to generate annotated data for detection tasks by using text-to-image DMs (Wu et al. 2023). In the first step, latent code and text-image representation obtained by DM are used to train a perceptual decoder. In the second step, a large language model generates several (diverse) prompts that aim at generating images while the perceptual decoder generates annotations. In particular, that method allows the generation of several annotations such as human pose, depth and semantic mask, instance mask and deep-fashion mask. Although DM approaches produce high-quality data, they require multiple steps to generate a single sample and therefore, they present a lower synthesis speed with respect to GANs (Croitoru et al. 2023).

4.2 Computer Graphics Engines

Software platforms used in game design and rendering tasks are currently capable of generating images and videos of various kinds of scenes with a high level of photo-realism. For this reason, they are also being used to generate synthetic data for various CV tasks. In particular, the two most famous game design platforms, Unity Unity Technologies (n.d.) and Unreal Engine Epic Games (n.d.) (Unreal for short), have been used to this aim, as well as the Script Hook V library, which allows using (offline) the video game Grand Theft Auto V (GTAV) Blade (n.d.); among rendering software, NOVA (Kerim et al. 2021a), BlackMagicDesign Blackmagic Design (n.d.) (BlackMagic for short), Blender Blender Online Community (n.d.), AutoDesk Autodesk Inc. (n.d.), Abobe Fuse CC Adobe (n.d.) (Adobe for short), MakeHuman (Community 2020), CARLA (Dosovitskiy et al. 2017) have been used.

Among game design platforms, Unreal is suited especially for developing large games, thanks to its fast rendering and memory and resource optimisation; however, it has a limited user community. Unity has a larger community instead, but its rendering is slower than Unreal and more suited to small and mid-size games. Both platforms are licence-free, although with some limitations. Images generated using the Script Hook V library allow the user to use the native functions of GTAV.

Concerning rendering software, Blender allows 3D modelling and 3D animation but is time-consuming in rendering. Autodesk and MakeHuman are effortless (with respect to other rendering software) and present a user-friendly interface. However, the former requires significant hardware resources, whereas the latter can only generate 3D human models but not complete scenes. CARLA is a simulator (based on Unreal) developed for autonomous driving.

To our knowledge, the above-mentioned software has been used to generate synthetic data in all the considered CV applications, with the exception of game design platforms for face recognition. In particular, synthetic data have been used mainly to train CV models without real data and to generate new benchmark data sets. Among game design platforms, *GTAV* has been used to generate bounding boxes of pedestrians from different scenes for Re-Id (Wan et al. 2020; Xiang et al. 2020), videos of normal and abnormal events for behaviour analysis (Lazaridis et al. 2018; Montulet and Briassouli 2020; Lin et al. 2021), and scenes for object detection (Ciampi et al. 2020; Johnson-Roberson et al. 2017) and tracking (Richter et al. 2017; Fabbri et al. 2018), as well as for crowd counting (Zhao et al. 2020; Zhang et al. 2021; Wang et al. 2021, 2019). Besides automatic generation and annotation of images and videos, GTAV has also been used for generating more challenging scenarios than the ones present in real data (e.g., night scenes and low light conditions) (Zhao et al. 2020), and to obtain a balanced data set in terms of weather conditions (e.g., clear and raining) (Montulet and Briassouli 2020). Finally, GTAV has also been used to generate benchmark data sets for crowd counting (Zhang et al. 2021; Wang et al. 2019), object detection (Ciampi et al. 2020; Johnson-Roberson et al. 2017) and tracking (Richter et al. 2017; Fabbri et al. 2018), crowd behaviour analysis (Montulet and Briassouli 2020) and Re-Id (Wan et al. 2020; Wang et al. 2019) (see Sect. 5). Similarly, *Unity* and *Unreal Engine* have been used to generate scenarios for object detection (Jaipuria et al. 2020; Linder et al. 2020) and benchmark data sets for Re-Id (Uner et al. 2021; Sun and Zheng 2019), behaviour analysis (Cheung et al. 2019; de Souza et al. 2017) and tracking (Cabon et al. 2020) (see Sect. 5).

With regard to rendering software, *NOVA* has been used to simulate adverse weather conditions for tracking (Kerim et al. 2021b) and to generate benchmark data sets for behaviour analysis (Courty et al. 2014), tracking (Kerim et al. 2021b) and face recognition (Wood et al. 2021); *AutoDesk* has been used to create scene-specific images for object and pedestrian detection (Hattori et al. 2015, 2018; Aranjuelo et al. 2021), to include more challenging views for the same task (Aranjuelo et al. 2021), and to generate a synthetic data with a small domain gap with respect to real images for face recognition (Wood et al. 2021); *BlackMagic* has been used to create more representative training images for crowd counting (Ghosh et al. 2017). *CARLA* has been used to create scenes of interest using pre-installed or pre-defined maps of different cities, allowing the user to set several parameters, e.g., weather and light conditions (Dosovitskiy et al. 2017).

Synthetic data generated through Blender have also been used for analysis purposes (Ledda et al. 2021; Han et al. 2020). In our previous work (Ledda et al. 2021), we used synthetic training images generated with Blender to evaluate how different degrees of realism in specific scene characteristics, i.e., background and human models, affect the performance of state-of-the-art CNN-based crowd counting models. Similarly, 3D synthetic images generated with Blender have been used to improve the robustness of face recognition models against occlusions (e.g., due to sunglasses) or variations in terms of scale, background etc (Han et al. 2020).

Finally, some authors exploited MakeHuman or Adobe to generate 3D human models, thanks to their ease of use for this purpose (see above), together with another rendering software to generate the background scene, for Re-Id (Wang et al. 2020; Zhang et al. 2021; Barbosa et al. 2018; Bak et al. 2018; Wang et al. 2022) (Blender, Unreal or Unity) and behaviour analysis (Villamizar et al. 2020) (Blender).

4.3 Image Composition

The last category of synthetic image generation approaches for the considered CV tasks is based on combining several real or synthetic images, without using GANs or computer graphics software. Images generated with this approach have been used to build fully synthetic training sets (Delussu et al. 2020; Ekbatani et al. 2017; Li et al. 2020; Kortylewski et al. 2019, 2018; Weyrauch et al. 2004; Delussu et al. 2022b), to augment existing training sets of real images (Li et al. 2020; Yaghoubi et al. 2021; Shang et al. 2018; Tripathi et al. 2019; Dwibedi et al. 2017; Kortylewski et al. 2019) or for fine-tuning a model previously trained on real data (Mokhayeri et al. 2019). We subdivide existing techniques into image *pasting* and image *fusion*.

Image pasting consists of adding new items to a real background image, such as objects or pedestrians, which can be either taken from other real images or synthetically generated. Image pasting has been used to force object detection models to learn object appearance and to improve their robustness to image artefacts by adding to a background image instances of specific object categories (e.g., cereal box and bottle) taken from other real images (Dwibedi et al. 2017). Global image consistency was purposely disregarded since it was found to be less relevant in this task than patch-level realism (i.e., it suffices that the bounding box of the added object patch is realistic to human eyes). A similar approach has been used by Li et al. (2020) for face recognition: several kinds of occlusions (e.g., face masks, hands, hats, sunglasses) are pasted onto real face images to improve the recognition capability on occluded faces. Similarly, for pedestrian detection, one or more real images of pedestrians have been pasted onto a background scene image to simulate overlapping among people and to improve robustness to occlusions (Shang et al. 2018; Tripathi et al. 2019). Image pasting has also been used to generate scene-specific training sets for crowd counting, including a higher variability in features such as crowd size (Ekbatani et al. 2017; Delussu et al. 2020). To this aim, pedestrian images were pasted onto the background of the target scene in random positions; such images were taken from the same data set (Ekbatani et al. 2017) or from a gallery of pedestrian images collected from the Web (Delussu et al. 2022b). In the latter case, user interaction was required to obtain the perspective map of the target scene, which is necessary to properly rescale the pasted pedestrian images (Delussu et al. 2020).

Image fusion This technique consists of generating a synthetic image by suitably combining or fusing two or more *real* images that exhibit specific features (e.g., a pedestrian or a face image). For Re-Id, two different fusion techniques have been proposed. One consists of linearly interpolating two randomly selected training images to obtain a larger training set and improve the generalisation capability (Li et al. 2020). The other technique aims at forcing the model to focus on important features, e.g., the image foreground (Yaghoubi et al. 2021); to this aim, two images are selected according to some constraints in terms of aspect ratio, pose and viewpoint, and are then fused to obtain a new pedestrian image with, e.g., mixed lower- or upper-body appearance with respect to the original images, or the full-body of one pedestrian over the background of the other one.

Image fusion techniques have also been used in face recognition. One goal was to generate synthetic images representative of different pose and illumination conditions (Mokhayeri et al. 2019; Weyrauch et al. 2004). To this aim, pose and illumination are first extracted from a real face image; second, a 3D morphable model is used to reconstruct the same face image according to a target pose and a target illumination. Similarly, 3D morphable models are used to generate synthetic faces according to some user-defined parameters (e.g., shape, texture, light, camera, head pose), whereas the background is selected from a database of real images (Kortylewski et al. 2019, 2018).

5 Synthetic Data Sets

As mentioned in Sects. 3 and 4, several data sets of synthetic images have been developed so far; in the following, we describe the data sets proposed for each of the considered VS tasks, highlighting the publicly available ones.

5.1 Crowd Counting

Most real data sets for crowd counting, such as ShanghaiTech (Zhang et al. 2016b),[4] WorldExpo'10 (Zhang et al. 2016),[5] UCF-QNRF (Idrees et al. 2018a),[6] JHU (Sindagi et al. 2022),[7] NWPU (Wang et al. 2021),[8] are multi-scene, i.e., they are made up of images taken from different scenes, with

[4] https://www.kaggle.com/datasets/tthien/shanghaitech.

[5] http://www.ee.cuhk.edu.hk/~xgwang/expo.html.

[6] https://www.crcv.ucf.edu/data/ucf-qnrf/.

[7] http://www.crowd-counting.com/.

[8] https://gjy3035.github.io/NWPU-Crowd-Sample-Code/.

the aim of promoting the generalisation capability of crowd counting models to different target scenes (see Sect. 3.1). Some of them (e.g., WorldExpo'10) contain a set of video frames from each scene. Other data sets (e.g., UCF-QNRF) are made up of images collected from the Web, each one from a different scene, with very different characteristics in terms of background, perspective, lighting and weather conditions, etc.

Although several authors have generated synthetic images for their experiments, in particular, to pre-train the proposed crowd counting models, to our knowledge, only four synthetic data sets have been generated so far for this task: GCC (GTAV Crowd Counting) (Wang et al. 2019), CVCS (Cross-View Cross-Scene) (Zhang et al. 2021), CrowdX (Hou et al. 2022) and CrowdXV (Hou et al. 2023). All such data sets have been generated following the same multi-scene approach as real ones, but they also present richer annotations that allow a subdivision into single-view subsets. Figure 4 shows some sample images. We point out that CrowdX and CrowdV were not available online at the time of writing this paper.

GCC (GTAV Crowd Counting) was generated using GTAV. It consists of 15,212 images containing a total of 7,625,843 pedestrians, collected from 100 scenes and 4 different views per scene. Each image has a resolution of 1080 × 1920 pixels and contains, on average, 501 people. Moreover, different lighting conditions have been considered according to different daytime hours (e.g., 9–12) and weather conditions (e.g., extra sunny, rain). Pedestrians have been rendered using 256 different human models with 6 different clothing appearances. GCC has been built in three steps: scene selection, setting and synthesis. First, different locations were chosen (e.g., mall, stadium and store). For each location, the following parameters were chosen: camera view (e.g., height and rotation), region of interest (ROI), i.e., the area of the camera view where pedestrians can appear, scene capacity (maximum number of pedestrians), weather conditions, and time. For each location and view, several images were then set up by choosing the number of pedestrians and their position in the ROI. Finally, pedestrians were rendered on the background scene, and their head point locations were automatically computed.

CVCS was also generated using GTAV. It contains 280,000 images with a resolution of 1920 × 1080 pixels, obtained from 31 locations, with 60 to 120 different views for each location. Each image contains 90 to 180 pedestrians. This data set was generated by selecting two sets of parameters. First, scene-related features were chosen, i.e., location, ROI, weather conditions, 3D human models, etc. Then, the views of each location and the related camera parameters were chosen; in particular, a top view of each scene was included.

CrowdX contains 24,000 crowd images with a resolution of 1024 × 768 pixels. Given its rendering capabilities, Unity3D engine was used to simulate the scenes, cameras, lights, etc, while the pedestrian images come from PersonX (see Sect. 5.5), a data set originally proposed for Re-Id that contains 1,266 3D pedestrian models with diverse height, weight, skin colour, hairstyle and clothing. Dynamic properties (such as standing, walking, and running) were also edited to make the generated images highly realistic. Three urban scenes containing buildings, urban roads, traffic lights, etc and five solid colour backgrounds were used to build CrowdX. Finally, for each scene, to favour generalisation, the camera pitch angle was varied by the following values 30°, 50°, 70°, and 90° and the number of pedestrians was randomly sampled from 1 to 1000 with steps of 100. This data set provides richer annotation information than other existing data sets. Indeed, the annotation parameters for each scene include the scene ID, type and camera position, rotation and resolution. Furthermore, annotations are also provided for each pedestrian including the pedestrian's ID, 3D position, 2D camera plane positions, height, and standing direction.

CrowdXV extends CrowdX into a synthetic video data set for the evaluation of video crowd counting algorithms. Like CrowdX, it is built using Unity3D and PersonX. It contains 10,000 video clips of five frames each, with a frame rate of 30 fps and a resolution of 1024 × 768. The average number of people per frame is 250. The pedestrians are randomly instantiated in the scene and move with direction and velocity that follow a Gaussian and a uniform distribution, respectively.

Table 1 reports a summary of the main features of synthetic data sets, as well as the above-mentioned benchmark real data sets, for comparison, in terms of their number of images, pedestrian count and number of scenes. It can be seen that GCC and CVCS contain a much larger (one to three orders of magnitude) total number of images than real data sets. They also contain images of the same locations from different views, contrary to real data sets. On the other hand, three real data sets (UCF-QNRF, JHU and NWPU) contain images with a much larger crowd size; two of them (JHU and NWPU) also contain a higher number of scenes. We point out that, although synthetic data sets for this task are much larger than existing real data sets (as one would expect from synthetic ones), they present some limitations. In particular, GCC is slightly unbalanced in terms of crowd size: the number of images showing a dense crowd is lower than those showing a medium or sparse crowd. In contrast, CVCS has a relatively small range of pedestrian counts and lacks crowded scenes (the largest pedestrian count is 180), whereas CrowdXV contains just five frames per video clip, even if it was proposed for the evaluation of video crowd counting algorithms.

Table 1 Statistics of synthetic data sets for crowd counting (top four rows): average image size, number of images, pedestrian count and number of scenes

	Data set	Avg. img. size	# Images	Pedestrian count				Scenes		
				Total	Min	Avg	Max	Total	Locations	Views
Synth	GCC (Wang et al. 2019)	1080 × 1920	15,212	7,625,843	0	501	3995	400	100	4
	CVCS (Zhang et al. 2021)	1080 × 1920	280,000	–	90	–	180	2800	31	Avg 90
	CrowdX (Hou et al. 2022)	1024 × 768	24,000	–	1	500	1000	36	8	4
	CrowdXV (Hou et al. 2023)	1024 × 768	50,000	–	–	250	–	–	–	–
Real	SHT A (Zhang et al. 2016a)	589 × 868	482	241,677	33	501	3139	482	482	1
	SHT B (Zhang et al. 2016a)	768 × 1024	716	88,488	9	123	578	716	716	1
	WorldExpo’10 (Zhang et al. 2016)	576 × 720	3980	199,923	1	50	253	108	1	108
	UCF-QNRF (Idrees et al. 2018b)	2013 × 2902	1525	1,251,642	49	815	12,865	1525	1525	1
	JHU (Sindagi et al. 2022)	910 × 1430	4372	1,510,000	0	346	25,000	4372	4372	1
	NWPU (Wang et al. 2021)	2191 × 3209	5109	2,133,375	0	418	20,033	5109	5109	1

For comparison, statistics of five benchmark real data sets are also reported

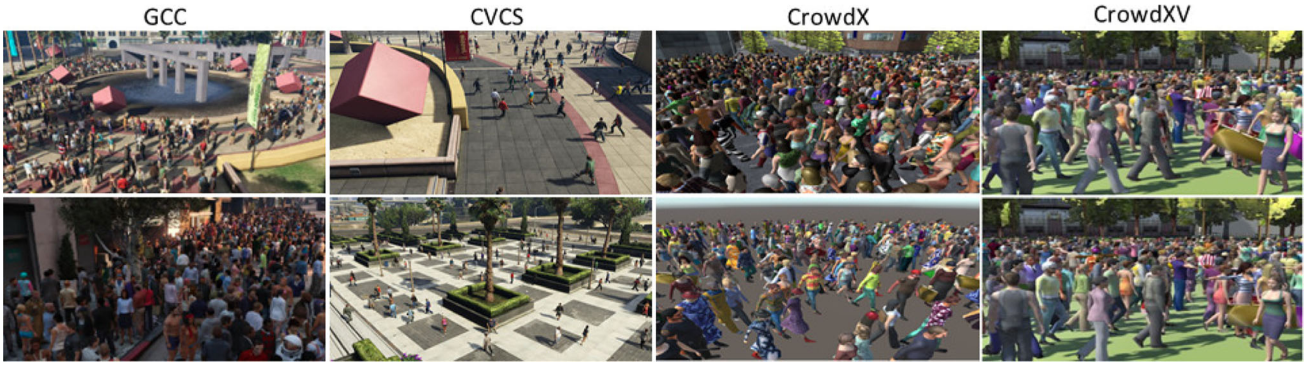

Fig. 4 Examples of images from synthetic data sets for crowd counting. From left to right: GCC (Wang et al. 2019) (source: https://gjy3035.github.io/GCC-CL/), CVCS (Zhang et al. 2021) (source: https://github.com/zqyq/Cross-view-cross-scene-multi-view-counting-CVPR2021), CrowdX (Hou et al. 2022) (not available online) and CrowdXV (Hou et al. 2023) (not available online)

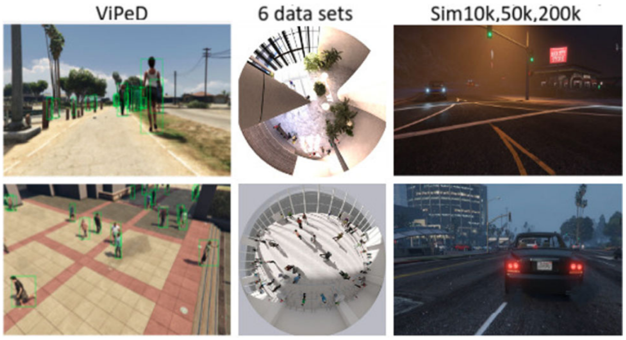

Fig. 5 Examples of images from synthetic data sets for detection tasks. From left to right: Virtual Pedestrian Data set (ViPeD) (Ciampi et al. 2020) (source: https://ciampluca.github.io/viped/), 6 data sets (Aranjuelo et al. 2021) (source: https://datasets.vicomtech.org/v4-osd/OSD_download.zip) and Sim10k, 50k, 200k (Johnson-Roberson et al. 2017) (source: https://fcav.engin.umich.edu/projects/driving-in-the-matrix)

5.2 Object and Pedestrian Detection

Two synthetic data sets have been generated for pedestrian detection: Virtual Pedestrian Dataset (ViPeD) (Ciampi et al. 2020) and “6 data sets” (Aranjuelo et al. 2021); and one for object detection, in three versions of different sizes: Sim10k, Sim50k and Sim200k (Johnson-Roberson et al. 2017). They all have been generated using computer graphics software. Figure 5 shows some examples of images from these data sets.

ViPeD (Ciampi et al. 2020) was built using GTAV, and consists of about 500,000 images acquired from 512 urban videos coming from the JTA synthetic data set previously developed for tracking tasks (Fabbri et al. 2018) (see Sect. 5.3). JTA also included skeleton information about 14 body parts of pedestrians appearing in its images. To build ViPeD, the coordinates of each pedestrian bounding box were automatically extracted, limited to pedestrians within 40 ms from the camera.

The AutoDesk rendering software has been used to generate *6 data sets* (Aranjuelo et al. 2021), focused on pedestrian detection in large indoor spaces. It contains 28,000 images subdivided into six sets, characterised by increasing difficulty levels. Images were obtained from different photo-realistic scenes with different locations, backgrounds, distortions, lighting conditions, pedestrian appearance and locations, objects and camera positions.

The GTAV video game was used to generate three data sets for object detection of different sizes, named *Sim10k* (i.e., made up of 10,000 samples), *Sim50k* and *Sim200k* (Johnson-Roberson et al. 2017), including several weather conditions (e.g., sunny and foggy) and lighting conditions (e.g., night). We point out that the original goal of that work was to evaluate the data set bias in object detection and classification (Johnson-Roberson et al. 2017).

Statistics about the above synthetic data sets are reported in Table 2, together with a comparison with well-known and widely used real ones: Cityscapes (Cordts et al. 2016),[9] KITTI (Geiger et al. 2012),[10] MOT17Det (Milan et al. 2016),[11] MOT19Det (Dendorfer et al. 2019) and COCOPerson (Lin et al. 2014).[12] It can be seen that even for this task, synthetic data sets are much larger than real ones. This task also points out several issues of real data sets, that can be overcome by synthetic ones. One is related to manual annotation: in some data sets (e.g., Cityscapes), groups of objects of the same category (e.g., people or vehicles) can appear close to each other or partially overlapping, to the extent that the boundary of individual objects is not clearly visible: in

[9] https://www.cityscapes-dataset.com/.

[10] https://www.cvlibs.net/datasets/kitti/.

[11] https://motchallenge.net/data/MOT17Det/.

[12] https://cocodataset.org/#home.

Table 2 Statistics of synthetic data sets for object and pedestrian detection, and a comparison with the main real data sets available for these tasks

		Data set	# Images
Object	Synthetic	Sim10k, 50k, 200k (Johnson-Roberson et al. 2017)	10k–50k–200k
	Real	Cityscapes (Cordts et al. 2016)	2975
		KITTI (Geiger et al. 2012)	7481
Pedestrian	Synthetic	ViPeD (Ciampi et al. 2020)	500,000
		6 data sets (Aranjuelo et al. 2021)	28,000
	Real	MOT17Det (Milan et al. 2016)	5316
		MOT19Det (Dendorfer et al. 2019)	8931
		COCOPerson (Lin et al. 2014)	66,000

this case, they are given together a single annotation as a group (e.g., a group of people). Another issue is that some data sets (e.g., KITTI) exhibit a notable class imbalance in favour of a particular class (e.g., car) with respect to others (e.g., truck and van) or skewed distribution of the number of pedestrians per image, with a large number of images containing a single pedestrian, and fewer ones containing two or more pedestrians. A further issue is the time required to collect real images: for instance, one of the goals of Cityscapes was to include images collected in different seasons, which required several months.

5.3 Object and Pedestrian Tracking

Three synthetic data sets have been built for object tracking: VIsual PERception (VIPER) (Richter et al. 2017), VirtualKITTI2 (Cabon et al. 2020) and SHIFT (Sun et al. 2022); and four for pedestrian tracking: Virtual Person Tracking Benchmark #1 (VirtualPTB1) (Kerim et al. 2021a), Person Tracking of Synthetic sequences (PTAW217Synth) (Kerim et al. 2021b), Joint Track Auto (JTA) (Fabbri et al. 2018) and MOTSynth (Fabbri et al. 2021). They have been built mainly using computer graphics software. We point out that VirtualPTB1 and JTA were not available online at the time of writing this paper.

VIPER (Richter et al. 2017) was generated using GTAV and presents 254,064 frames with a resolution of 1920 × 1080 pixels, under different weather conditions and times of day (e.g., sunset and night). Besides the common annotations for tracking (see in Sect. 3.3), such data set also presents annotations for semantic and instance segmentation, optical flow and 3D bounding box for each object in the scene.

VirtualKITTI2 (Cabon et al. 2020) was generated using Unity and contains five scenes characterised by different illumination and weather conditions (e.g., clear and fog). Similarly to VIPER, it also contains annotations for different CV tasks: depth information, class segmentation (e.g., terrain, guardrail, van, car), instance segmentation, and optical flow.

SHIFT (Sun et al. 2022) was built using the CARLA simulator (Dosovitskiy et al. 2017) and the open-source web annotation tool Scalabel Zurich (n.d.). This data set consists of 2.5M frames of 4850 tracking sequences which were acquired in eight virtual cities. Moreover, different weather and illumination conditions are considered. This data set presents several annotations such as semantic and instance segmentation, 2D and 3D bounding boxes, depth maps, optical flow, key-points for human poses. Due to this information, SHIFT can also be employed in other tasks, e.g., object detection and semantic segmentation.

VirtualPTB1 (Kerim et al. 2021a) was generated using Nova and consists of 108 video sequences and more than 13,000 photo-realistic frames generated under different weather conditions and times. Also this data set contains further annotations useful for various CV tasks: optical flow, surface normal, depth map, object identifiers, semantic segmentation and body part segmentation.

PTAW217Synth (Kerim et al. 2021b) was built as well using Nova, but it presents a higher number of frames with respect to VirtualPTB1 (217 videos and 108,547 frames) and higher variability in terms of weather conditions (36,182 frames on average were generated for foggy, snowy and rainy), time of day, camera type, number and density of pedestrians.

JTA (Fabbri et al. 2018) was built using GTAV and contains more than 460,000 frames of urban scenes, taken from 512 videos, each containing 21 pedestrians on average. It also contains 3D coordinates annotations and the distinction between occluded and visible body parts.

MOTSynth (Fabbri et al. 2021) was generated using GTA5 and contains 1382k frames of urban scenes acquired using different weather conditions (e.g., overcast). This data set contains bounding boxes, segmentation and depth masks. Due to its annotations, it can also be used for other tasks such as pedestrian detection.

Table 3 reports the main statistics of the above synthetic data sets (see Fig. 6) for pedestrian and object tracking. Some critical aspects are shared with detection data sets, e.g., distribution and balancing of categories. In addition to these

Table 3 Statistics of synthetic data sets, and of some benchmark real data sets, for object and pedestrian tracking tasks

		Data set	#Frames
Object	Synthetic	VIPER (Richter et al. 2017)	254,064
		VirtualKITTI2 (Cabon et al. 2020)	–
		SHIFT (Sun et al. 2022)	2,500,000
	Real	Cityscapes (Cordts et al. 2016)	2975
		KITTI (Geiger et al. 2012)	7481
Pedestrian	Synthetic	VirtualPTB1 (Kerim et al. 2021a)	> 13,000
		PTAW217Synth (Kerim et al. 2021b)	108,547
		JTA (Fabbri et al. 2018)	460,000
		MOTSynth (Fabbri et al. 2021)	1,382,000
	Real	MOT17Det (Milan et al. 2016)	5316
		MOT19Det (Dendorfer et al. 2019)	8931

factors, we point out two further problems of real data sets. One is the presence of a small number of videos per scene, i.e., only one scene for training and one for testing [e.g., MOT17Det (Milan et al. 2016)]. Another problem is that each video might contain a different number of pedestrians per frame or can be acquired using different cameras (e.g., moving cameras). The use of synthetic images can alleviate the above problems since it allows monitoring any aspect, including the frame rate, if necessary.

5.4 Human and Crowd Behaviour Analysis

To the best of our knowledge, seven synthetic data sets have been built so far for human or crowd behaviour analysis: AGORASET (Courty et al. 2014), CrowdFlow (Schroder et al. 2019), Labelled Crowd Videos (LCrowdV) (Cheung et al. 2019), Procedural Human Action Videos (PHAV) (de Souza et al. 2017), Abnormal Crowd Data set (ACD) (Lazaridis et al. 2018), SyntHetic Abnormality DatasEt (SHADE) (Lin et al. 2021) and GTA5Event (Montulet and Briassouli 2020). They have been generated using several computer graphics software, i.e., Unreal, Unity, GTAV and AutoDesk.

It is worth highlighting that a few data sets present individual behaviour annotations (i.e., LCrowd and PHAV), whereas the others contain crowd behaviour information. In the following, we first describe the data sets which contain individual behaviour information. *LCrowdV* (Cheung et al. 2019) was built using Unreal Engine. It consists of more than 1,000,000 videos and 20,000,000 frames generated in indoor (e.g., mall) and outdoor (e.g., park) environments, under different lighting and weather conditions (e.g., sunny), and various crowd density levels. The following crowd behaviours, obtained by combining individual behaviours (extracted using a personality-based model), have been simulated: aggressive, assertive, shy, active, tense and impulsive. The following annotations are present globally and frame per frame: pedestrian trajectories and bounding boxes, head point annotations, pedestrian count, and flow estimation. *PHAV* (de Souza et al. 2017) was generated using Unity and contains 55 h of videos with about 6,000,000 frames and more than 1000 videos for each behaviour of interest, which include car hit, escape, gunshot, walking and climbing stairs. Each scene contains two or more people. This data set presents the following annotations per frame: depth map, semantic and instance segmentation, 2D and 3D bounding boxes, pose, and muscle information (e.g., muscular strength). *AGORASET* (Courty et al. 2014) is the first synthetic data set proposed for this task. It has been generated using AutoDesk, and consists of seven scenes with medium- and high-density crowds, taking place in very simple virtual environments defined by stylised walls and floors rendered with uniform colours; various pedestrian appearances (e.g., different clothes) are considered. Each scene simulates normal crowd behaviours, such as people flow, cross flow and rotation, and abnormal behaviours, such as evacuation and dispersion. Each behaviour is shown under different velocities, views, occlusions and lighting conditions. The data set contains individual trajectories, density, velocity and per-frame segmentation annotations. *CrowdFlow* (Schroder et al. 2019) was built using Unreal Engine at resolution 1280 × 720 and 25 fps. It contains 10 video sequences with lengths ranging between 300 and 450 frames for a total of 3200 frames. The video sequences are related to 5 scenes that were rendered twice: with a static point of view and a dynamic camera (simulating drone/UAV cameras). Each sequence has the following ground-truth data: optical flow fields, person trajectories (up to 1451) and dense pixel trajectories. Each sequence contains between 371 and 1451 independently moving individuals and covers different kinds of crowd movement: structured behaviour with either a single crowd or two crowds moving in different directions as well as fully unstructured movements of the individuals. The other data sets were all generated using GTAV. *ACD* (Lazaridis et al. 2018) contains 14 videos (globally annotated) of dense crowds, showing seven normal

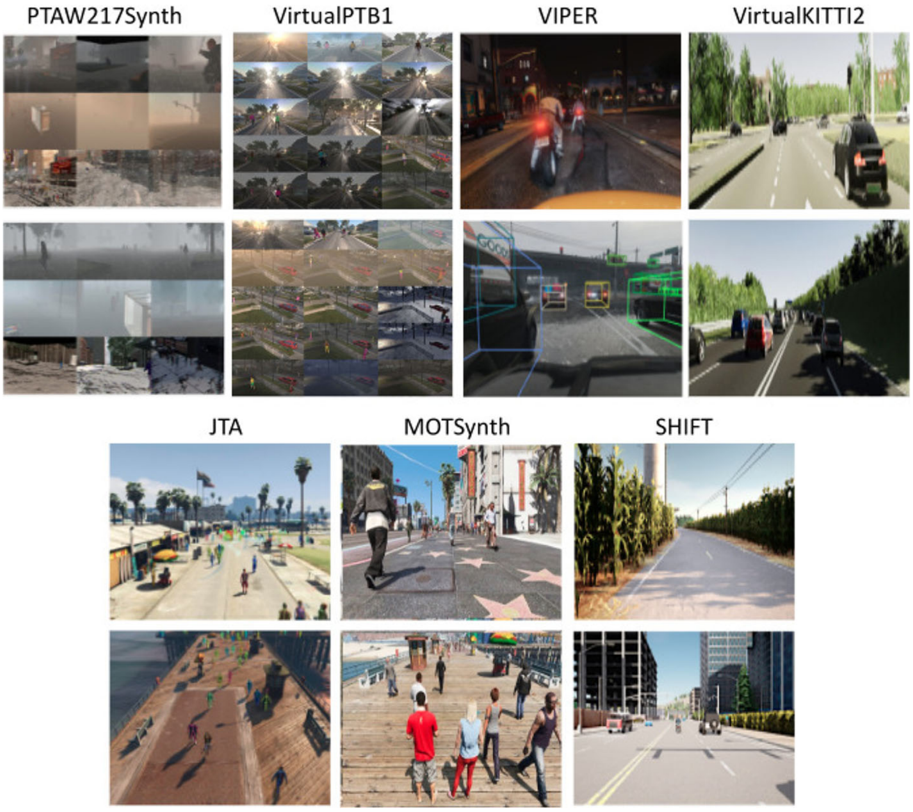

Fig. 6 Examples of frames from synthetic data sets for tracking tasks. From left to right: PTAW217Synth (Kerim et al. 2021b) (source: https://graphics.cs.hacettepe.edu.tr/NOVA-Adverse/), VirtualPTB1 (Kerim et al. 2021a) (source: https://github.com/A-Kerim/NOVA), VIPER (Richter et al. 2017) (source: https://github.com/LucasVandroux/pfb2kitti), VirtualKITTI2 (Cabon et al. 2020) (source: https://europe.naverlabs.com/research/computer-vision/proxy-virtual-worlds-vkitti-2/), JTA (Fabbri et al. 2018) (source: https://aimagelab.ing.unimore.it/imagelab/page.asp?IdPage=25), MOTSynth (Fabbri et al. 2021) (source: https://aimagelab.ing.unimore.it/imagelab/page.asp?IdPage=42), SHIFT (Sun et al. 2022) (source: https://www.vis.xyz/shift/)

behaviours and two abnormal ones, fight and panic. *SHADE* (Lin et al. 2021) contains 2149 videos, 879,932 frames and 4 views for each scene, with sparse and medium crowds. Several normal behaviours are shown, including social activities, walking, and the following anomalous ones: arrest, chase, fight, scatter, knock-down, run and shooting. For each activity, 200 videos are present under different weather conditions, times of day, locations, etc. *GTA5Event* (Montulet and Briassouli 2020) consists of about 24,000 frames and 54 scenes (e.g., beach) involving medium-size crowds. Such scenes are characterised by several camera settings (e.g., different heights and angles), different weather conditions and time of day. Different behaviours are shown, such as walking, standing, social activities, etc.; every video corresponds to a single behaviour, and the corresponding label is given to the whole video. This data set also contains the following annotations: pedestrian bounding boxes, identities and trajectories, depth map, and scene information (e.g., location and camera position) and information about the frame where, e.g., an abnormal behaviour occurs. Table 4 shows the main statistics of the above synthetic data sets, together with a comparison with the benchmark real data sets UCSD Anomaly Detection (Chan et al. 2008),[13] CUHK Avenue (Lu et al. 2013),[14] UMN (Mehran et al. 2009)[15] and UCF Crime (Sultani et al. 2018).[16]

The above data sets differ in terms of the number of videos, frames, events, image resolution and annotations. Moreover, in most cases, the anomalous behaviours occurring in real data sets differ from those occurring in synthetic data sets. It is worth noting that just one synthetic data set (i.e., PHAV) presents a higher number of events with respect to real data sets (except for UCSD, whose abnormal behaviours are related to non-pedestrian entities only), and that the lowest number of anomalous behaviours overall appears in a synthetic data set (i.e., ACD). Moreover, in our opinion, the degree of realism exhibited by most of the above synthetic data sets is somewhat limited (see Fig. 7).

The type of annotations represents another difference between real and synthetic data sets: in most real data sets, videos are globally annotated with a specific label (e.g., normal, abnormal), whereas in synthetic data sets, few of them present this type of annotation (i.e., ACD and GTA5Event). The other synthetic data sets contain frame annotations; therefore, it is possible to extract the exact time span in which the anomaly occurs.

[13] http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm.

[14] http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html.

[15] https://www.crcv.ucf.edu/projects/Abnormal_Crowd/.

[16] https://www.crcv.ucf.edu/projects/real-world/.

Table 4 Statistics of synthetic data sets for human and crowd behaviour analysis, and a comparison with benchmark real data sets for the same task

	Data set	# Videos	# Frames	# Events	Abnormal behaviours
Synthetic	SHADE (Lin et al. 2021)	2149	879,932	7	Arrest, chase, fight, scatter, knockdown, run, shoot
	ACD (Lazaridis et al. 2018)	14	–	2	Fight, panic
	AGORASET (Courty et al. 2014)	–	250	5	Obstacles, evacuation, dispersion, rotation, crossing
	CrowdFlow (Schroder et al. 2019)	10	3200	3	Flow, crossing flow, independent flow
	LCrowdV (Cheung et al. 2019)	1 M	20 M	6	Aggressive, shy, assertive, active, tense, impulsive
	PHAV (de Souza et al. 2017)	39,982	6 M	34	Car hit, bump into each other, escape, hop, etc
	GTA5Event (Montulet and Briassouli 2020)	54	24,000	3	Fight, flee random, flee same
Real	UCSD anomaly detection (Chan et al. 2008)	100	2000	52	Non-pedestrian entities
	CUHK avenue (Lu et al. 2013)	37	35,240	14	Running, throwing objects, loitering, abnormal object, etc
	UMN (Mehran et al. 2009)	11	7725	3	Speed, direction
	UCF crime (Sultani et al. 2018)	1900	13 M	13	Fighting, road accident, burglary, robbery, etc

In particular, #Event represents the number of behaviours contained in the corresponding data set

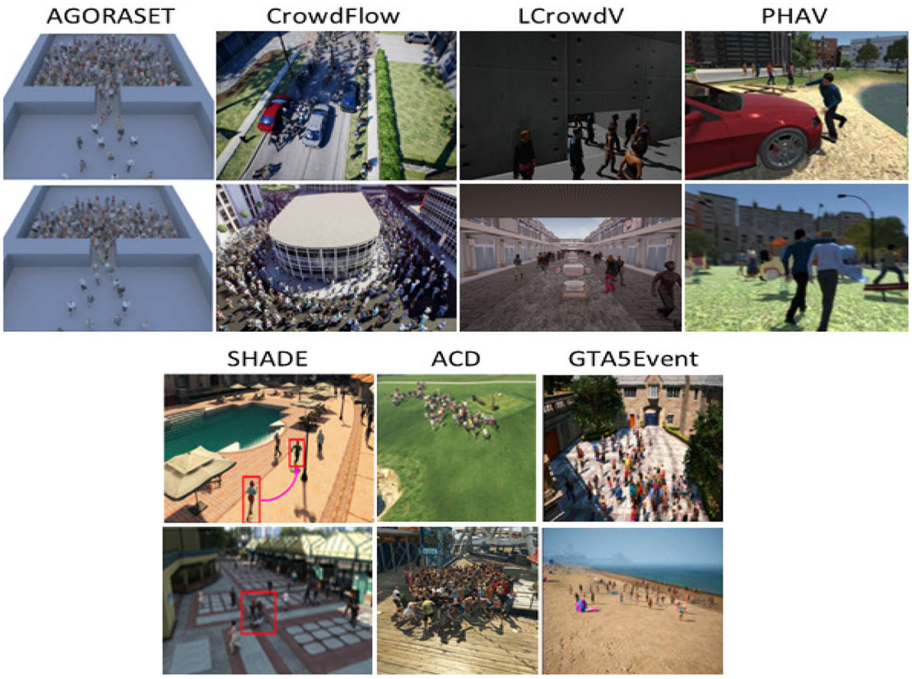

Fig. 7 Examples of frames from synthetic data sets for human and crowd behaviour analysis. From left to right: AGORASET (Courty et al. 2014) (source http://www.sites.univ-rennes2.fr/costel/corpetti/agoraset/Site/AGORASET.html), CrowFlow (Schroder et al. 2019) (source https://github.com/tsenst/CrowdFlow), LCrowdV (Cheung et al. 2019) (source http://gamma.cs.unc.edu/LCrowdV/#dataset), PHAV (de Souza et al. 2017) (source http://adas.cvc.uab.es/phav/), SHADE (Lin et al. 2021) (not available online), ACD (Lazaridis et al. 2018) (not available online), GTA5Event (Montulet and Briassouli 2020) (source https://github.com/RicoMontulet/GTA5Event)

Finally, we point out that, among synthetic data sets, only LCrowdV is available online.

5.5 Person Re-Identification

Re-Id is the task for which the highest number of synthetic data sets has been generated, due to the considerable interest it has received by the scientific community in recent years and probably also to privacy restrictions that make the acquisition of real data sets particularly difficult for this task (Uner et al. 2021). The existing data sets are: Synthetic18k (Uner et al. 2021), PersonX (Sun and Zheng 2019), RandPerson (Wang et al. 2020), Virtually Changing-Clothes (VC-Clothes) (Wan et al. 2020), GTA Person Re-Id (GPR) (Xiang et al. 2020), GPR+ (Xiang et al. 2021), FineGPR (Xiang et al. 2023), Synthetic person Re-Id (SyRI) (Bak et al. 2018), SOMAset (Barbosa et al. 2018), UnrealPerson (Zhang et al. 2021), WePerson (Li et al. 2021), ClonedPerson (Wang et al. 2022). They have been mostly generated using computer graphics engines (Synthetic18k, RandPerson, PersonX, VC-Clothes, GPR, GPR+, FineGPR, WePerson), mainly Unity and GTAV. In particular, a single engine has been used for Synthetic18k, PersonX, SOMAset, GPR, GPR+, FineGPR and WePerson, whereas two engines have been used for RandPerson, SyRI and ClonedPerson. In the former case, either Unity or GTAV were used to create both the virtual environments and the human models.

Synthetic18K, PersonX, RandPerson and ClonedPerson have been generated using Unity. *Synthetic18K* (Uner et al. 2021) contains four synthetic environments (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 individual pedestrians. *PersonX* (Sun and Zheng 2019) contains three different environments, as well as the same pedestrian images with a uniform background; this data set focuses on viewpoint changes, i.e., the cameras move to capture pedestrians from different viewpoints. To increase diversity in pedestrian appearance, the 1266 identities of this data set (547 females and 719 males) present different ages, body forms, skin colour, etc. In total, 273,456 bounding boxes of individual pedestrians are present. *RandPerson* (Wang et al. 2020) has been built using Unity, jointly with MakeHuman, to generate the human models. It contains 11 environments (8 outdoor and 3 indoor), 8000 identities and a total of 1,801,816 bounding boxes of individual pedestrians, acquired from 19 cameras (more than one camera view was considered for some environments). *ClonedPerson* (Wang et al. 2022) consists of 887,766 bounding boxes of 5621 identities, acquired by 24 cameras. It was generated by importing into Unity human models created using MakeHuman.

VC-Clothes, GPR and WePerson have been generated using GTAV, instead. *VC-Clothes* (Wan et al. 2020) 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 for this task, VC-Clothes focuses on *clothing-independent* Re-Id, and therefore, images of each identity differ in clothing appearance and attributes. *GPR* (Xiang et al. 2020) was generated with GTAV. This data set is composed of 443,352 images of 754 identities collected using 12 cameras, under 12 weather conditions (e.g., cloudy and foggy), 8 different illuminations (e.g., afternoon and midnight) and 26 typical locations, such as beach, street, school and mall. Two further versions of GPR were generated: *GPR+* (Xiang et al. 2021) and *FineGPR* (Xiang et al. 2023). The former contains more identities and images than GPR; the latter contains more fine-grained details, mostly attributes of pedestrian models such as upper- and lower-body clothing colours, hats and bags. *WePerson* (Li et al. 2021) has been built using GTAV. It is composed of 4 million bounding boxes of 1500 identities acquired by 560 cameras. The images have been generated using seven weather conditions (e.g., cloudy, snow), seven illumination conditions (e.g., afternoon, night), 36 viewpoints and 14 scenes (e.g., street, subway), among which 10 outdoor and 4 indoor. Moreover, several occlusions among different pedestrians or objects have been simulated. *SyRI* (Bak et al. 2018) was generated as well using two computer graphics softwar, Unreal and Adobe. It contains 100 identities and 56,000 images acquired under more than 100 different illumination conditions. The computer graphics software MakeHuman has been used to generate SOMAset and UnrealPerson, together with, respectively, Blender and Unreal. *SOMAset* (Barbosa et al. 2018) focuses on clothing-independent or long-term Re-Id, analogously to VC-clothes. To this aim, it contains 50 identities, each 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* (Zhang et al. 2021) contains four scenes (three urban outdoor and one indoor) built using Unreal under different illumination conditions and 3000 human models generated using MakeHuman, presenting more than 200 types of clothes and in some cases different accessories (e.g., masks, glasses and hats). In total, it contains 120,000 bounding boxes acquired from 34 cameras.

Examples of images from the above data sets are shown in Fig. 8. Table 5 summarises their main statistics and compares them with three widely used benchmark data sets of real images, Market-1501 (Zheng et al. 2015),[17] CUHK03 (Li et al. 2017)[18] and MSMT17 (Wei et al. 2018).[19] It is worth noting that one of the most used data set, DukeMTMC (Ergys et al. 2016), has been retracted due to privacy constraints; for this reason, it is not included in Table 5.

It is also worth noting that some synthetic data sets (SOMAset, PersonX and VC-Clothes) contain a lower number of images than the largest real data set (MSMT17) and that only Synthetic18k contains a larger number of identities than the four real data sets reported in Table 5.

We also point out that, among the existing synthetic data set, only RandPerson and UnrealPerson reproduce a real environment with interactions between people or static objects. Indeed, they were generated under a set-up similar to real VS systems, with multiple people moving simultaneously and partial occlusions by static objects or by other people.

An interesting, related issue is the degree of photo-realism of the different data sets, which can be observed from the examples in Fig. 8. Focusing on human model appearance, in our opinion, SyRI presents better visual details. In contrast, PersonX looks less photo-realistic than the other data sets in terms of body shape and textures. In terms of the background scene, RandPerson, UnrealPerson and GPR exhibit the highest degree of photo-realism. Nevertheless, all data sets try to attain a certain degree of realism by focusing on one or more visual aspects. For instance, SyRI focused on illumination conditions, whereas PersonX paid attention to pose, illumination, background and viewpoint.

However, understanding what degree of realism is beneficial to Re-Id models is still an open issue. This can be observed by the fact that existing synthetic data sets focus on *different* visual factors, as pointed out above. In particular, only some of them present a high degree of realism in *all* image components (human models, background, etc.). Accordingly, an interesting direction for future work is to investigate the relationship between the degree of realism of the different image components and the performance of Re-Id models. We carried out a preliminary analysis of this issue in a previous work (Delussu et al. 2022a).

We finally point out that some synthetic data sets turned out to be unavailable at the time of writing this manuscript, whereas, for some other data sets, we observed some discrepancies with the information in the respective papers. In particular: (i) GPR, and GPR+ have not yet been released; (ii) Synthetic18k does not include camera information, which is necessary for training Re-Id models; (iii) in several cases, the downloaded data sets present different statistics in terms of the number of images, identities and cameras with respect to the ones reported in the respective papers; for instance, FineGPR should contain more than 2 million images and 36 cameras, whereas the downloaded version contains only about 4 thousand images and 4 cameras.

5.6 Face Recognition

Currently, numerous synthetic face data sets are available for tasks such as feeling, emotion and pain recognition, age estimation, and more recently also for deep fake detection (Mirsky and Lee 2021). However, although synthetic face images were used in a large number of works specifically for face recognition, to our knowledge, only five synthetic data sets have been built for this task: Kortilesky (Kortylewski

[17] https://www.kaggle.com/datasets/pengcw1/market-1501.

[18] http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html.

[19] https://www.pkuvmc.com/dataset.html.

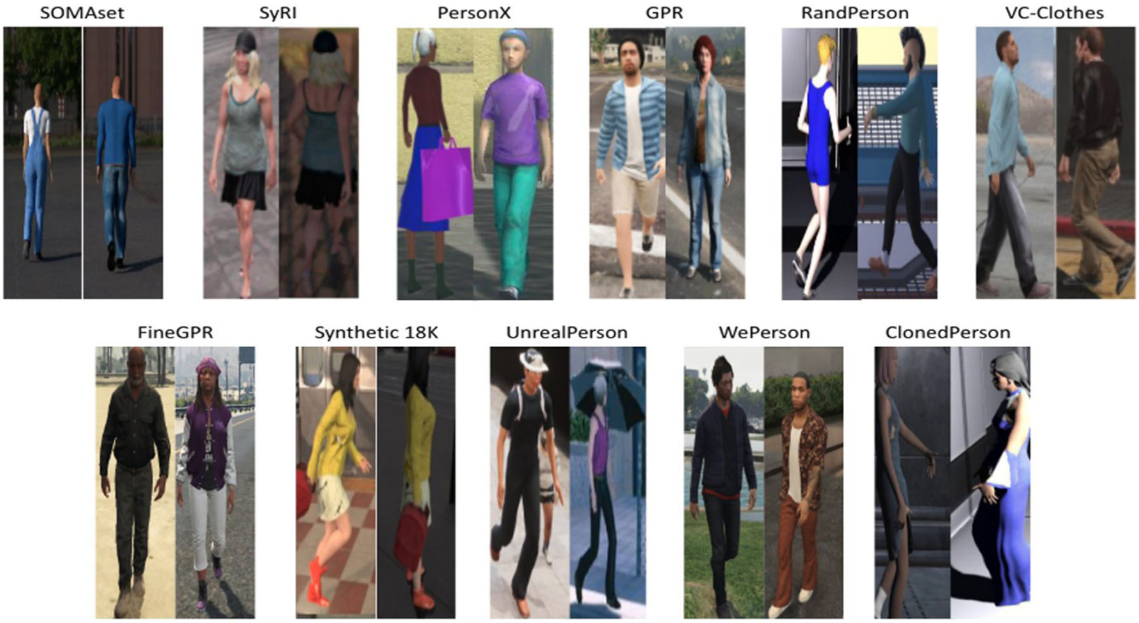

Fig. 8 Examples of images from synthetic data sets for Re-Id. From left to right: SOMAset (Barbosa et al. 2018) (source: https://www.kaggle.com/vicolab/somaset), SyRI (Bak et al. 2018) (source: https://github.com/swbak/SyRI), PersonX (Sun and Zheng 2019) (source: https://github.com/sxzrt/Instructions-of-the-PersonX-dataset), GPR (Xiang et al. 2020) (source: https://github.com/JeremyXSC/GPR/), RandPerson (Wang et al. 2020) (source: https://github.com/VideoObjectSearch/RandPerson), VC-Clothes (Wan et al. 2020) (source: https://wanfb.github.io/dataset.html), FineGPR (Xiang et al. 2023) (source: https://github.com/JeremyXSC/FineGPR), Synthetic18k (Uner et al. 2021) (source: https://hucvl.github.io/synthetic18k/), UnrealPerson (Zhang et al. 2021) (source: https://github.com/FlyHighest/UnrealPerson), WePerson (Li et al. 2021) (source: https://github.com/lihe404/WePerson), ClonedPerson (Wang et al. 2022) (source: https://github.com/Yanan-Wang-cs/ClonedPerson)

Table 5 Statistics of synthetic Re-ID data sets, and of the main benchmark data sets of real images

Data set		#Identities	# Images	#Cam	View
Synthetic	SOMAset (Barbosa et al. 2018)	50	100,000	250	N
	SyRI (Bak et al. 2018)	100	1,680,000	–	N
	PersonX (Sun and Zheng 2019)	1266	273,456	6	Y
	PersonX$_{123,456}$	1266	136,728	3	Y
	PersonX$_{12,13}$	1266	91,152	2	Y
	PersonX$_{45,46}$	1266	91,152	2	Y
	VC-Clothes (Wan et al. 2020)	512	19,060	4	N
	GPR (Xiang et al. 2020)	754	443,352	12	Y
	GPR+ (Xiang et al. 2021)	808	475,104	12	Y
	FineGPR (Xiang et al. 2023)	1150	2,028,600	36	Y
	WePerson (Li et al. 2021)	1500	4,000,000	560	N
	UnrealPerson (Zhang et al. 2021)	3000	120,000	34	Y
	ClonedPerson (Wang et al. 2022)	5621	887,766	24	N
	RandPerson (Wang et al. 2020)	8000	132,145	19	N
	Synthetic18k (Uner et al. 2021)	18,306	1,408,600	4	–
Real	Market-1501 (Zheng et al. 2015)	1501	32,668	6	N
	CUHK03 (Li et al. 2017)	1467	14,096	2	N
	MSMT (Wei et al. 2018)	4101	124,068	15	–

“View” denotes whether the data set has viewpoint labels et al. 2019), FaceSynthetics (Wood et al. 2021), Syn10K50 (Qiu et al. 2021), USynthFace (Boutros et al. 2023a) and SFace (Boutros et al. 2022). Most of them have been generated using GANs, and all of them present a significant variety in terms of ethnicity, age, accessories (e.g., sunglasses and hats) and background. *Kortilesky* (Kortylewski et al. 2019) has been generated using a 3D Morphable model in order to control shape, texture as well as pose, illumination and facial expression. The head pose is sampled according to a uniform pose distribution on the yaw, pitch and roll angles in the respective ranges: yaw $\in [-90°, 90°]$, pitch $\in [-30°, 30°]$ and roll $\in [-15°, 15°]$. The 3D models are finally rendered in 2D to create the data set, which consists of 1 million face images with 20 thousand different identities, and 100 images per identity. *FaceSynthetics* (Wood et al. 2021) has been generated using AutoDesk. It contains 100,000 images of 100,000 identities with a resolution of 512×512 pixels, and 70 standard landmark and semantic class annotations (e.g., background, skin, nose and neck), besides identity, which makes this data set useful to other related tasks such as face parsing and landmark localisation. *Syn10K50*, have been generated with DiscoFaceGAN (Qiu et al. 2021) by randomly sampling latent variables from the standard normal distribution for identity, expression, pose and illumination coefficients, respectively, which leads to the same person with different expressions, poses and illuminations in the same class. It presents 10K different identities with 50 samples per identity, for a total of 500K images. *USynthFace* (Boutros et al. 2023a) has also been generated using DiscoFaceGAN in combination with other GANs for data augmentation. This data set presents just one image per identity, leaving the users the task of generating further images if necessary, e.g., by data augmentation. The full data set contains a total of 400K images. *SFace* (Boutros et al. 2022) has been generated by translating a real public data set through StyleGAN2-ADA, which is a style-based GAN (Style-GAN2) with adaptive discriminator augmentation (ADA) to increase the diversity of the training data set. The real data set used both for training and translation is CASIA-WebFace (Yi et al. 2014). Nevertheless, the number of synthetically generated images is 634K, which is higher than that of CASIA-WebFace; such images are equally distributed across identities (60 images per identity).

Examples of images from the above data sets are shown in Fig. 9. Table 6 reports their main statistics taken from the respective papers or from the data set documentation, and a comparison with some well-known benchmark data sets of real images, LFW (Huang et al. 2008),[20] Extended Yale B (Georghiades et al. 2001),[21] SCface (Grgic et al. 2011),[22] CASIA-WebFace (Yi et al. 2014),[23] CFPW (Sengupta et al. 2016),[24] AgeDB (Moschoglou et al. 2017)[25] and FFHQ (Karras et al. 2021).[26] It is worth noting that some of the largest real data sets for face recognition [MegaFace (Kemelmacher-Shlizerman et al. 2016), MS-Celeb-1 M (Guo et al. 2016) and VGGFace2 (Cao et al. 2018)] have been retracted by their creators based on increasing ethical and legal grounds. Since they are no longer usable, they are not included in Table 6.

From Fig. 9, it can be noted that the data sets generated with GAN-based methods are more realistic. Kortyleski disregards the background and focuses only on facial appearance, which looks more realistic than that of FaceSynthetics.

From Table 6 it can be noted that synthetic data sets contain a much larger number of images than real data sets, except for CASIA-WebFace. The number of identities is also larger, in some cases much larger, which typically favours the generalisation capability of face recognition models. In addition, it can be seen that, unlike synthetic data sets, real ones are highly unbalanced in terms of the number of images per identity, which is detrimental to representativeness.

6 Effectiveness of Synthetic Data

As mentioned above, the main purpose of existing synthetic data sets in the CV field (as well as in other fields) is to be used as training data for machine learning-based models. Accordingly, a relevant issue in the context of this survey is to assess how effective they are as training data for CV models that will process real data after deployment. To comprehensively investigate the above issue, experiments on all the considered CV tasks should be carried out, using all the available synthetic data sets for training and at least the main benchmark data sets of real data for testing, which, however, would be very onerous. Accordingly, in the following, we report results from the papers where synthetic data sets have been proposed. We point out that not all of these papers present valuable quantitative results, e.g. due to the absence of a common testing data set, therefore, the following subsections do not cover all the considered VS tasks. Also, the results reported in different works are often not directly comparable due to different experimental settings, including the real data sets used for testing and the CNN architectures used. Moreover, synthetic images have not always been used

[20] https://www.kaggle.com/datasets/jessicali9530/lfw-dataset.

[21] https://www.kaggle.com/datasets/souvadrahati/extended-yale-dataset-b.

[22] https://www.scface.org/.

[23] https://www.kaggle.com/datasets/ntl0601/casia-webface.

[24] http://www.cfpw.io/.

[25] https://ibug.doc.ic.ac.uk/resources/agedb/.

[26] https://github.com/NVlabs/ffhq-dataset.

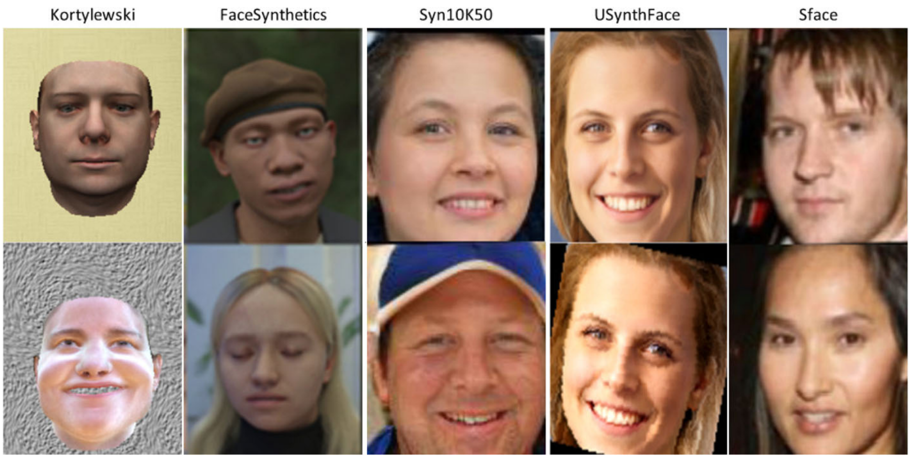

Fig. 9 Examples of images from synthetic data sets for face recognition. From left to right: Kortilesky (not available online), FaceSynthetics (source https://datagen.tech/blog/microsofts-face-analysis-in-the-wild-using-synthetic-data-alone-summarized/), Syn10K50 (not available online), USynthFace (not available online) and SFace (not available online)

Table 6 Statistics of synthetic data sets for face recognition, and comparison with the main benchmark, real data sets

	Data set	# Images	# Identities	# Images per id	Source
Synt	Kortilesky (Kortylewski et al. 2019)	1,000,000	20,000	100	3DMM
	FaceSynthetics (Wood et al. 2021)	100,000	100,000	1	RS
	Syn10K50 (Qiu et al. 2021)	500,000	10,000	50	DiscoFaceGAN
	SFace (Boutros et al. 2022)	634,000	10,575	60	StyleGAN2-ADA
	USynthFace (Boutros et al. 2023a)	400,000	400,000	1	DiscoFaceGAN
Real	Extended Yale B (Georghiades et al. 2001)	2414	> 38	–	–
	LFW (Huang et al. 2008)	13,233	5749	1/2.3/530	Web
	SCface (Grgic et al. 2011)	4160	130	–	VSC
	CASIA-WebFace (Yi et al. 2014)	494,414	10,575	2/46.8/804	Web
	CFPW (Sengupta et al. 2016)	7000	500	14	Web
	AgeDB (Moschoglou et al. 2017)	16,516	570	1/21/101	Web
	FFHQ (Karras et al. 2021)	70,000	–	–	Flickr

The column *Source* indicates either the approach for synthetic image generation (RS: rendering software) or the image source of real data (VSC: video surveillance cameras; information not available for Extended Tale B). The three entries in the 'Images per id.' column are the minimum/average/maximum number of images per identity

as the only data source for training. They were often used just for pre-training followed by fine-tuning on real images or directly mixed with real images for a single training process. Nevertheless, such results provide interesting hints on the effectiveness of synthetic vs real training data.

6.1 Crowd Counting

In Tables 7 and 8, we report the results obtained on one of the most widely used benchmark data sets (ShanghaiTech), for different CNN architectures, in terms of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Relative Error (MRE) (Jiang et al. 2022). In detail, Table 7 presents the performance reported by Hou et al. (2022) on ShanghaiTech B with or without pre-training on the synthetic data sets GCC and CrowdX. These results have been obtained for three CNN architectures: MCNN, CSRNet and ESA-Net. MCNN (Zhang et al. 2016a) is a multi-column CNN trained from scratch, while CSRNet (Li et al. 2018) and ESA-Net (Hou et al. 2022) are single-column architectures based on a VGG-16 (Simonyan and Zisserman 2015) backbone with different dilation and attention modules. The aim of such experiments was to assess the contribution of synthetic data sets in data representation and enhancement. Models pre-trained with synthetic data outperform the models without pre-training, which provides evidence of the effectiveness of synthetic data sets in data enhancement. Similarly, Table 8 shows the performance reported by Jiang et al. (2022) on ShanghaiTech A and ShanghaiTech B, using the CNN architecture PSDENet, which was proposed by the same authors, based on a VGG-16 backbone. Such architecture can also exploit the annotations related to the pedestrian binary mask to improve the localisation and counting performance. In this case, the authors evaluated not only the crowd counting model with or without pre-training on synthetic data sets, but also including (w/mask) or not (w/o mask) the annotation related to the binary mask. To this aim, they synthesised 1452 images (not publicly available) with resolution 1080 × 1920. The reported results provide further evidence that synthetic data can improve counting performance in real scenes, and that the use of pedestrian binary masks allows PSDENet to focus on image regions containing pedestrians, thus reducing the counting error. Note that such binary masks are not related to the head, but to the whole body, which is easily extracted from synthetic images, whereas it would require considerable annotation effort in real images.

6.2 Pedestrian Detection

For the pedestrian detection task, we did not find common real testing data sets to evaluate the performance of the three synthetic data sets mentioned in Sect. 5.3. Nevertheless, synthetic pedestrian tracking data sets can also be used for pedestrian detection, due to the presence of the corresponding annotations (i.e., bounding boxes). In Table 9, we report a limited comparison of some synthetic tracking data sets used for the detection task. In particular, we report performances obtained by Fabbri et al. (2021) on MOT17 (Milan et al. 2016), using different models trained on one synthetic data set (either VIPER, JTA or MOTSynth), using two common metrics, i.e., Average Precision (AP) and Multi-Object Detection Accuracy (MODA) (Bernardin and Stiefelhagen 2008). Reported results show that MOTSynth outperforms the other synthetic data sets in terms of AP and MODA. This can be due to the higher variability in the frames used for training (MOTSynth contains more than 1.3 million frames) with respect to the other data sets. This variability might improve the generalisation capability of the detector. The higher performances attained using VIPER can be due to the fact that this data set has been specifically generated for object tracking, instead of pedestrian tracking.

6.3 Person Re-Identification

As previously mentioned, the highest number of synthetic data sets is related to the Re-Id task, and all of them, except for VC-Clothes, have been evaluated in the respective papers on one or more common real benchmarks, i.e., Market-1501 and MSMT17 (see Sect. 5.5). Nevertheless, we point out that also for this task, a direct comparison between all the different synthetic data sets is hindered by some differences in the deep learning architectures used in the respective papers, as well as in the training protocols: either direct transfer, or fine-tuning on real data from the target data set, or using real auxiliary data together with synthetic ones for training, and then fine-tuning on real data from the target data set.

For the purpose of this work, we included only the cross-data set results, extracted from the respective papers, attained by training on synthetic data and testing on real data (i.e., Market-1501 and MSMT17). For comparison, we also included cross-data set experiments on Market-1501 and MSMT17, i.e., each of them was used in turn for training and the other for testing (Delussu et al. 2023).

Results are reported in Table 10 in terms of the two most common performance metrics for Re-Id: Cumulative Matching Curve (CMC) at rank 1 and mean Average Precision (mAP). Note that the latter provides a more complete account of the performance of a Re-Id system when more than one image of the query individual is present in the gallery (which is a realistic scenario in many real-world applications), whereas the CMC only considers the top-ranked one.

First, it can be observed that, for the same real data set used for testing, the best results were attained when synthetic-to-real domain adaptation techniques were used, as one could expect. Moreover, in such cases, models trained on both synthetic and real data always outperformed the ones trained only on real data, although the former was favoured by fine-tuning on the same target data, whereas a cross-data set setting was used for the latter.

The behaviour exhibited by models trained only on synthetic data sets (PersonX, RandPerson, FineGPR and UnrealPerson) is very different, instead. Notably, models trained on RandPerson attained performances close to the ones attained by training on real data, whereas models trained on UnrealPerson outperformed by a large margin, on all three target data sets, models trained on real data. Instead, when FineGPR and PersonX were used (especially the latter), the performances were clearly worse than using real training data.

To sum up, considering the above-mentioned limitations of our experimental evaluation, these results suggest that in the Re-Id task, synthetic data sets may already be capable of covering, or even overcoming, the performance gap with real training data, at least on benchmark data sets.

6.4 Face Recognition

Also for this task, a direct comparison between all the different synthetic data sets is hindered by some differences in the deep learning architectures used in the respective papers, as well as in the training and testing protocols. For such reason, we report the results of three different comparisons, which correspond to three different experimental protocols used in previous work: pre-training on synthetic data and fine-tuning on real auxiliary data, different from target data (Table 11); mixing real auxiliary data together with synthetic data for training (Table 12); training on synthetic data and testing on real ones (direct transfer approach), sometimes comparing the performance obtained with different training partitions (Table 13).

Table 11 presents the results reported by Kortylewski et al. (2019). They have been obtained by testing on LWF data set

Table 7 Crowd counting performances (MAE and RMSE) reported by Hou et al. (2022) on ShanghaiTech B using thee CNN architectures (MCNN, CSRNet and ESA-Net) with or without pre-training on synthetic data sets

Pre-training	MCNN		CSRNet		ESA-Net	
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
None/ImageNet	26.4	41.31	10.6	16.0	8.3	12.9
GCC (Wang et al. 2019)	8.8	28.21	10.1	15.7	–	–
CrowdX (Hou et al. 2022)	6.9	24.4	9.7	14.6	7.6	11.8

Table 8 Crowd counting performances (MAE, RMSE and MRE) reported by Jiang et al. (2022) on ShanghaiTech A (Part A) and ShanghaiTech B (Part B) using the PSDENet architecture with or without pre-training on synthetic data (not publicly available), including (w/mask) or not (w/o mask) the segmentation mask

Pre-training	Part A			Part B		
	MAE ↓	RMSE ↓	MRE ↓	MAE ↓	RMSE ↓	MRE ↓
None/ImageNet	68.54	111.48	0.1553	10.74	19.85	0.0852
Synt. w/o mask	65.22	106.27	0.1542	9.68	17.27	0.0789
Synt. w/mask	63.35	100.80	0.1516	9.40	17.12	0.0724

Table 9 Pedestrian detection performances (AP and MODA) reported by Fabbri et al. (2021) on the real data set MOT17 (Milan et al. 2016), using synthetic data sets for training different models (YOLOv3, CenterNet and Faster R-CNN)

Model	Training set	AP ↑	MODA ↑
YOLOv3 (Redmon and Farhadi 2018)	VIPER (Richter et al. 2017)	26.65	22.02
	JTA (Fabbri et al. 2018)	53.18	48.77
	MOTSynth (Fabbri et al. 2021)	71.90	64.51
CenterNet (Zhou et al. 2019)	VIPER (Richter et al. 2017)	44.58	39.92
	JTA (Fabbri et al. 2018)	60.15	45.38
	MOTSynth (Fabbri et al. 2021)	70.49	55.25
Faster R-CNN (Ren et al. 2015)	VIPER (Richter et al. 2017)	60.93	42.87
	JTA (Fabbri et al. 2018)	69.69	38.38
	MOTSynth (Fabbri et al. 2021)	78.98	54.96

Table 10 Performances (rank-1 and mAP) attained on two real benchmarks for person re-identification (Market-1501 and MSMT17) using synthetic data sets for training (results extracted from respective papers), and comparison with the performances attained by training on the same real data sets in a cross-domain setting [source: Delussu et al. (2023)]

	Training set	Architecture	Market-1501		MSMT17	
			mAP ↑	rank-1 ↑	mAP ↑	rank-1 ↑
Real	Market-1501 *** (Zheng et al. 2015)	ResNet50	–	–	6.44	19.08
	MSMT17 *** (Wei et al. 2018)	ResNet50	38.33	69.77	–	–
Synthetic	SOMAset* (Barbosa et al. 2018)	SOMAnet	53.5	77.49	–	–
	SyRI* (Bak et al. 2018)	ResNet-18	–	54.3	–	–
	PersonX*** (Sun and Zheng 2019)	PCB	20.4	44.0	3.6	11.7
	GPR** (Xiang et al. 2020)	ResNet-50†	50.8	76.2	–	–
	RandPerson*** (Wang et al. 2020)	ResNet-50	28.8	55.6	6.3	20.1
	FineGPR*** (Xiang et al. 2023)	ResNet-50	24.6	50.5	3.9	12.5
	Synthetic18k* (Uner et al. 2021)	DenseNet-121	77.3	91.6	–	–
	UnrealPerson*** (Zhang et al. 2021)	ResNet-50†	54.3	79.0	15.3	38.5
	ClonedPerson*** (Wang et al. 2022)	ResNet-50†	59.9	84.5	18.5	49.1
	WePerson*** (Li et al. 2021)	ResNet-50†	–	–	18.9	46.4

Training protocols for synthetic data: fine-tuning on the same real data set (*); using a different, auxiliary real data set during training and fine-tuning on the same real data set used for testing (**); direct transfer (***). † indicates that a modified ResNet-50 architecture was used after pre-training the FaceNet (Schroff et al. 2015) architecture with synthetic data only, or fine-tuning it with different percentages of real data from CASIA-WebFace. For comparison, they also reported the results obtained after training with the same percentages of real data only. The pre-trained models considerably outperform the models without pre-training, even when the full real data set was used, providing evidence that such a process improves generalisation capability, as well as data efficiency, since the amount of real data needed to achieve competitive performance was significantly reduced.

Table 12 presents the results reported by Qiu et al. (2021). They have been obtained by testing on LWF data set as well using however a ResNet-50 architecture (He et al. 2016) trained either on synthetic data only, or only on different subsets of real data from CASIA-WebFace, or mixing ($MixNS$) synthetic data with different sub-sets of real images (where N denotes the number of identities and S the number of samples per identity). The reported results show that mixing synthetic with real images can improve the appearance of synthetic images with real-world attributes (e.g., blur and illumination), which alleviates the domain gap. Moreover, by enlarging the intra-class variations of synthetic data the performance can be further improved.

Finally, Table 13 presents the results reported by Boutros et al. (2022, 2023a). They have been obtained by testing on LWF, AgeDB-30, CFPW, CA-LFW (Cross-Age LFW) and CP-LFW (Cross-Pose LFW) data sets with a ResNet-100 architecture (He et al. 2016) trained either only on real data from CASIA-WebFace, or only on synthetic data. In this case, the authors did not fine-tune or mix with real images, but evaluated different synthetic image sub-sets. In particular, the evaluation was aimed at assessing how much the performances of the considered model, trained under a fully supervised or unsupervised setting, changes by varying the number of images per identity (Boutros et al. 2022) or the total number of images (Boutros et al. 2023a), and, simultaneously, to evaluate to what extent GANs can generate identity-separable face images. Reported results show that increasing the size of synthetic training data in terms of samples per identity does generally increase face recognition performance, both for supervised and unsupervised models. Furthermore, such data sets exhibited, to a certain degree, the same identity discriminant information that is present in real data, even if with a convergent trend as the number of total images or the number of images per identity increases.

7 Discussion

In previous sections, we described the main issues of VS applications of CV related to the use of real data, approaches used to generate synthetic data, and existing synthetic data sets. In this section, we discuss to what extent synthetic data address the above-mentioned issues (see Sects. 1, 3), and propose some possible future research directions.

7.1 Existing Issues

In the following, we discuss the main issues common to all the considered applications: manual annotation effort, limited amount of real data, limited representativeness, data imbalance, and privacy restrictions.

Manual annotation is time-consuming, extremely laborious and prone to errors. It is worth noting that several kinds of annotations are required for the considered CV tasks, from local [up to pixel-level, e.g., object tracking (Cabon et al. 2020)] to global ones, i.e., related to a whole image or even a whole sequence of frames (e.g., a frame or a whole video labelled as showing a normal or abnormal crowd behaviour), or even both [e.g., in crowd counting (Wang et al. 2019)]. The generation of synthetic data completely solves this issue in all such tasks since all kinds of annotations of interest can be automatically generated with the desired accuracy. This is the case also for GANs and DMs, for the CV tasks where they have been used.

Data set size The possibility of automatic and accurate annotation of synthetic data sets, together with the availability of efficient generation techniques, allows generating synthetic data sets much larger than real ones, at least in terms of the bare number of images or videos. This is witnessed by the currently available synthetic data sets (see Sect. 5). For instance, the largest synthetic data sets contain over 280,000 images for crowd counting, 500,000 images for detection tasks, over 1.3 million frames for tracking tasks, 1 million videos for behaviour analysis, 4 million images for Re-Id and 1 million images for face recognition. This issue can therefore be considered to be solved for all the considered tasks.

Data imbalance In terms of the bare number of images or videos, a relatively large data set size is necessary to have a sufficient amount of examples of the patterns of interest in a given application. However, this is not sufficient to achieve a satisfactory generalisation capability, which is also affected, among other factors, by data imbalance. This issue can also affect synthetic data. In particular, although generative models can, in principle, generate an unlimited number of samples, currently they allow much more limited control of the different aspects of generated images (e.g., pedestrian pose and body shape) than other methods like computer graphics engines. This may result in some aspects being over-represented with respect to others. For instance, the USynthFace synthetic data set for face recognition (Boutros et al. 2023a) was generated from real face images using GANs, which in this task may allow obtaining more realistic images than other approaches. However, if the real input images are biased, e.g., in terms of gen-

Table 11 Accuracy on LWF reported by Kortylewski et al. (2019) after training the FaceNet architecture with synthetic data only, or fine-tuning it with different percentages of real data from CASIA-WebFace

Pre-training on synth. data	% of real data used for training	Accuracy ↑
Yes	0	80.1
No	100	94.1
Yes	100	95.8
No	25	89.1
Yes	25	93.6
No	10	85.1
Yes	10	91.8

For comparison, results obtained after training with the same percentages of real data only are also reported

Table 12 Accuracy on LWF data set reported by Qiu et al. (2021) after training the ResNet-50 architecture with either synthetic data only, real data only (CASIA-WebFace), or mixing (MixNS) synthetic data with different sub-sets of real images (where N denotes the number of identities and S the number of samples per identity)

Training set	Real ident	Real img. per ident	Total real img	Accuracy ↑
Real	10,575	47	494,414	99.18
Syn10K50	0	0	0	91.97
Real1K10	1K	10	10k	87.50
Mix1K10	1K	10	10k	92.28
Real1K20	1K	20	20k	92.53
Mix1K20	1K	20	20k	95.05
Real2K10	2K	10	20k	91.22
Mix2K10	2K	10	20k	95.78

For comparison, results obtained after training only with the same real data sub-sets are reported

der, ethnicity and age, the obtained synthetic images may exhibit a similar bias. This issue can instead be avoided if computer graphics engines or image composition techniques are used, since they allow a higher level of control than generative models, including the degree of data imbalance. However, we point out that this aspect is to some extent disregarded in some of the existing synthetic data sets. This can be observed, for instance, in the GCC and SHADE data sets for crowd counting and behaviour analysis, respectively, generated using GTAV. GCC (Wang et al. 2019) contains about 7 million pedestrians, but most of them show sparse or non-dense crowds; thus dense crowds are under-represented. Similarly, SHADE (Lin et al. 2021) night-time scenes are under-represented, since most of its videos represent scenes taking place between 7:30 a.m. and 5:30 p.m. To sum up, data imbalance could be avoided or at least mitigated for most of the considered VS tasks using computer graphics engines and image composition techniques, whereas it remains an open issue for synthetic data sets generated using generative models.

Representativeness Although synthetic images can be generated in large quantities, thus solving or limiting the issues of data set size and imbalance, it remains infeasible to consider and include all the possible patterns of interest in a *single*, *general-purpose* synthetic data set, e.g., all possible relevant events for behaviour analysis or all possible crowd sizes. This problem has been discussed by Yuille and Liu (2021), although in a slightly different context, related to the capability of deep neural networks to deal with the combinatorial explosion caused by the complexity of natural images. As an example, with the aim of rendering images of a *single* object from different viewpoints and illuminations and in a limited number of background scenes, using computer graphics, by setting 13 parameters related to camera pose, lighting, texture, material and scene layout; assuming that 1000 values are used for each of them, it is estimated (Yuille and Liu 2021) that a total of 10^{39} different images can be generated. A similar combinatorial explosion would clearly occur in the considered VS applications, where synthetic images or videos should present, e.g., different weather and illumination conditions, time of day, backgrounds, perspectives, poses etc. As an example, consider 7 different weather conditions (i.e., clear, clouds, rain, foggy, thunder, overcast, extra sunny), 8 daytime, 100 values for the camera distance, elevation, and angle parameters, and 100 illumination conditions, and assume that one wants to generate images in 100 different locations. This would amount to generating about 10^9 images. Moreover, a specific application, such as Re-Id, is likely to involve additional parameters with respect to the ones mentioned above; for instance, adding just four different pedestrian poses (front, back and both sides) and clothing appearance for both lower and upper body, would lead to generating at least five orders of magnitude more images. Accordingly, the representativeness issue cannot be considered to be solved by synthetic images, although they allow to mitigate it with respect to the use of real images.

Table 13 Accuracy on LWF, AgeDB, CFPW, CA-LFW and CP-LFW data sets reported by Boutros et al. (2022, 2023a) after training the ResNet-100 architecture with either real (CASIA-WebFace) data only, or different sub-sets of synthetic data

Source	Superv	Training set	Ident	Img. per ident	Total	LFW	AgeDB	CFPW	CA-LFW	CP-LFW
Boutros et al. (2022)	Yes	CASIA-WebFace	10,575	≈ 47	494,414	99.55	95.31	94.55	93.78	89.95
		SFace (10%)	10,575	10	105K	87.13	63.30	68.84	73.47	66.82
		SFace (20%)	10,575	20	211K	90.50	69.17	73.33	76.35	71.17
		SFace (40%)	10,575	40	423K	91.43	69.87	73.10	76.92	73.42
		SFace (60%)	10,575	60	634K	91.87	71.68	73.86	77.93	73.20
Boutros et al. (2023a)	No	USynthFace	100K	1	100K	91.52	69.30	78.46	75.35	71.93
		USynthFace	200K	1	200K	91.93	71.23	78.03	76.73	72.27
		USynthFace	400K	1	400K	92.23	71.62	78.56	77.05	72.03

Privacy Synthetic images are instead an effective solution to achieve compliance with privacy regulations [e.g., EU's GDPR (European Commission 2020)] when people are involved. To this aim, synthetic data sets can be generated in two main ways: (i) manipulating real images (e.g., using GANs), provided they had been acquired with user consent (EiC of Pattern Recognition 2022); (ii) generating images containing synthetic pedestrians, either in a virtual environment (e.g., using computer graphics engines) or using a real environment image (e.g., using image composition techniques). In the first case, a possible solution is to generate new images [e.g., with different background (Chen et al. 2019; Hussin and Yildirim 2021) or camera style (Tian et al. 2021; Liu et al. 2019)]. Basically, this strategy tends to enlarge a specific real data set by combining real and synthetic images and has already been adopted using GANs (Tian et al. 2021; Hussin and Yildirim 2021). In the second case, images generated using computer graphic engines and image composition approaches can optionally be refined using GANs to increase their degree of realism (Wang et al. 2019).

7.2 Future Research Directions

In the following, we present three interesting directions for future work on synthetic data (not limited to the considered VS tasks).

Degree of realism Although several synthetic data sets have been proposed so far, only a few works have thoroughly investigated the factors that affect their effectiveness, particularly their degree of realism. Preliminary investigations on the degree of realism have been carried out for crowd counting (Ledda et al. 2021; Delussu et al. 2020), object detection (Dvornik et al. 2021; Wu et al. 2023), Re-Id (Hussin and Yildirim 2021; Delussu et al. 2022a), face recognition (Kim et al. 2023). Results of our investigations on crowd counting (Ledda et al. 2021; Delussu et al. 2020) suggest that the lower the distance of pedestrians from the camera, the higher the degree of realism in their appearance (e.g., clothing details) required to attain a certain accuracy. For the Re-Id task, we found that the quality of human models and virtual environments seems more relevant than the total number of training images (Delussu et al. 2022a).

For object detection, an interesting insight was provided by Dvornik et al. (2021), who distinguished between object *instance* detection, which is a fine-grained task, and object *category* detection, which has to account for large intra-class variability. They pointed out that an image pasting approach based on placing instances of the objects of interest in *random* positions of existing scenes (Dwibedi et al. 2017), which is characterised by a relatively low degree of realism (e.g., a bottle appearing in the sky), is effective for *instance* detection, but is ineffective for *category* detection. For the latter, instead, it is also necessary to place objects in the appropriate context. On the other hand, illumination conditions and blending artefacts turned out to negatively affect instance detection, whereas they are not critical for category detection.

The realism of images generated by generative models has been analysed for the Re-Id (Hussin and Yildirim 2021) and face recognition (Kim et al. 2023) tasks. To this aim, the Fréchet Inception Distance (FID) (Heusel et al. 2017) was used. The FID metric was proposed to evaluate the effectiveness of GANs by measuring the distance between the feature distribution of generated data and real data (not limited to images). In particular, the lower the FID, the closer the distribution of synthetic data to the one of real data, which is expected to result in a better performance of a model trained on synthetic data. This behaviour was empirically validated by Hussin and Yildirim (2021) and Kim et al. (2023). It is worth highlighting that most of the works about diffusion models focus on the quality of generated data by using several metrics besides FID (Croitoru et al. 2023). Considering that many of these metrics, such as FID, are based on the use of the Inception layer, in our opinion it would be appropriate to investigate them accurately, in order to define which metrics can be used for a specific task to compare results or performances fairly. Also, we believe that developing FID-like metrics tailored to the other approaches for synthetic image generation discussed in this work, besides generative-based ones, is an interesting direction for future work. Indeed, such metrics could be beneficial to evaluate whether a given synthetic data set (generated with approaches different from generative models) is suitable for a specific task and to provide a measure of its degree of realism.

According to existing evidence, the required degree of realism is likely to be not only application-specific [see, in particular, the above-mentioned work by Dvornik et al. (2021)] but also related to the specific CNN model, as suggested by our previous work on crowd counting (Ledda et al. 2021). In particular, we observed that the type of model, i.e., either regression- or detection-based (see Sect. 3.1), influences crowd counting accuracy, and that the relationship between accuracy and degree of realism can be even counter-intuitive (Ledda et al. 2021). For instance, in our experiments, detection-based models attained a *worse* performance when trained on synthetic images with a *higher* degree of realism, i.e., a realistic background vs a uniform one, and a rich pedestrian's clothing appearance vs simple human models rendered with a uniform colour.

The above results also suggest that different degrees of realism may be required for different image components or aspects (e.g., image background, human models and illumination conditions), depending on the task. Therefore, the envisaged investigation should also focus on the degree of realism of the different image components and aspects, besides considering the degree of realism of an image as a whole.

Using synthetic data for testing As mentioned earlier (and as shown in detail in Sect. 6) synthetic data at present is mainly used for pre-training or training (also mixed with auxiliary real data) of CNN-based methods. A further interesting research direction is to investigate if they can also be used as a possible alternative to real data sets for benchmarking, i.e., as *testing* data.

So far, synthetic images have been used for testing mainly in specific simulations for which no real data was available, e.g., in early work based on the analysis by synthesis approach (see Sect. 1). In particular, to analyse the discriminant capability of complex textures by the human visual system (Pratt et al. 1978), to evaluate the robustness of algorithms for optical flow estimation (Horn and Schunck 1981) and analyse the properties of land cover for remote sensing tasks, under different types of acquisition conditions (e.g., in terms of illumination and weather conditions) (Woodham et al. 1985). In principle, synthetic testing data should exhibit similar advantages as synthetic training data, e.g., larger data set size, no manual annotation effort and no privacy issues. To our knowledge, the only work where this issue has been addressed so far in the context of CV tasks related to VS is the one by Kang (2023), which focused on Re-Id. Its goal was to assess the "reliability" of synthetic data sets in the evaluation of the generalisation capability of Re-Id models, in a cross-data set setting. To this aim, ten different Re-Id models were trained on each of three real data sets (CUHK03, Market-1501 and MSMT17), and two synthetic data sets (RandPerson and UnrealPerson, see Sect. 5.5), and were then tested on the remaining two real data sets, and on the ClonedPerson synthetic data set (which is the only synthetic data set for Re-Id including a testing partition). Then, the Kendall rank correlation coefficient τ was computed to evaluate the pairwise similarity between the ranking of the ten models, in terms of rank-1 accuracy and mAP, on each pair of testing data sets. Finally, the non-parametric two-sample Kolmogorov-Smirnov test was used to evaluate whether the distribution of Kendall's τ between pairs of real testing data was statistically identical to that between real and synthetic testing data. Reported results show that this was the case, with a significance level $\alpha = 0.05$. Quoting from Kang (2023), the conclusion was that "the synthetic data set ClonedPerson can be reliably used to benchmark generalisable person re-identification algorithms, with no statistical difference to real-world data sets."

The above results are very promising, although they are still limited to a single task (Re-Id), and to a single synthetic data set. Therefore, extending these experiments to other data sets and tasks is very interesting. To this aim, the only issue to be addressed is the definition of standard training and testing partitions for synthetic data sets. Indeed, for most of them, no separate testing partition has been defined, precisely because they were designed for training purposes only. A standard partition could be easily defined for large enough data sets, e.g., in terms of identities and images per identity for the Re-Id task; otherwise, a new testing set can be generated using the same approach as the original data set.

A related issue is to investigate whether and to what extent the performance score (e.g., rank-1 accuracy or mAP, for Re-Id) of a given model on a synthetic testing set is representative of its performance on real data. In particular, one may expect that the performance of synthetic data sets is less representative if they exhibit a larger visual gap to real data (i.e., a lower degree of realism). For instance, a relatively higher visual gap can be observed between real data sets and computer graphics-based synthetic data sets (e.g. crowd counting, Re-Id, etc.), whereas the lowest visual gap seems to occur with GAN-based synthetic data sets for face recognition (see Sect. 5.6).

Synthetic image generation tools Another future direction is the development of software tools based on computer graphics or other rendering software, to allow end users to generate their own data sets with low effort, similar to the approaches proposed in our previous work for crowd counting (Delussu et al. 2022b) and by Hattori et al. (2018) for pedestrian detection. Some companies already propose this kind of service. Although it may be difficult to develop a single tool suitable for several different VS tasks, such as all the ones considered in this survey, it may be possible to include related tasks such as object/pedestrian tracking and detection, or crowd counting and crowd behaviour analysis.

One of the functionalities we envisage for this kind of tool is generating *scene-specific* synthetic data sets, which can be useful for application scenarios involving the deployment of a system (e.g., for crowd counting) on a new scene, without collecting and manually annotating real data (Delussu et al. 2022b; Hattori et al. 2018).

As an example, we sketch here some guidelines to implement this functionality for the crowd counting tasks, based on our previous experience on it (Delussu et al. 2022b). The envisaged tool should allow the user to use a *real* background image of the scene of interest: this would guarantee a high degree of realism in this aspect, without the effort of building a synthetic model of the scene background. The tool should also allow the user to provide information useful to reproduce the scene perspective corresponding to the camera position and to consistently place pedestrians in the correct image regions. For instance, to this aim, the user could select the region of the background image where people can appear; she could also use a real image of the target scene, including some pedestrians in different locations, and draw a bounding box around a few of them to allow to tool to automatically compute the perspective map, which in turn allows to re-scale synthetic pedestrians accordingly. The above functionality could also be useful to the other VS tasks considered in this survey, except for face recognition. Moreover, other task-specific information may be acquired from the user. For instance, for crowd counting, the user could provide the maximum expected number of pedestrians in the scene, which would allow the synthetic image generation tool to generate a data set representative of the different crowd sizes of interest. Analogous information can be acquired by the envisaged tool for tracking, detection and Re-Id, e.g., to generate a data set with a desired degree of occlusions. In the case of behaviour analysis tasks, the tool should embed a predefined set of events (e.g., panic escape and fight), among which the user can choose the ones of interest.

It is worth noting that researchers have already developed a few such prototype tools, for Re-Id (Zhang et al. 2021; Chen et al. 2022), object detection (Wu et al. 2023) and crowd counting (Wang et al. 2019). In particular, the Re-Id tool by Zhang et al. (2021) can also be used for other VS tasks such as pedestrian tracking. It allows generating 3D human models, scenes and animations, as well as extracting bounding boxes and tracklets. The crowd counting tool (Wang et al. 2019) consists of a data collector and labeller, which automatically generates head points annotations. An open-source web annotation tool, called Scalabel, has been used to annotate synthetic data for tracking task (Sun et al. 2022). The tool can also be used for other tasks since it allows to provide several annotations, e.g., 2D and 3D bounding boxes, and 2D instance segmentation. The diffusion model-based tool proposed in Wu et al. (2023) allows to provide data with several annotations, e.g., human pose, depth, and semantic, instance and deep-fashion masks.

Concerning face recognition, the Adobe software Adobe (n.d.) could be exploited inside a specific tool for generating data sets of synthetic face images: it provides a library of 3D face models and allows users to refine them by tuning a large number of parameters, such as eye dimension and skin roughness.

We finally point out that the envisaged synthetic image generation tools could benefit from the results of the investigation discussed at the beginning of this section, aimed at understanding how the different aspects of synthetic data (e.g., the degree of realism) affect their effectiveness for model training.

8 Conclusion

In this work, we surveyed the use of synthetic training data focusing on applications related to video surveillance. In particular, we focused on crowd counting, object and pedestrian detection and tracking, behaviour analysis, person re-identification and face recognition. In such applications, the use of synthetic images is even more relevant to address specific issues arising from unconstrained acquisition conditions, as well as well-known issues of real data.

We first described each of the applications we focused on, emphasising the requirements in terms of data and task-specific issues. We then categorised and discussed the existing methods for creating synthetic data, highlighting their main pros and cons for the mentioned video surveillance applications. We also analysed the synthetic data sets proposed in the literature for each of the considered applications and provided an overview of their effectiveness as training data. Finally, we discussed whether and to what extent the existing synthetic data sets mitigate the issues of real data, highlight existing open issues, and suggested future research directions in this field.

Our main finding is that the data synthesis allows to completely solve some issues, i.e., manual annotation and data set size, while others are partially solved, i.e., data imbalance, representativeness, and privacy, and therefore require further work. Moreover, we identified interesting directions for further work related to the following main aspects: (i) investigating what kind and what degree of realism is required from synthetic data to be effective as training data; some work has already evaluated the photorealism of images generated by GANs and DMs, although it was not related to their use as training data; (ii) investigating whether and under what conditions synthetic data can also be used also to *assess* the performance of machine learning models, i.e., as *testing* set; (iii) the development of tools to allow users to generate ad hoc synthetic data sets.

Acknowledgements This work was supported by the projects: “Law Enforcement agencies human factor methods and Toolkit for the Security and protection of CROWDs in mass gatherings” (LETSCROWD), EU Horizon 2020 programme, Grant Agreement No. 740466; “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. This work was partially supported by the Italian Ministry of Enterprises and Made in Italy (MIMIT), within the 5 G technology support program, on axis 1 “House of Emerging Technologies” (CTE), Project Name “Cagliari Digital Lab” (ID: G27F22000040008).

Funding Open access funding provided by Università degli Studi di Cagliari within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

References

Abbass, M. Y., Kwon, K., Kim, N., Abdelwahab, S. A. S., El-Samie, F. E. A., & Khalaf, A. A. M. (2021). A survey on online learning for visual tracking. *The Visual Computer, 375*, 993–1014. https://doi.org/10.1007/s00371-020-01848-y

Abdolahnejad, M., & Liu, P. (2020). Deep learning for face image synthesis and semantic manipulations: A review and future perspectives. *Artificial Intelligence Review, 538*, 5847–5880. https://doi.org/10.1007/s10462-020-09835-4

Adobe. (n.d.). Adobe fuse. https://www.adobe.com/it/wam/fuse.html.

Ainam, J., Qin, K., Liu, G., & Luo, G. (2019a). Person re-identification through clustering and partial label smoothing regularization. In *ACM international conference proceeding series* (pp. 189–193).

Ainam, J., Qin, K., Liu, G., & Luo, G. (2019b). Sparse label smoothing regularization for person re-identification. *IEEE Access, 7*, 27899–27910. https://doi.org/10.1109/ACCESS.2019.2901599

Aranjuelo, N., García, S., Loyo, E., Unzueta, L., & Otaegui, O. (2021). Key strategies for synthetic data generation for training intelligent systems based on people detection from omnidirectional cameras. *Computers & Electrical Engineering, 92*, 107105. https://doi.org/10.1016/j.compeleceng.2021.107105

Autodesk Inc. (n.d.). Autodesk. https://www.autodesk.eu/.

Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., & Fleet, D. J. (2023). Synthetic data from diffusion models improves imagenet classification. CoRRabs/2304.08466. https://doi.org/10.48550/ARXIV.2304.08466.

Bak, S., Carr, P., & Lalonde, J. (2018). Domain adaptation through synthesis for unsupervised person re-identification. *European Conference on Computer Vision (ECCV), 11217*, 193–209.

Barbosa, I. B., Cristani, M., Caputo, B., Rognhaugen, A., & Theoharis, T. (2018). Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *Computer Vision and Image Understanding, 167*, 50–62. https://doi.org/10.1016/j.cviu.2017.12.002

Bernardin, K., & Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*. https://doi.org/10.1155/2008/246309

Blackmagic Design. (n.d.). Black magic design. https://www.blackmagicdesign.com/uk.

Blade, A. (n.d.). Grand theft auto V, Script Hook V. http://www.dev-c.com/gtav/scripthookv/.

Blender Online Community. (n.d.). Blender. https://www.blender.org/.

Boutros, F., Huber, M., Siebke, P., Rieber, T., & Damer, N. (2022). Sface: Privacy-friendly and accurate face recognition using synthetic data. *International Joint Conference on Biometrics (IJCB)*. https://doi.org/10.1109/IJCB54206.2022.10007961

Boutros, F., Klemt, M., Fang, M., Kuijper, A., & Damer, N. (2023). Unsupervised face recognition using unlabeled synthetic data. *International Conference on Automatic Face and Gesture Recognition*. https://doi.org/10.1109/FG57933.2023.10042627

Boutros, F., Struc, V., Fierrez, J., & Damer, N. (2023). Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing, 135*, 104688. https://doi.org/10.1016/j.imavis.2023.104688

Cabon, Y. , Murray, N., & Humenberger, M. (2020). Virtual KITTI 2. CoRRabs/2001.10773.

Cao, Q., Shen, L., Xie, W., Parkhi, O.M., & Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *Proceedings—13th IEEE international conference on automatic face and gesture recognition, FG67-74*. https://doi.org/10.1109/FG.2018.00020.

Chan, A. B., Liang, Z. -S. J., & Vasconcelos, N. (2008). Privacy preserving crowd monitoring: Counting people without people models or tracking. In *International conference on computer vision and pattern recognition, CVPR* (pp. 1–7).

Chen, K., Chen, W., He, T., Du, R., Wang, F., Sun, X., & Ding, G. (2022). Tagperson: A target-aware generation pipeline for person re-identification. In *MM: The 30th ACM international conference on multimedia* (pp. 560–571).

Chen, Y., Zhu, X., & Gong, S. (2019). Instance-guided context rendering for cross-domain person re-identification. In *International conference on computer vision, ICCV* (pp. 232–242).

Cheung, E., Wong, A., Bera, A., Wang, X., & Manocha, D. (2019). Lcrowdv: Generating labeled videos for pedestrian detectors training and crowd behavior learning. *Neurocomputing, 337*, 1–14. https://doi.org/10.1016/j.neucom.2018.08.085

Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *International conference on computer vision and pattern recognition, CVPR* (pp. 8789–8797).

Ciampi, L., Messina, N., Falchi, F., Gennaro, C., & Amato, G. (2020). Virtual to real adaptation of pedestrian detectors. *Sensors, 2018*, 5250. https://doi.org/10.3390/s20185250

Community, M. (2020). MakeHuman: Open source tool for making 3D characters. http://www.makehumancommunity.org.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *International conference on computer vision and pattern recognition, CVPR* (pp. 3213–3223).

Courty, N., Allain, P., Creusot, C., & Corpetti, T. (2014). Using the agoraset dataset: Assessing for the quality of crowd video analysis methods. *Pattern Recognition Letters, 44*, 161–170. https://doi.org/10.1016/j.patrec.2014.01.004

Croitoru, F., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 459*, 10850–10869. https://doi.org/10.1109/TPAMI.2023.3261988

de Souza, C. R., Gaidon, A., Cabon, Y., & Peña, A. M. L. (2017). Procedural generation of videos to train deep action recognition networks. In *International conference on computer vision and pattern recognition, CVPR* (pp. 2594–2604).

Delussu, R., Putzu, L., & Fumera, G. (2020). Investigating synthetic data sets for crowd counting in cross-scene scenarios. In *International joint conference on computer vision, imaging and computer graphics theory and applications, VISIGRAPP: Visapp* (Vol. 4, pp. 365–372).

Delussu, R., Putzu, L., & Fumera, G. (2022a). On the effectiveness of synthetic data sets for training person re-identification models. In *26th international conference on pattern recognition, ICPR* (pp. 1208–1214).

Delussu, R., Putzu, L., & Fumera, G. (2022b). Scene-specific crowd counting using synthetic training images. *Pattern Recognition, 124*, 108484. https://doi.org/10.1016/j.patcog.2021.108484

Delussu, R., Putzu, L., & Fumera, G. (2023). Human-in-the-loop cross-domain person re-identification. *Expert Systems with Applications, 226*, 120216. https://doi.org/10.1016/j.eswa.2023.120216

Dendorfer, P., Rezatofighi, S. H., Milan, A., Shi, J., Cremers, D., Reid, I. D., & Leal-Taixé, L. (2019). CVPR19 tracking and detection challenge: How crowded can it get?. CoRRabs/1906.04567.

Ding, G., Zhang, S., Khan, S. H., & Tang, Z. (2018). Center based pseudo-labeling for semi-supervised person re-identification. In *ICME workshops* (pp. 1–6).

Ding, G., Zhang, S., Khan, S. H., Tang, Z., Zhang, J., & Porikli, F. (2019). Feature affinity-based pseudo labeling for semi-supervised person re-identification. *IEEE Transactions on Multimedia, 2111*, 2891–2902. https://doi.org/10.1109/TMM.2019.2916456

Dong, G., Liao, G., Liu, H., & Kuang, G. (2018). A review of the autoencoder and its variants: A comparative perspective from target recognition in synthetic-aperture radar images. *IEEE Geoscience and Remote Sensing Magazine, 63*, 44–68. https://doi.org/10.1109/MGRS.2018.2853555

Dosovitskiy, A., Ros, G., Codevilla, F., López, A. M., & Koltun, V. (2017). CARLA: An open urban driving simulator. In *Conference on robot learning, corl* (Vol. 78, pp. 1–16).

Dvornik, N., Mairal, J., & Schmid, C. (2021). On the importance of visual context for data augmentation in scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 436*, 2014–2028. https://doi.org/10.1109/TPAMI.2019.2961896

Dwibedi, D., Misra, I., & Hebert, M. (2017). Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *International conference on computer vision, ICCV* (pp. 1310–1319).

EiC of Pattern Recognition. (2022). Expression of concern: "what-and-where to match: Deep spatially multiplicative integration networks for person re-identification". *Pattern Recognition 76*, 727–738.

Ekbatani, H. K., Pujol, O., & Seguí, S. (2017). Synthetic data generation for deep learning in counting pedestrians. In *International conference on pattern recognition applications and methods, ICPRAM* (pp. 318–323).

Elbishlawi, S., Abdelpakey, M. H., ElTantawy, A., Shehata, M. S., & Mohamed, M. M. (2020). Deep learning-based crowd scene analysis survey. *Journal of Imaging, 69*, 95. https://doi.org/10.3390/jimaging6090095

Epic Games. (n.d.). Unreal engine. https://www.unrealengine.com/en-US/.

Ergys, R., Solera, F., Zou, R., Rita, C., & Carlo, T. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV workshops* (pp. 17–35).

European Commission. (2020). On artificial intelligence-a European approach to excellence and trust.

European Union Agency for Fundamental Rights. (2019). Facial recognition technology: Fundamental rights considerations in the context of law enforcement. Publications Office of the European Union.

Fabbri, M., Brasó, G., Maugeri, G., Cetintas, O., Gasparini, R., Osep, A., & Cucchiara, R. (2021). Motsynth: How can synthetic data help pedestrian detection and tracking?. In *IEEE/CVF international conference on computer vision, ICCV* (pp. 10829–10839).

Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., & Cucchiara, R. (2018). Learning to detect and track visible and occluded body joints in a virtual world. In *European conference on computer vision, ECCV* (Vol. 11208, pp. 450–466).

Farooq, M., Dailey, M. N., Mahmood, A., Moonrinta, J., & Ekpanyapong, M. (2021). Human face super-resolution on poor quality surveillance video footage. *Neural Computing and Applications, 3320*, 13505–13523. https://doi.org/10.1007/s00521-021-05973-0

Frolov, S., Hinz, T., Raue, F., Hees, J., & Dengel, A. (2021). Adversarial text-to-image synthesis: A review. *Neural Networks, 144*, 187–209. https://doi.org/10.1016/j.neunet.2021.07.019

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The kitti vision benchmark suite. In *IEEE conference on computer vision and pattern recognition, CVPR* (pp. 3354–3361).

Georghiades, A. S., Belhumeur, P. N., & Kriegman, D. J. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 236*, 643–660. https://doi.org/10.1109/34.927464

Ghosh, S., Amon, P., Hutter, A., & Kaup, A. (2017). Pedestrian counting using deep models trained on synthetically generated images. In *International joint conference on computer vision, imaging and computer graphics theory and applications (VISIGRAPP): Visapp* (pp. 86–97).

Goodfellow, I. J. (2017). NIPS 2016 tutorial: Generative adversarial networks. CoRRabs/1701.00160.

Grgic, M., Delac, K., & Grgic, S. (2011). Scface–surveillance cameras face database. *Multimedia Tools and Applications, 51*(3), 863–879.

Guo, G., & Zhang, N. (2019). A survey on deep learning based face recognition. *Computer Vision and Image Understanding*. https://doi.org/10.1016/j.cviu.2019.102805

Guo, T., Huynh, C. P., & Solh, M. (2019). Domain-adaptive pedestrian detection in thermal images. In *IEEE international conference on image processing, ICIP* (pp. 1660–1664).

Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV9907 LNCS* (pp. 87–102). https://doi.org/10.1007/978-3-319-46487-9_6/FIGURES/5.

Han, H., Ma, W., Zhou, M., Guo, Q., & Abusorrah, A. (2021). A novel semi-supervised learning approach to pedestrian reidentification. *IEEE Internet Things Journal, 8*(4), 3042–3052. https://doi.org/10.1109/JIOT.2020.3024287

Han, J., Karaoglu, S., Le, H., & Gevers, T. (2020). Object features and face detection performance: Analyses with 3d-rendered synthetic data. In *International conference on pattern recognition, ICPR* (pp. 9959–9966).

Hattori, H., Boddeti, V. N., Kitani, K. M., & Kanade, T. (2015). Learning scene-specific pedestrian detectors without real data. In *International conference on computer vision and pattern recognition, CVPR* (pp. 3819–3827).

Hattori, H., Lee, N., Boddeti, V. N., Beainy, F., Kitani, K. M., & Kanade, T. (2018). Synthesizing a scene-specific pedestrian detector and pose estimator for static video surveillance: Can we learn pedestrian detectors and pose estimators without real data? *International Journal of Computer Vision, 1269*, 1027–1044. https://doi.org/10.1007/s11263-018-1077-3

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *International conference on computer vision and pattern recognition (cvpr)* (pp. 770–778).

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems 30: Annual conference on neural information processing systems* (pp. 6626–6637).

Ho, J. , Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, (neurips)*.

Horn, B., & Schunck, B. (1981). Determining optical flow. *Artificial Intelligence, 17*(1–3), 185–203. https://doi.org/10.1016/0004-3702(81)90024-2

Hou, Y., Li, C., Lu, Y., Zhu, L., Li, Y., Jia, H., & Xie, X. (2022). Enhancing and dissecting crowd counting by synthetic data. *ICASSP Proceedings, 2022*, 2539–2543. https://doi.org/10.1109/ICASSP43922.2022.9747070

Hou, Y., Zhang, S., Ma, R., Jia, H., & Xie, X. (2023). Frame-recurrent video crowd counting. *IEEE Transactions on Circuits and Systems for Video Technology*. https://doi.org/10.1109/TCSVT.2023.3250946

Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'real-life' images: detection, alignment, and recognition*.

Hussin, S. H. S., & Yildirim, R. (2021). StyleGAN-ISRO method for person re-identification. *IEEE Access, 9*, 13857–13869. https://doi.org/10.1109/ACCESS.2021.3051723

Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Máadeed, S., Rajpoot, N. M., & Shah, M. (2018a). Composition loss for counting, density map estimation and localization in dense crowds. In *European conference on computer vision, ECCV* (Vol. 11206, pp. 544–559).

Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., & Shah, M. (2018b). Composition loss for counting, density map estimation and localization in dense crowds. In *European conference on computer vision, ECCV* (Vol. 11206, pp. 544–559).

Jaipuria, N., Zhang, X., Bhasin, R., Arafa, M., Chakravarty, P., Shrivastava, S., & Murali, V. N. (2020). Deflating dataset bias using synthetic data augmentation. In *CVPR workshops* (pp. 3344–3353).

Jiang, X., Liu, H., Zhang, L., Li, G., Xu, M., Lv, P., & Zhou, B. (2022). Transferring priors from virtual data for crowd counting in real world. *Frontiers of Computer Science, 16*, 1–8. https://doi.org/10.1007/S11704-021-0387-8/METRICS

Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S. N., Rosaen, K., & Vasudevan, R. (2017). Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?. In *International conference on robotics and automation, ICRA* (pp. 746–753).

Júnior, J. C. S. J., Musse, S. R., & Jung, C. R. (2010). Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine, 27*(5), 66–77. https://doi.org/10.1109/MSP.2010.937394

Kang, C. (2023). Is synthetic dataset reliable for benchmarking generalizable person re-identification?. In *International joint conference on biometrics (ijcb)* (pp. 1–8).

Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O. I., & Radke, R. J. (2019). A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(3), 523–536. https://doi.org/10.1109/TPAMI.2018.2807450

Karras, T., Laine, S., & Aila, T. (2021). A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(12), 4217–4228. https://doi.org/10.1109/TPAMI.2020.2970919

Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., & Brossard, E. (2016). The megaface benchmark: 1 million faces for recognition at scale. In *International conference on computer vision and pattern recognition, CVPR* (pp. 4873–4882).

Kerim, A., Aslan, C., Celikcan, U., Erdem, E., & Erdem, A. (2021). Nova: Rendering virtual worlds with humans for computer vision tasks. *Computer Graphics Forum*. https://doi.org/10.1111/cgf.14271

Kerim, A., Celikcan, U., Erdem, E., & Erdem, A. (2021). Using synthetic data for person tracking under adverse weather conditions. *Image and Vision Computing, 111*, 104187. https://doi.org/10.1016/j.imavis.2021.104187

Kim, M., Liu, F., Jain, A. K., & Liu, X. (2023). Dcface: Synthetic face generation with dual condition diffusion model. In *Conference on computer vision and pattern recognition, CVPR* (pp. 12715–12725).

Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., & Vetter, T. (2019). Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *CVPR workshops* (pp. 2261–2268).

Kortylewski, A., Schneider, A., Gerig, T., Egger, B., Morel-Forster, A., & Vetter, T. (2018). Training deep face recognition systems with synthetic data. CoRRabs/1802.05891.

Lazaridis, L., Dimou, A., & Daras, P. (2018). Abnormal behavior detection in crowded scenes using density heatmaps and optical flow. In *European signal processing conference, EUSIPCO* (pp. 2060–2064).

Ledda, E., Putzu, L., Delussu, R., Loddo, A., & Fumera, G. (2021). How realistic should synthetic images be for training crowd counting models?. In *International conference on computer analysis of images and patterns (CAIP)* (Vol. 13053, pp. 46–56).

Leng, Q., Ye, M., & Tian, Q. (2020). A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology, 30*(4), 1092–1108. https://doi.org/10.1109/TCSVT.2019.2898940

Li, B., Huang, H., Zhang, A., Liu, P., & Liu, C. (2021). Approaches on crowd counting and density estimation: A review. *Pattern Analysis and Applications, 24*(3), 853–874. https://doi.org/10.1007/s10044-021-00959-z

Li, C., Ge, S., Zhang, D., & Li, J. (2020). Look through masks: Towards masked face recognition with de-occlusion distillation. In *International conference on multimedia* (pp. 3016–3024).

Li, H. , Ye, M., & Du, B. (2021). Weperson: Learning a generalized re-identification model from all-weather virtual data. In *MM: ACM multimedia conference* (pp. 3115–3123).

Li, W. (2021). Image synthesis and editing with generative adversarial networks (GANs): A review. In *World conference on smart trends in systems security and sustainability, worlds* (Vol. 4, pp. 65–70).

Li, W., Mahadevan, V., & Vasconcelos, N. (2014). Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(1), 18–32. https://doi.org/10.1109/TPAMI.2013.111

Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *International conference on computer vision and pattern recognition, CVPR* (pp. 152–159).

Li, X., Dong, N., Huang, J., Zhuo, L., & Li, J. (2021). A discriminative self-attention cycle GAN for face super-resolution and recognition. *IET Image Processing, 15*(11), 2614–2628. https://doi.org/10.1049/ipr2.12250

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *International conference on computer vision and pattern recognition, CVPR* (pp. 3204–3213).

Li, Y., Zhang, X., & Chen, D. (2018). Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Cvpr* (pp. 1091–1100).

Li, Z., Guo, J., Jiao, W., Xu, P., Liu, B., & Zhao, X. (2020). Random linear interpolation data augmentation for person re-identification. *Multimedia Tools and Applications, 79*(7–8), 4931–4947. https://doi.org/10.1007/s11042-018-7071-5

Lin, C., Kew, J., Chan, C. S., Lai, S., & Zach, C. (2023). Cycle-object consistency for image-to-image domain adaptation. *Pattern Recognition, 138*, 109416. https://doi.org/10.1016/j.patcog.2023.109416

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European conference on computer vision, ECCV* (Vol. 8693, pp. 740–755).

Lin, W., Gao, J., Wang, Q., & Li, X. (2021). Learning to detect anomaly events in crowd scenes from synthetic data. *Neurocomputing, 436*, 248–259. https://doi.org/10.1016/j.neucom.2021.01.031

Linder, T., Pfeiffer, K. Y., Vaskevicius, N., Schirmer, R., & Arras, K. O. (2020). Accurate detection and 3d localization of humans using a novel yolo-based RGB-D fusion approach and synthetic training data. In *International conference on robotics and automation, ICRA* (pp. 1000–1006).

Liu, J. , Zhou, Y. , Sun, L., & Jiang, Z. (2019). Similarity preserved camera-to-camera GAN for person re-identification. In *ICME workshops* (pp. 531–536).

Liu, L., Ouyang, W., Wang, X., Fieguth, P. W., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection:

A survey. *International Journal of Computer Vision, 128*(2), 261–318. https://doi.org/10.1007/s11263-019-01247-4

Lu, C. , Shi, J., & Jia, J. (2013). Abnormal event detection at 150 FPS in MATLAB. In *International conference on computer vision, ICCV* (pp. 2720–2727).

Meharban, M., Sabu, M., & Krishnan, S. (2021). Introduction to medical image synthesis using deep learning: A review. In *International conference on advanced computing and communication systems, ICACCS, 2021* (pp. 414–419).

Mehran, R., Oyama, A., & Shah, M. (2009). Abnormal crowd behavior detection using social force model. In *International conference on computer vision and pattern recognition, CVPR* (pp. 935–942).

Milan, A., Leal-Taixé, L., Reid, I. D., Roth, S., & Schindler, K. (2016). MOT16: A benchmark for multi-object tracking. CoRRabs/1603.00831.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Survey, 5417*(1–7), 41. https://doi.org/10.1145/3425780

Mokhayeri, F., Granger, E., & Bilodeau, G. (2019). Domain-specific face synthesis for video face recognition from a single sample per person. *IEEE Transactions on Information Forensics and Security, 14*(3), 757–772. https://doi.org/10.1109/TIFS.2018.2866295

Mokhayeri, F., Kamali, K., & Granger, E. (2020). Cross-domain face synthesis using a controllable GAN. In *Winter conference on applications of computer vision, WACV* (pp. 241–249).

Montulet, R., & Briassouli, A. (2020). Densely annotated photorealistic virtual dataset generation for abnormal event detection. *ICPR Workshops and Challenges, 12664*, 5–19.

Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., & Zafeiriou, S. (2017). Agedb: The first manually collected, in-the-wild age database. In *International conference on computer vision and pattern recognition workshops, [CVPR]2017* (pp. 1997–2005). https://doi.org/10.1109/CVPRW.2017.250.

Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *Proceedings of the 38th international conference on machine learning, ICML* (Vol. 139, pp. 8162–8171).

Nikolenko, S. (2021). *Synthetic data for deep learning* (Vol. 174). Berlin: Springer.

Pratt, W., Faugeras, O., & Gagalowicz, A. (1978). Visual discrimination of stochastic texture fields. *IEEE Transactions on Systems, Man, and Cybernetics, 8*(11), 796–804. https://doi.org/10.1109/TSMC.1978.4309867

Qiu, H., Yu, B., Gong, D., Li, Z., Liu, W., & Tao, D. (2021). Synface: Face recognition with synthetic data. In *International conference on computer vision, ICCV* (pp. 10860–10870). https://doi.org/10.1109/ICCV48922.2021.01070

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th international conference on learning representations, ICLR*.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. CoRRabs/1804.02767.

Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems 28: Annual conference on neural information processing systems* (pp. 91–99).

Richter, S. R. , Hayder, Z., & Koltun, V. (2017). Playing for benchmarks. In *International conference on computer vision, ICCV* (pp. 2232–2241).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Conference on computer vision and pattern recognition, CVPR* (pp. 10674–10685).

Saez-Trigueros, D., Meng, L., & Hartnett, M. (2021). Generating photorealistic training data to improve face recognition accuracy. *Neural Networks, 134*, 86–94. https://doi.org/10.1016/j.neunet.2020.11.008

Saleh, S. A. M., Suandi, S. A., & Ibrahim, H. (2015). Recent survey on crowd density estimation and counting for visual surveillance. *Engineering Applications of Artificial Intelligence, 41*, 103–114. https://doi.org/10.1016/j.engappai.2015.01.007

Sam, D. B., Peri, S. V., Sundararaman, M. N., Kamath, A., & Radhakrishnan, V. B. (2020). Locate, size and count: Accurately resolving people in dense crowds via detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2020.2974830

Sánchez, F. L., Hupont, I., Tabik, S., & Herrera, F. (2020). Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. *Information Fusion, 64*, 318–335. https://doi.org/10.1016/j.inffus.2020.07.008

Schroder, G., Senst, T., Bochinski, E., & Sikora, T. (2019) Optical flow dataset and benchmark for visual crowd analysis. In *Proceedings AVSS*. https://doi.org/10.1109/AVSS.2018.8639113.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the international conference on computer vision and pattern recognition, CVPR07-12-June-2015* (pp. 815–823). https://doi.org/10.1109/CVPR.2015.7298682.

Sengupta, S., Chen, J. C., Castillo, C., Patel, V. M., Chellappa, R., & Jacobs, D. W. (2016). Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision, WACV 2016*. https://doi.org/10.1109/WACV.2016.7477558.

Shamsolmoali, P., Zareapoor, M., Granger, E., Zhou, H., Wang, R., Celebi, M. E., & Yang, J. (2021). Image synthesis with adversarial networks: A comprehensive survey and case studies. *Information Fusion, 72*, 126–146. https://doi.org/10.1016/j.inffus.2021.02.014

Shang, C. , Ai, H. , Zhuang, Z., & Chen, L. C. R. (2018). Improving pedestrian detection in crowds with synthetic occlusion images. In *ICME workshops* (pp. 1–4).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of international conference on learning representations, ICLR 2015*.

Sindagi, V., & Patel, V. M. (2017). A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters, 107*, 3–16. https://doi.org/10.1016/j.patrec.2017.07.007

Sindagi, V. A., & Patel, V. M. (2018). A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Lett., 107*, 3–16. https://doi.org/10.1016/j.patrec.2017.07.007

Sindagi, V. A., Yasarla, R., & Patel, V. M. (2022). JHU-CROWD++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(5), 2594–2609. https://doi.org/10.1109/TPAMI.2020.3035969

Sindagi, V. A., Yasarla, R., Sam, D. B., Babu, R. V., & Patel, V. M. (2020). Learning to count in the crowd from limited labeled data. In *European conference on computer vision, ECCV* (Vol. 12356, pp. 212–229).

Smeulders, A. W. M., Chu, D. M., Cucchiara, R., Calderara, S., Dehghan, A., & Shah, M. (2014). Visual tracking: An experimental survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(7), 1442–1468. https://doi.org/10.1109/TPAMI.2013.230

Sultana, M., Mahmood, A., & Jung, S. K. (2020). Unsupervised moving object detection in complex scenes using adversarial regularizations. *IEEE Transactions on Multimedia, 23*, 2005–2018. https://doi.org/10.1109/TMM.2020.3006419

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *International conference on computer vision and pattern recognition, CVPR* (pp. 6479–6488).

Sun, T., Segù, M., Postels, J., Wang, Y., Gool, L. V., Schiele, B., & Yu, F. (2022). SHIFT: A synthetic driving dataset for continuous multi-task domain adaptation. In *Conference on computer vision and pattern recognition, CVPR* (pp. 21339–21350).

Sun, X., & Zheng, L. (2019). Dissecting person re-identification from the viewpoint of viewpoint. In *International conference on computer vision and pattern recognition, CVPR* (pp. 608–617).

Sun, Z., Chen, J., Liang, C., Ruan, W., & Mukherjee, M. (2021). A survey of multiple pedestrian tracking based on tracking-by-detection framework. *IEEE Transactions on Circuits and Systems for Video Technology, 31*(5), 1819–1833. https://doi.org/10.1109/TCSVT.2020.3009717

Tian, J., Teng, Z., Zhang, B., Wang, Y., & Fan, J. (2021). Imitating targets from all sides: An unsupervised transfer learning method for person re-identification. *International Journal of Machine Learning and Cybernetics, 12*(8), 2281–2295. https://doi.org/10.1007/s13042-021-01308-6

Trabucco, B., Doherty, K., Gurinas, M., & Salakhutdinov, R. (2023). Effective data augmentation with diffusion models. CoRRabs/2302.07944. https://doi.org/10.48550/ARXIV.2302.07944.

Tripathi, G., Singh, K., & Vishwakarma, D. K. (2019). Convolutional neural networks for crowd behaviour analysis: A survey. *Visual Computing, 35*(5), 753–776. https://doi.org/10.1007/s00371-018-1499-5

Tripathi, S., Chandra, S., Agrawal, A., Tyagi, A., Rehg, J. M., & Chari, V. (2019). Learning to generate synthetic data via compositing. In *International conference on computer vision and pattern recognition, CVPR* (pp. 461–470).

Uner, O. C., Aslan, C., Ercan, B., Ates, T., Celikcan, U., Erdem, A., & Erdem, E. (2021). Synthetic18k: Learning better representations for person re-id and attribute recognition from 1.4 million synthetic images. *Signal Processing: Image Communication, 97*, 116335. https://doi.org/10.1016/j.image.2021.116335

Unity Technologies. (n.d.). Unity. https://unity.com/.

Verma, A., Subramanyam, A. V., Wang, Z., Satoh, S., & Shah, R. R. (2023). Unsupervised domain adaptation for person re-identification via individual-preserving and environmental-switching cyclic generation. *IEEE Transactions on Multimedia, 25*, 364–377. https://doi.org/10.1109/TMM.2021.3126404

Villamizar, M., Martínez-González, Á., Canévet, O., & Odobez, J. (2020). Watchnet++: Efficient and accurate depth-based network for detecting people attacks and intrusion. *Machine Vision and Applications, 31*(6), 41. https://doi.org/10.1007/s00138-020-01089-y

Wan, F., Wu, Y., Qian, X., Chen, Y., & Fu, Y. (2020). When person re-identification meets changing clothes. In *CVPR workshops* (pp. 3620–3628).

Wang, M., & Deng, W. (2021). Deep face recognition: A survey. *Neurocomputing, 429*, 215–244.

Wang, Q., Gao, J., Lin, W., & Li, X. (2021). Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(6), 2141–2149. https://doi.org/10.1109/TPAMI.2020.3013269

Wang, Q., Gao, J., Lin, W., & Yuan, Y. (2019). Learning from synthetic data for crowd counting in the wild. In *International conference on computer vision and pattern recognition, CVPR* (pp. 8198–8207).

Wang, Q., Gao, J., Lin, W., & Yuan, Y. (2021). Pixel-wise crowd understanding via synthetic data. *International Journal of Computer Vision, 129*(1), 225–245. https://doi.org/10.1007/s11263-020-01365-4

Wang, Y., Liang, X., & Liao, S. (2022). Cloning outfits from real-world images to 3d characters for generalizable person re-identification. In *Conference on computer vision and pattern recognition, CVPR* (pp. 4890–4899).

Wang, Y., Liao, S., & Shao, L. (2020). Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In *International conference on multimedia* (pp. 3422–3430).

Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer GAN to bridge domain gap for person re-identification. In *International conference on computer vision and pattern recognition, CVPR* (pp. 79–88).

Weyrauch, B., Heisele, B., Huang, J., & Blanz, V. (2004). Component-based face recognition with 3d morphable models. In *CVPR workshops* (p. 85).

Wood, E., Baltrusaitis, T., Hewitt, C., Dziadzio, S., Cashman, T. J., & Shotton, J. (2021). Fake it till you make it: Face analysis in the wild using synthetic data alone. In *International conference on computer vision, ICCV* (pp. 3661–3671).

Woodham, R., Catanzariti, E., & Mackworth, A. (1985). Analysis by synthesis in computational vision with application to remote sensing. *Computational Intelligence, 11*, 71–79. https://doi.org/10.1111/j.1467-8640.1985.tb00060.x

Wu, D., Zheng, S., Zhang, X. S., Yuan, C., Cheng, F., Zhao, Y., & Huang, D. (2019). Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing, 337*, 354–371. https://doi.org/10.1016/j.neucom.2019.01.079

Wu, Q., Dai, P., Chen, P., & Huang, Y. (2021). Deep adversarial data augmentation with attribute guided for person re-identification. *Signal, Image and Video Processing, 154*, 655–662. https://doi.org/10.1007/s11760-019-01523-3

Wu, W., Zhao, Y., Chen, H. , Gu, Y., Zhao, R., He, Y., & Shen, C. (2023). Datasetdm: Synthesizing data with perception annotations using diffusion models. CoRRabs/2308.06160. https://doi.org/10.48550/ARXIV.2308.06160.

Xiang, S., Fu, Y., You, G., & Liu, T. (2020). Unsupervised domain adaptation through synthesis for person re-identification. In *International conference on multimedia and expo, ICME* (pp. 1–6).

Xiang, S., Fu, Y., You, G., & Liu, T. (2021). Taking a closer look at synthesis: Fine-grained attribute analysis for person re-identification. In *International conference on acoustics, speech and signal processing, ICASSP* (pp. 3765–3769).

Xiang, S., Qian, D., Guan, M., Yan, B., Liu, T., Fu, Y., & You, G. (2023). Less is more: Learning from synthetic data with fine-grained attributes for person re-identification. *ACM Transactions on Multimedia Computing, Communications and Applications, 19*(5s), 173:1-173:20. https://doi.org/10.1145/3588441

Yaghoubi, E., Borza, D., Kumar, S. V. A., & Proença, H. (2021). Person re-identification: Implicitly defining the receptive fields of deep learning classification frameworks. *Pattern Recognition Letters, 145*, 23–29. https://doi.org/10.1016/j.patrec.2021.01.035

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. H. (2022). Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 446*, 2872–2893. https://doi.org/10.1109/TPAMI.2021.3054775

Yi, D. , Lei, Z. , Liao, S., & Li, S. Z. (2014). Learning face representation from scratch. CoRR.

Yuille, A. L., & Liu, C. (2021). Deep nets: What have they ever done for vision? *International Journal of Computer Vision, 129*(3), 781–802. https://doi.org/10.1007/s11263-020-01405-z

Zahra, A., Perwaiz, N., Shahzad, M., & Fraz, M. M. (2023). Person re-identification: A retrospective on domain specific open challenges and future trends. *Pattern Recognition, 142*, 109669. https://doi.org/10.1016/j.patcog.2023.109669

Zhang, C., Kang, K., Li, H., Wang, X., Xie, R., & Yang, X. (2016). Data-driven crowd understanding: A baseline for a large-scale crowd dataset. *IEEE Transactions on Multimedia, 18*(6), 1048–1061. https://doi.org/10.1109/TMM.2016.2542585

Zhang, Q., & Chan, A. B. (2019). Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In *International conference on computer vision and pattern recognition, CVPR* (pp. 8297–8306).

Zhang, Q. , Lin, W., & Chan, A. B. (2021). Cross-view cross-scene multi-view crowd counting. In *International conference on computer vision and pattern recognition, CVPR* (pp. 557–567).

Zhang, S., & Hu, H. (2023). Unsupervised person re-identification using unified domanial learning. *Neural Processing Letters*. https://doi.org/10.1007/s11063-023-11242-z

Zhang, T. , Xie, L. , Wei, L. , Zhuang, Z. , Zhang, Y. , Li, B., & Tian, Q. (2021). Unrealperson: An adaptive pipeline towards costless person re-identification. In *International conference on computer vision and pattern recognition, CVPR* (pp. 11506–11515).

Zhang, Y., Zhou, D., Chen, S., et al. (2016a). Single-image crowd counting via multi-column convolutional neural network. In *International conference on computer vision and pattern recognition, CVPR* (pp. 589–597).

Zhang, Y., Zhou, D., Chen, S., Gao, S., & Ma, Y. (2016b). Single-image crowd counting via multi-column convolutional neural network. In *International conference on computer vision and pattern recognition, CVPR* (pp. 589–597).

Zhao, Z., Han, T., Gao, J., Wang, Q., & Li, X. (2020). A flow base bi-path network for cross-scene video crowd understanding in aerial view. *ECCV Workshops, 12538*, 574–587.

Zheng, A., Chen, Z., Li, C., Tang, J., & Luo, B. (2021). Learning deep RGBT representations for robust person re-identification. *International Journal of Automation and Computing, 18*(3), 443–456. https://doi.org/10.1007/s11633-020-1262-z

Zheng, L., Shen, L., Tian, L. , Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *International conference on computer vision, ICCV* (pp. 1116–1124).

Zhou, R., Jiang, C., & Xu, Q. (2021). A survey on generative adversarial network-based text-to-image synthesis. *Neurocomputing, 451*, 316–336. https://doi.org/10.1016/j.neucom.2021.04.069

Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. CoRRabs/1904.07850.

Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International conference on computer vision, ICCV* (pp. 2242–2251).

Zurich, E. (n.d.). https://www.scalabel.ai/.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.