future internet

MDPI

Review

Survey on Videos Data Augmentation for Deep Learning Models

Nino Cauli * and Diego Reforgiato Recupero *

Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124 Cagliari, Italy
* Correspondence: nino.cauli@unica.it (N.C.); diego.reforgiato@unica.it (D.R.R.)

Abstract: In most Computer Vision applications, Deep Learning models achieve state-of-the-art performances. One drawback of Deep Learning is the large amount of data needed to train the models. Unfortunately, in many applications, data are difficult or expensive to collect. Data augmentation can alleviate the problem, generating new data from a smaller initial dataset. Geometric and color space image augmentation methods can increase accuracy of Deep Learning models but are often not enough. More advanced solutions are Domain Randomization methods or the use of simulation to artificially generate the missing data. Data augmentation algorithms are usually specifically designed for single images. Most recently, Deep Learning models have been applied to the analysis of video sequences. The aim of this paper is to perform an exhaustive study of the novel techniques of video data augmentation for Deep Learning models and to point out the future directions of the research on this topic.

Keywords: data augmentation; deep learning; domain randomization; simulation

Citation: Cauli, N.; Reforgiato Recupero, D. Survey on Videos Data Augmentation for Deep Learning Models. *Future Internet* **2022**, *14*, 93. https://doi.org/10.3390/fi14030093

Academic Editors: Vijayakumar Varadarajan, Rajanikanth Aluvalu and Ketan Kotecha

Received: 22 February 2022
Accepted: 13 March 2022
Published: 16 March 2022

1. Introduction

We live in a world where most of our actions are constantly captured by cameras. Video cameras are spread almost everywhere: in smartphones, computers, drones, surveillance systems, cars, robots, intercoms, etc. Image Processing (IP) and Computer Vision (CV) models, able to extract and analyse information from images, are becoming more and more important. With the advent of Deep Learning (DL) and the increase in computational power, classical CV algorithms are quickly being replaced by Convolutional Neural Networks (CNN) or other DL models [1,2]. Typically, DL models possess a huge number of parameters that need to be trained. The risk of overfitting with such big models is very high and big datasets with high variability are needed for networks to be able to generalise.

Unfortunately, collecting a big collection of images or videos and labellng them is both resource and time consuming, and, in some cases, even impossible. In medical image analysis, data such as computerized tomography (CT) and magnetic resonance imaging (MRI) scans are expensive and time consuming to collect. Moreover, medical data are protected by strict privacy protocols, making it difficult to obtain past recordings from hospitals. In robotics, a prolonged operation of robots for collecting data can result in the wearing or damaging of components, labour intensive procedures and dangerous interactions between machines and operators. Collecting data for autonomous vehicles control have similar problems. Data collection in this case consists of running a vehicle (car, drone, boat) with a camera mounted on top in various environmental conditions (weather, time of the day, city versus countryside, etc.). This process can take a conspicuous amount of time, it is expensive, the vehicle can be damaged and special permissions to operate in restricted areas are often needed. From these examples, it is clear how data collection can become a complex and troublesome process, but it is only part of the problem. In order to generate a dataset for supervised learning models, data need to be labelled. In many occasions, the labelling process cannot be automatised, and each image needs to be labelled manually by humans (e.g., medical images segmentation).

The consequence of the aforementioned problems in data collection and labelling is the generation of small and unbalanced datasets. Several techniques exist to tone down this problem, reducing the overfitting and improving the generalisation capabilities of the models. For some problems like object recognition, face recognition and autonomous driving, big generic and public datasets have already been collected [3–6]. Pretraining is a technique where models are first trained on big existing datasets built for more generic tasks. In this way, pretrained models can learn a base knowledge to be transferred to a specific problem. A pretrained model is able to converge faster on a new dataset, needing less data [7]. A similar approach is Transfer Learning: models pretrained on a dataset for a specific data distribution are able to transfer part of the acquired knowledge to a different distribution with small or no fine-tuning. Data regularization techniques (Dropout and Batch normalization) are other approaches to reduce overfitting. Using a combination of these techniques, tasks where data are scarce can be more easily handled by DL models.

However, none of the previous methods directly solve the problems of shortage of data and unbalanced datasets. Data augmentation techniques, on the other hand, address the lack of data artificially generating new ones. The most basic technique of data augmentation for image analysis is noise injection: the dataset is expanded creating duplicates of the original images injected of random values in the RGB space. Since the introduction of AlexNet in 2012 [8], geometric and color space transformations are common data augmentation techniques used to improve the performance of DL models for image analysis. Cropping, flipping, rotating, translating, histogram and RGB values alteration all fall in this category.

With the improvements in Neural Networks (NN) and DL, more advanced data augmentation methods increased. Strategies based on generative modeling are able to generate new input images belonging to a similar distribution of the original dataset. These strategies use Generative adversarial networks (GANs) to generate the new images [9]. A GAN consists of two networks, a generator and a discriminator that compete against each other during training: the generator tries to produce an image belonging to a distribution of interest from input images, while the discriminator tries to distinguish generated images from the ones belonging to the true data distribution. After training, the generator can be used to augment the original dataset with newly generated images from the same distribution of the original dataset. Neural Style Transfer is another DL based methodology able to augment the size of image datasets. The idea is to alter the latent space of an Encoder/Decoder CNN in order to generate images with different styles. The output image of the Decoder is similar to the input one but with a difference in style that depends on the changes applied to the latent layer. In [10], the authors propose a GAN architecture, based on Neural Style Transfer, able to generate photo realistic images that can be used to create synthetic datasets.

We already mentioned several data augmentation techniques to improve training and reduce overfitting of image processing DL models. A common trait of all these methodologies is the use of images from the original dataset as a base for generating the new images. A different approach is to generate the images for the augmented dataset from physical models that approximate the world. In this case, detailed models of the environment, the physics and the cameras are defined by the researcher and used to generate synthetic approximation of real images. The researchers are able to randomise the scene and create varied simulated images tuning the models' parameters. This ability to tune the models offers a higher flexibility compared to other data augmentation techniques. Graphical and physical simulators are powerful tools and several researchers are using them for the generation of artificial datasets [11]. This process has been accelerated by the advancement in the video games industry. Modern game engines (Unity [12] and Unreal Engine [13] among others) are able to render in a few milliseconds photo-realistic images, simulate realistic physical interaction between objects, and they offer powerful scripting and designing tools to recreate detailed artificial scenes. Exploiting game engines, researchers can create varied and faithful to reality synthetic datasets to train large DL

models for image analysis. Tremblay et al. [11] used Unreal Engine to create a photo realistic synthetic dataset for training an object detection model. The dataset was generated randomising objects position, scene, illumination and camera position. However, photo realism is not always the way to go to create a synthetic dataset. Domain Randomization (DR) is a method to generate synthetic data in simulation [14]. If the variability in the generated dataset is significantly high, models trained on it will be able to generalise to the real world. In order to achieve this result, the scene parameters need to be randomized to an extent that the generated images can fully cover the desired data distribution. In this way, to overtake the reality gap (intrinsic difference between simulated and real worlds), simulated datasets are not generated to be as realistic as possible, but to contain a high variability in lighting, object shapes, textures, camera position, and physics behaviours.

Video analysis adds the temporal dimension to the images problem, resulting in a very complex challenge. With the introduction of industry 4.0, robotics and autonomous vehicles, video analysis is becoming a focal problem for the research community. In this case, the input of the DL models is not single images but streams of multiple images with temporal and spatial correlation between each others. While some of the models meant for image analysis can be used out of the box to analyse videos, usually some changes have to be done to take into account the temporal dimension. Optical flow [15], 3D convolutions [16] and Recurrent Neural Networks (RNN) [17,18] are the most common methods used to handle image sequences. However, the correlation in time and space in between images of the same sequence needs to be taken into account not only in the design of the DL models, but also in the design of the datasets. Geometric and color space transformation can usually be applied to videos keeping them constant for the entire image sequence, but, for more complex methods, the changes need to be more significant. In generative modeling, the Generator network needs to keep some information of the past frames. The DL models used to analyse image sequences (Optical flow, 3D convolutions and RNN) are a proper solution. In simulation, the physical interaction between objects needs to be taken into account. If the focus is in human action recognition or prediction, the skeletal animation of the subjects is needed to simulate the motion. In domain randomization methods, camera motion must be taken into account and the variation in textures, illumination and objects shapes must be constant or coherent through the entire video sequence.

Reviews on image data augmentation for DL models have already been published [19–23]. In their paper, Shorten et al. [19] realised a complete survey on image data augmentation for DL, covering both basic image manipulation and DL approaches. The basic manipulation approaches are composed of geometrical and color space transformations, kernel filters, noise injection, mixing images and random erasing. In their review, the authors focus also on more recent DL approaches: feature space augmentation, adversarial training, GAN-based and Neural Style Transfer. More recently, Khalifa et al. [20] grouped the papers of their review on image data augmentation in a similar fashion. In addition, the authors present an analysis of the state of the art specific to different application domains. On the contrary, Wang et al. [21] focused their review on the more specific problem of face data augmentation. The authors reviewed the state of the art on face data augmentation under several points of view: transformation types, transformation methods, and evaluation metrics. In face data augmentation, likewise for general image data augmentation, generative models have recently been the primary choice, replacing or enhancing most of the other methods. A different application area was tackled by Chlap et al. [22] in their review of medical image data augmentation. The paper analyses the state of the art of CT and MRI image data augmentation for DL applications. The reviewed works were divided into three groups. The first group are basic augmentation techniques that correspond to geometric and color space transformations. The second group are deformable augmentation techniques, consisting of scaling and warping masks applied to the original images. The last group are the DL augmentation techniques, further divided into GAN-based and others. The author pointed out that basic and deformable methods are still more popular being easy to apply on medical data and due to the availability of

several software libraries for their implementation. On the other hand, DL methods are evolving rapidly and will gain more popularity given their ability to tackle complex data synthesis problems such as cross-domain image synthesis. Image mixing and deleting data augmentation strategies are reviewed by Humza Naveed in his survey [23]. The reviewed papers are split into three categories: erasing image patches; cut image regions and replacing them with patches from other images; and mixing multiple images. The author compared the performances of the reviewed methods for image classification, object detection, and fine-grained image recognition on publicly available datasets. Unfortunately, as far as our knowledge goes, an exhaustive review on data augmentation for DL based video analysis is still missing. The goal of this paper is to fill this gap. In Section 2, the methodology used to review the state of the art on video data augmentation is presented. This section also contains a statistical analysis of the literature selected for reviewing. In Section 3, the papers selected are presented based on the data augmentation method they use. Lastly, future research challenges and directions are discussed in Section 4.

2. Methods Used and Overview of the Literature

The focus of this review paper is to comprehensively cover the scientific publications on video data augmentation for DL models. We are going to center our attention specifically on data augmentation methods designed to handle the temporal dimension of video streams. Nowadays, several search engines for academic publications are available. We decided to use Scopus [24] to examine the scientific literature and select the best set of papers published in recent years on video data augmentation. Scopus has been chosen based on the size and the quality of its scholarly literature database. The final set of selected publications was obtained after few iterative refinement steps:

1. An initial search was performed, resulting in a collection of 570 publications. The criteria used to select a paper were the following:
   (a) title, abstract or main text must contain the set of words (“video” “data augmentation”) or (“video” “synthetic” “data” “generation”) or (“video” “simulation” “data” “generation”)
   (b) papers must be published from 2012 to 2022;
   (c) papers must be written in English;
   (d) book chapters were excluded.
2. Duplicated entries and papers with the titles and abstracts not relevant with the topic were removed, resulting in a pruned set of 76 papers.
3. The full text of the remaining 76 paper was evaluated. Several of the papers applied standard image data augmentation strategies without focusing specifically on the problem of video data augmentation. For this reason, only 33 papers were finally selected.
4. The set of 33 papers was extended with two more papers which we felt had an impact on the survey. The final number of paper selected for the review is 35.

Analysing more in details the search results, we can point out some interesting findings. Figure 1 depicts a bar chart of the number of publication for each year. We have not found any relevant publication in the interval between 2012 and 2015; for this reason, only the years from 2016 to 2022 are shown. It is possible to notice an exponential growth in the number of papers that introduce data augmentation strategies tailored for video sequences. This is a clear sign on how this topic is significant and is attracting the interest of the research community. In addition, 77% of the papers selected for this review has been published after 2019, demonstrating the novelty of this research area.

Some insight can be also drawn by the analysis of the most frequently occurring keywords (Figure 2). Besides the obvious high number of occurrences of keywords like “Data Augmentation”, “Deep Learning” and “Computer Vision”, it is worth noticing the presence in the plot of application areas highly dependent on time (e.g., “Action Recognition” and “Human-action Recognition”). Data augmentation techniques specifically

tailored for video sequences are particularly interesting in applications dealing with subjects in motions and scenarios dynamically changing through time.

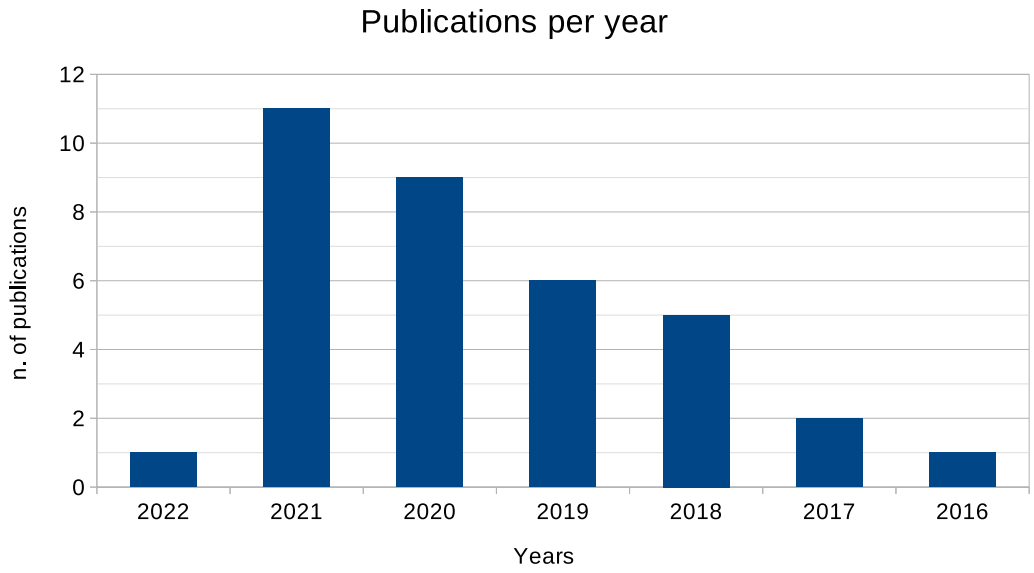

Figure 1. Number of the selected publications per year in the interval between 2016–2022. No relevant publications on video data augmentation has been found between 2012–2015.

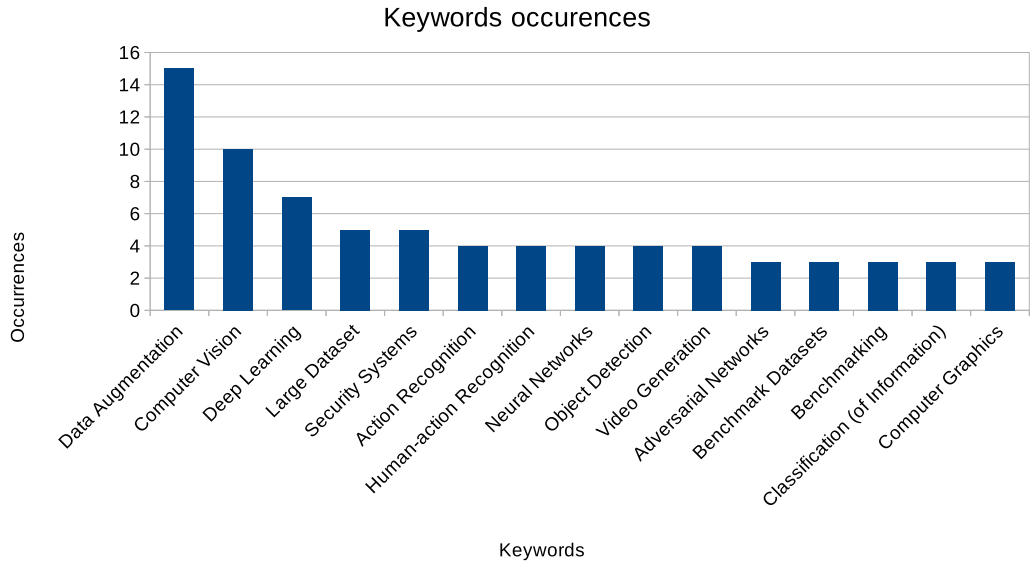

Figure 2. Number of occurrences of the 15 most frequently occurring keywords in the selected papers.

3. Review on Video Data Augmentation

In this section, we are going to review and analyse the 35 papers selected from the literature. The papers in this collection range from straightforward time domain extensions of geometric and color space transformations, to complex DL generative models able to render realistic synthetic videos. Some of the works focus primarily on data augmentation; for others, data augmentation is only a block of a bigger system, while some others present architectures that are not primarily intended for data augmentation, even if they can be used to expand or create a video dataset. For all the reviewed papers, we will analyse exclusively their application on video data augmentation. We decided to organise the review grouping the papers according to the methodology they use. We define five classes of methodologies for video data augmentation: basic transformations (geometric, color

space, temporal, erasing and mixing), feature space augmentation, DL models, simulation, and methods that improve data generated though simulation using Generative Adversarial Networks. Tables 1 and 2 list all the 35 papers with their application area, data augmentation methodology, and DL model used for the testing and dataset used for data augmentation or comparison. The next subsections will analyse each of these classes.

Table 1. List of the reviewed papers part 1. The first 20 papers ordered by year of publication. Please read the analysis of the papers in Section 3 for a better understanding of the fields.

Paper	Application Area	Data Augmentation Method	Model Tested	Dataset
Charalambous et al., 2016 [25]	Gait recognition	Simulated avatars animated with mocap data	SVM	Self collected
Wang et al., 2017 [26]	Action recognition	Temporal cropping	Three-stream CNN	UCF101, HMDB51, Hollywood2, Youtube
De Souza et al., 2017 [27]	Action recognition	Simulated scene (Unity)	TSN	UCF-101, HMDB-51
Wang et al., 2018 [28]	Salient regions detection	Optical flow warping	Encoder/Decoder CNNs	FBMS, DAVIS
Griffith et al., 2018 [29]	Aerial surveillance	simulated wide area aerial imagery	-	-
Lu et al., 2018 [30]	Synthesis of Shaking Videos	Dynamic 3D scene modeling	-	-
Dong et al., 2018 [31]	Video recommendation	Feature space	InceptionV3	Hulu Challenge 2018
Angus et al., 2018 [32]	Road-scene synthetic annotation	Simulated road scenes	FCN, SegNet	URSAS (generated by the authors), CityScapes, PFB, Synthia
Aberman et al., 2019 [33]	Video-based cloning	Double branches GAN	-	-
Rimboux et al., 2019 [34]	Pedestrian detection	Background subtraction + 3D synthetic models of persons	ResNet-101, RPN+	Town Centre
Fonder et al., 2019 [35]	Drone video analysis	Simulated aerial scenes (Unreal Engine and AirSim)	-	-
Wu et al., 2019 [36]	Action recognition	GAN for dynamic image generation	2D and 3D CNNs	UCF101, KTH
Sakkos et al., 2019 [37]	Background subtraction	Changes in illumination	Encoder/Decoder CNN	SABS
Li et al., 2019 [38]	Hand gesture recognition	Temporal cropping	mdCNN	VIVA
Sakkos et al., 2020 [39]	Background subtraction	Changes in illumination	Encoder/Decoder CNN	SABS
Kwon et al., 2020 [40]	Novel view synthesis of human performance videos	Two-tower siamese encoder/decoder	-	MVHA, PVHM, ShapeNet
Chai et al., 2020 [41]	Crowd video generation	CrowdGAN	MCNN, CSRNet, SANet, CAN	Mall and FDST

Table 1. *Cont.*

Paper	Application Area	Data Augmentation Method	Model Tested	Dataset
De Souza et al., 2020 [42]	Action recognition	Simulated scene (Unity)	TSN	UCF-101, HMDB-51
Namitha et al., 2020 [43]	Video surveillance	Coloured boxes superposition	-	Sherbrooke, i-Lids, M-30, Car
Isobe et al., 2020 [44]	Person re-identification	Random cropping, flipping and erasing	Swallow network	ILIDSVID, PRID 2011, MARS

Table 2. List of the reviewed papers part 2. The last 15 papers ordered by year of publication. Please read the analysis of the papers in Section 3 for a better understanding of the fields.

Paper	Application Area	Data Augmentation Method	Model Tested	Dataset
Zhang et al., 2020 [45]	Video data augmentation	WGAN for dynamic image generation	2D and 3D CNNs	LS-HMDB4
Yun et al., 2020 [46]	Action recognition	Image mixing	SlowFast-50	Mini-Kinetics
Ye et al., 2020 [47]	Semantic segmentation	Simulated fisheye model	SwiftNet-18	CityScapes
Wang et al., 2021 [48]	Crowd understanding	Simulation + GAN	Several CNNs	ShanghaiTech A/B,UCF-CC-50, UCF-QNRF, WorldExpo'10, CityScapes
Hwang et al., 2021 [49]	Action recognition	Simulated scene and avatar (Unreal Engine)	Glimpse, ST-GCN, VA-CNN	ETRI-Activity3D, NTU RGB+D 120
Tsou et al., 2021 [50]	Photoplethysmography Estimation	Encoder/Decoder deep networks	rPPG network (3DCNN)	PURE, UBFC-RPPG
Wei et al., 2021 [51]	Human video motion transfer	GAN	-	Self collected, iPER dataset
Chen et al., 2021 [52]	Self driving	Video images + 3D car models	PSPNet, DeepLabv3	UrbanData (Collected by the authors)
Dong et al., 2021 [53]	Video relevance prediction	Feature space	InceptionV3	Hulu Challenge 2018
Hu et al., 2021 [54]	Object detection	Background extraction and geometrical transformations	ResNet-18, ResNet-50	LaSOT, GOT-10k, TrackingNet, OTB-100, UAV123
Kerim et al., 2021 [55]	Person tracking	Simulated scene (NOVA engine)	DiMP, ATOM, KYS, PrDiM	PTAW172Real, PTAW217Synt (Collected and generated by the authors)
Varol et al., 2021 [56]	Action recognition	Synthetic 3D human avatars generation	3D ResNet-50	RGB+D Dataset
Hu et al., 2021 [57]	Object tracking	Simulated scene (GTAV)	Pixel2Mesh, Pix2Vox, MeshR-CNN, Video2Mesh	SAIL-VOS 3D (proposed synthetic dataset), Pix3D.
Bongini et al., 2021 [58]	Object detection	-	YOLOv3	FLIR ADAS
Otberdout et al., 2022 [59]	Facial expression generation	MotionGAN and TextureGAN	LSTM	CASIA

3.1. Basic Transformations

A simple technique for temporal data augmentation in videos was proposed in [26]. The paper focuses on the problem of action recognition from videos. The authors augment the training set for their model applying iteratively a temporal cropping several times to each original video sequence. They temporally sub-sampled each video sequence of length l with a stride s, obtaining s new sequences of length l/s. A three-stream CNN was trained with and without data augmentation. The accuracy of both networks was evaluated on four different datasets: UCF101, HMDB51, Hollywood2 and Youtube. The network trained with data augmentation improved the accuracy on all the datasets (+1.3% on UCF101, +1.1% on HMDB51, +1.2% on Hollywood2 and +2.5% on Youtube). Data augmentation using temporal cropping is proposed also by Lee et al. [38]. The authors augment a video dataset of hand gestures splitting the original 12 frames videos in 3 videos of 8 frames each (1st to 8th, 3rd to 10th and the 5th to 12th frame). They also invert the temporal order of the frames obtaining an augmented dataset six times larger than the original. The proposed data augmentation strategy was used to augment the VIVA dataset. Their mdCNN trained on the augmented dataset improved the accuracy of 6% over the same network trained without data augmentation.

Applying commonly used image-level data augmentation strategies to video sequences may introduce unnecessary noise corrupting the temporal cues of intra-clip frames. In [44], the authors solve the problem applying the same transformation to all the frames of a mini-batch clip instead of randomly changing it for each frame. Random cropping, flipping and erasing are used to augment a video dataset for person re-identification.

Image mixing techniques (e.g., Mixup [60] and CutMix [61]) have been widely used for image data augmentation. These types of approaches generate the augmented images mixing the pixel values from two different images of the original dataset. Some algorithms, (i.e., Mixup), averages the RGB values of the two images, while methods like CutMix replace randomly shaped patches of one image with the other. In order to extend image mixing techniques to video data augmentation, temporal cues in between frames must be taken into account. VideoMix [46] is a data augmentation method proposed by Yun et al. that extends CutMix to video data augmentation. The temporal consistency is preserved keeping the patch size and position the same for all the frames of each video clip. The authors tested VideoMix on three tasks (action recognition, localization and detection) training different 3D CNNs. They compared the performances of their algorithm against the vanilla CutMix method. After training the SlowFast-50 network on the Mini-Kinetics dataset, VideoMix achieved the best improvement in accuracy (+2.4%) for action recognition.

A different approach to generate synthetic video is warping some key frames with the use of optical flow fields. This technique is proposed by [28] to augment a dataset for salient regions detection in videos. Starting from video frames and their saliency masks, the authors generate a synthetic optical flow field. They then use these optical flow fields to warp the original frames in order to generate new synthetic data. The proposed data augmentation method was tested on FBMS and DAVIS datasets. To show the effectiveness of data aumentation, three versions of training sets were created: only synthetic images, mixed real and synthetic images, and only real images. After training on the three training sets, the saliency maps mean absolute error (MAE) was calculated for each version. Using only synthetic data resulted in a small increase in MAE compared to the mixed training set (7.65 $\rightarrow$ 9.27 on FBMS, 6.36 $\rightarrow$ 7.53 on DAVIS), while the real images training set suffered from severe overfitting.

Sakkos et al. [37,39] achieved data augmentation through changes in illumination of video datasets for background subtraction. New synthetic images are generated applying local and global illumination masks to the original frames that simulate dynamical changes in illumination (spot light switch, global darkening and brightening, etc.). The data augmentation method was tested on the Stuttgart Artificial Background Subtraction dataset (SABS). Even if the dynamic light models were simplistic, an Encoder/Decoder CNN for

background extraction improved its intersection over union (IoU)trained on the augmented dataset rather than the original one: +25% (0.3594 → 0.6161) on IoU.

In video synopsis applications, motion information is more important than video fidelity. Namitha et al. [43] proposed a toolbox for data augmentation able to generate synthetic surveillance videos of static cameras for video synopsis analysis. The synthetic videos are composed superimposing to an extracted background a series of coloured rectangular boxes that represent moving objects or persons. The toolbox permits to choose number, size, trajectory and speed of the boxes added to the synthetic video. In order to test the efficiency of their data augmentation method, the authors compared real camera footage from different real-world video datasets to their synthetic counterparts. When evaluated on frame compact ratio (CR), total true collision area (TCA) and total false overlapping area (FOA) metrics, the results obtained by both real-world and synthetic data were close, demonstrating the validity of the data augmentation method. In their paper, Hu et al. [54] introduced AMMC (Augmentation by Mimicking Motion Change), a data augmentation strategy for object tracking that takes into consideration tracking motion features. AMMC first separates the target and background from the images. The cropped target images are transformed with operations like rotation, projection, resizing, blurring, and occlusion that reflect motion changes. The augmented target images are then superimposed on the background images at a random position in order to obtain new synthetic data. The authors trained ATOM and DiMP trackers on their simulated dataset, and they perform comprehensive experiments on five popular tracking benchmarks: LaSOT, GOT-10k, TrackingNet, OTB-100 and UAV123.

In several video analysis applications (e.g., autonomous driving or surveillance), a wider field of view gives more information about the surrounding environment. For this reason, fisheye cameras are often used to capture the videos to analyse. Unfortunately, due to the high level of distortion of fisheye camera images, common data augmentation methods based on geometrical transformations cannot be directly applied. Ye et al. [47] proposed a data augmentation method able to generate synthetic fisheye images from rectilinear ones. The authors defined a geometrical projection model that simulates a fisheye camera. The model possesses seven degrees of freedom (DoF): six DoFs are the relative rotation (3 DoFs) and translation (3 DoFs) between the world coordinate system and the camera coordinate system, while the last one is the focal length. The augmented data are generated randomising each DoF in order to vary translations, rotations, zooming and distortion of the simulated camera. Through a detailed ablation study, the paper evaluated the importance of each DoF calculating the mean intersection over union (mIoU) for semantic segmentation. Focal length is the most relevant, but using a 7-DoF model gives the best mIoU. However, for tests of the augmented dataset on real fisheye images, the paper gives only qualitative results.

3.2. Feature Space

DL models often extract a one-dimensional, feature vector from the input images. Sometimes, it is more convenient to perform the data augmentation on the feature space instead than on the image space (lack of availability of the original videos due to privacy constraints, ad hoc organization of the feature space, etc.). In their works, Dong et al. [31,53] proposed a data augmentation strategy for a content-based video recommendation challenge. The authors did not have access to the RGB video frames and applied the data augmentation directly on the feature vector extracted from an InceptionV3 deep network. They propose a data augmentation technique similar to the one used by Wang et al. [26] for video action recognition. Their frame-level data augmentation sub-samples each feature sequence skipping frames with a stride s. Repeating the process starting from a different frame of the original feature sequence, they are able to generate s distinct new sequences. The authors compared the performance metric scores (recall/hit scores) of the network trained with and without data augmentation on the Hulu Content-based Video Relevance Prediction Challenge 2018. In the most recent work, the network trained with data augmen-

tation achieved an improvement of the performances both for TV-Shows (2.708 → 3.092) and Movies datasets (2.030 → 2.289).

3.3. Dl Models

In [33], the authors used a double branches GAN to generate synthetic videos of a subject performing new dancing moves. The network is trained on a specific person, and it is able to generate images of that person performing moves acted by different subjects. Even if it is possible to use this system to augment an existing video dataset, one drawback is that the network needs to be retrained every time that the principal actor changes. Each synthetic frame can be generated at 12.5 fps using an NVIDIA GeForce GTX Titan Xp GPU (12 GB).

A GAN is also used by [36] to augment video datasets for action recognition. For each video sequence representing an action, the generator outputs a single frame that encodes all the information regarding motion features. The generated frames and original datasets are then joined together to obtain the augmented training set. The GAN features generator can enlarge the differences between similar classes. The data augmentation model was tested on UCF101 and KTH action recognition datasets. A 2DCNN and 3DCNN were trained with and without data augmentation, with the data augmentation networks obtaining an increase in accuracy on both datasets with respect to the one trained on the original ones: 2DCNN +35% on KTH and +26% on UCF101, 3DCNN +37% on KTH and +21% on UCF101. The work of Zhang et al. [45] shares the idea of utilising a GAN to generate dynamic images compressing the motion information of video sequences. The authors propose a data augmentation framework that generates new synthetic dynamic images from videos using a WGAN. The augmented dataset of real and synthetic dynamic images can be used to train video classification models. In an ablation study, the proposed method was compared with other two data augmentation strategies: corner cropping with scale jittering (CCS) and horizontal flipping (HF). The three data augmentation strategies were applied separately to the LS-HMDB4 dataset and the proposed method obtained the better accuracy of 71.01% (70.28% for CCS and 69.69% for HF).

More recently, Wei et al. [51] presented a novel GAN based model for Appearance-Controllable Human Video Motion Transfer. The GAN model is able to generate a novel video from a source motion video and multiple target appearance videos. The innovation of their technique is the ability to control the appearance of the subject and the background in the generated synthetic videos without any retraining of the model. To achieve this result, the input are first preprocessed, extracting the skeletal poses sequence from the source motion video together with the appearance of face, upper garment and lower garment from the target appearance videos. Using the preprocessed inputs, a GAN generates a synthetic video of a new subject performing the source action. This video is then superimposed to a selected background to generate the final video sequence.

Kwon et al. [40] designed and implemented a two-tower siamese encoder/decoder network able to synthesise, from videos of a human performing the same action taken from multiple reference viewpoints, novel videos of the same scene from different viewpoints. The network first encodes the 2D videos in a 3D volumetric latent layer. The volumetric latent layer is then used to generate the synthetic new viewpoint videos.

Instead of focusing on a single subject, CrowdGAN [41] is a DL model able to recursively generate synthetic crowd videos starting from few initial context frames. The model contains two modules, one directly predicts the next frame while the other predicts the optical flow map used to warp the pixels of the starting frame. The output of the two modules is fused together to obtain the next frame. Iterating the process, using the output frames as next inputs, CrowdGAN can generate longer realistic video sequences for crowd motion analysis. In order to evaluate the applicability of the model for data augmentation, the authors tested it on two datasets for crowd counting: Mall and FDST. Several state-of-the-art counting methods (i.e., MCNN, CSRNet, SANet and CAN) were

trained with and without augmenting the datasets. The results of all the tests on data augmentation outperformed the ones on real data.

Tsou et al. [50] focused their attention on the specific problem of Remote Photoplethysmography Estimation (rPPG), in order to detect blood volume changes from videos of human faces. They propose a data augmentation methodology to generate synthetic videos of the face of a subject that represents a target rPPG signal. The facial synthetic videos are generated by two Encoder/Decoder deep networks. The inputs needed by the models are images of the face of the subject and the target rPPG signal. The authors evaluate their model on PURE and UBFC-RPPG datasets over several metrics (e.g., Pearson correlation coefficient (R), Mean absolute error (MAE), and Root mean square error (RMSE)). They trained on three different data: the source data, the data augmented with traditional methods (i.e., random rotation, brightness, and saturation), and the data augmented with the proposed method. The model trained using the proposed method for data augmentation improved the performances on all the metrics. Lastly, the authors of [59] presented a model able to generate synthetic images of facial expressions. The model is composed of two GANs: a MotionGAN able to generate a series of synthetic facial landmark sequences that represent facial expressions; and a TextureGAN that generates novel facial expressions videos from the generated landmark sequences together with a source natural face image. The performances of MotionGAN in data augmentation were tested on the CASIA dataset, training an LSTM network to recognise facial expressions. The LSTM trained on the original training set achieved an accuracy of 87.5%, while the same network trained on an augmented dataset using MotionGAN achieved 92.7%.

3.4. Simulation

The great success of the video game industry is leading to an exponential improvement of graphic cards and real-time rendering systems. Several graphic and physic engines exist that are able to render photo realistic scenes at high frame rates. Game engines like Unreal Engine [13] and Unity [12] not only produce high quality synthetic videos, but they also come with a powerful, programmable and user friendly interface, making them the perfect tool to generate augmented simulated datasets. In robotics, simulators are often used to test and train the control models and 3D robotic simulator, which have existed for more than two decades. As far as DL model training is concerned, Reinforcement Learning (RL) agents have often been trained in simulations, due to their need to continuously explore the environment that surrounds them [62].

One of the first attempts to generate a video simulated dataset for gait recognition was made by Charalambous et al. in 2016 [25]. The authors used Vicon's motion capture data extracted from recordings of humans walking and running on a treadmill. The Vicon data were then imported into Blender [63] and attached to randomly generated avatars (with differences in age, sex, weight, etc.). Using Blender, it was possible to automatically label the data. Compared to a more recent simulated dataset, the images were quite simplistic, with a single avatar centered in the frame and with a plain grey background. De Souza et al. [27] made a step ahead generating a diverse, realistic, and physically plausible dataset of human action videos, called PHAV. The authors used Unity to render the videos, and they were able to randomise the scene based on different parameters and preset assets (environment, camera position, weather, lighting, time of the day, number of actors). The approach is not limited to existing motion capture sequences, but it procedurally defines synthetic actions via a combination of atomic motions. In their follow up paper [42], the authors improve and deeper describe the generative 3D model and the procedural algorithm to randomise the scene and generate the actions. The improved framework is also able to generate multiple sensor modalities like semantic segmentation and optical flow. The proposed parametric simulation tool is able to generate fully annotated action videos at 3.6 FPS using one consumer-grade gaming GPU (NVIDIA GTX 1070). The authors tested data augmentation performances of the model on two main stream action recognition datasets: UCF-101 and HMDB-51. A Temporal Segment Network (TSN) was trained with and

without data augmentation, with the former (named CoolTSN) obtaining higher accuracy on both datasets: TSN on UCF-101 93.6%; CoolTSN on UCF-101 94.2%; TSN on HMDB-51 66.6%; and CoolTSN on HMDB-51 69.5%.

ElderSim, a synthetic data generation platform for human action recognition was implemented by Hwang et al. [49]. The authors used Autodesk Maya and Unreal Engine to model and simulate virtual scenes for eldercare applications. The realistic elders models are placed inside detailed 3D house environments and they are animated using motion capture data. The platform presents a user interface for generating synthetic video sequences with randomised camera viewpoints, illumination, objects, and avatars' appearance and movements. The data augmentation tool was tested on two elder activity recognition datasets (ETRI-Activity3D and NTU RGB+D 120) training three different action recognition models (Glimpse, ST-GCN, VA-CNN). The recognition networks were trained both on the original datasets and on the same datasets augmented with synthetically generated videos. In all the tests performed, the networks trained on augmented datasets improved the accuracy, reaching an increase of +16.39% on a cross-dataset test using the Glimpse model. In a similar fashion, Hu et al. [57] use the open world game Grand Theft Auto V (GTAV) to create a synthetic dataset for object selection and 3D mesh reconstruction from videos named SAIL-VOS 3D. The dataset provides video frames, camera matrices, depth data, instance level segmentation, instance level amodal segmentation and the corresponding 3D object shapes. No quantitative experiments on real-world datasets were performed. Kerim et al. [55], on the other hand, created their own rendering engine (NOVA) based on Unity to allow researchers with no experience in computer graphics to generate high quality datasets with accurate and dense annotations. The authors collected a real-world dataset named PTAW172Real, and they used NOVA to generate a synthetic one called PTAW217Synt. They tested several state-of-the-art person trackers (DiMP, ATOM, KYS and PrDiM) training them on both PTAW172Real and PTAW217Synt. Overall, the models trained on the synthetic dataset obtained better IoU results compared to the same models trained on PTAW172Real.

Surveillance video analysis is another area of application of simulated video data augmentation. In this case, the synthetic data need to be generated based on an accurate crowd and traffic simulation. The area to be watched can be wide, as in aerial surveillance imagery. Virtual flight simulators can come in handy to render photo realistic videos of wide areas. Griffith et al. [29] proposed a system for the generation of synthetic wide area aerial surveillance imagery. The authors simulated the traffic in an urban environment using Matlab and a traffic simulator (SUMO). The 3D models of city buildings were extruded from Open Street Map data and exported, together with the traffic simulations, to X-Plane flight simulator, which has been used as the main visualization tool. The system was able to generate simulated aerial videos of wide urban areas.

With the rise of autonomous vehicles, one of the most relevant applications for computer vision and DL are self driving cars. Angus et al. [32] explore using commercial video games to generate large-scale, high-fidelity training data for semantic segmentation in autonomous driving scenarios. The authors use GTAV to simulate, render and annotate a synthetic dataset of cars driving in urban roads (URSA dataset). In-game AI is used to drive the vehicles in simulation and an optimised annotation algorithm is proposed to segment the synthetic frames. The performances of URSA dataset used as a traning set for semantic segmentation were tested on the popular CityScapes dataset. The evaluation metric used was the class-specific intersection over union (c-IoU) averaged over all of the 19 classes of the dataset, and the networks used for segmentation were FCN and SegNet. The FCN model trained on an URSA synthetic dataset obtained an IoU in line with other state-of-the-art synthetic datasets: URSA = 0.139; Playing For Benchmarks (PFB) = 0.170; Synthia = 0.126. The performances are still far from the one obtained training directly FCN on CityScapes (c-IoU 0.449). However, using URSA plus only 10% of CityScapes dataset for training, the mean c-IoU gets closer to the baseline (0.422). Autonomous drones have now become very popular. Fonder et al. [35] used Unreal Engine and the AirSim plugin to

generate a simulated dataset of flying drones. The authors simulated several sensors other than RGB cameras (GPS, depth sensor, accelerometer and gyroscope), and they were able to produce several pieces of simulated data (RGB on-board camera images, and depth, normal and semantic segmentation maps). Different weather conditions and seasonal changes were present.

In problems where camera motion is the central focus, 3D simulated data are a good solution for data augmentation. Lu et al. [30] presented a framework to generate synthetic shaking videos to train video stabilization and deblurring models. As seen in previous papers, the authors generate various camera paths and motions starting from motion capture data. Matlab was used to process the original shaking motion capture data and produce the new synthetic camera motions. Autodesk Maya was used to render the synthetic videos of a camera moving in different static environments and with different illumination.

Simulated data do not need to render the entire scene. Some methods superimpose new simulated actors or objects on top of existing videos. In [34], Rimboux et al. generate an augmented synthetic dataset of video from surveillance cameras. The idea is to extract the background from a surveillance video sequence, calculate the 3D plane of the ground and superimpose 3D simulated pedestrians on top of it. This method is able to create a realistic video sequence, but it needs significant preprocessing to extract the background and calculate the 3D walkable area of the scene. The data generation model was tested in a person detection scenario using the Town Centre data set. A pretrained version of ResNet-101 and RPN+ networks was used as baseline. The two networks were then trained on the proposed synthetic dataset improving the average accuracy compared to the baselines (ResNet-101 39.09% $\rightarrow$ 50.04%, RPN+ 20.77% $\rightarrow$ 21.32%). A similar approach to generate video simulation for self driving cars (GeoSim) has been proposed by Chen et al. [52]. The idea behind their work is to superimpose realistic dynamic objects (cars) on existing on-board camera videos in order to obtain novel synthetic data. The authors create a dataset of 3D dynamic objects from real images and LiDAR point clouds. Starting from camera video footage, LiDAR point clouds, and HD maps, GeoSim adds and moves the 3D models of the cars into the 3D representation of the source video. The 3D scene is then used to add the car model to the source video generating a new realistic synthetic sequence. In addition, the authors collected a real-world dataset from cameras mounted on a fleet of self driving cars (UrbanData). To test the synthetic data generation, they trained two semantic segmentation networks (PSPNet, DeepLabv3) on the UrbanData dataset, both with and without augmenting it with synthetic data. The networks trained on the augmented dataset obtained an increase in the mean IoU: PSPNet +1.8%; DeepLabv3 +0.2%. Varol et al. [56] automatically estimate 3D human motions from videos and use that information to generate synthetic videos of realistic 3D human avatars performing those actions. The authors tested the data augmentation quality of their synthetic data training a 3D ResNet-50 network on the NTU RGB+D dataset of human actions. The network was tested with and without data augmentation. The model trained on real+synthetic data improved the mean accuracy over the one trained only on real data (74.8% $\rightarrow$ 81.7%).

3.5. Solving the Reality Gap (Simulation + GAN)

The reality gap is the subtle discrepancy between reality and simulation that prevents DL models to properly learn from simulated images. One way to alleviate the problem is to exploit the recent advancement in generative adversarial networks. GAN models can be used to refine synthetic images to be visually closer to real ones. Reently, Wang et al. [48] used a similar idea in their data augmentation framework for crowd videos. They created two synthetic datasets. The first one is a large synthetic video training set with labels generated using the video game GTAV; the second one is a smaller dataset of synthetic images generated by a CycleGAN. The CycleGAN takes as input real and simulated images and generates realistic images based on the two. CycleGAN generated dataset preserves the labels of the original simulated videos. The large synthetic dataset was used to pretrain

a CNN crowd understanding model. The crowd model was then fine-tuned on the smaller refined dataset.

Bongini et al. [58] used Unity game engine to augment a thermal imagery video dataset with synthetic images. In order to improve the fidelity of the synthetic images generated in Unity, the authors used a LSGAN model. Starting from a synthetic image and its segmentation mask, the generative network is able to output a refined thermal image.

4. Future Directions and Conclusions

Video analysis is a rising research topic and DL is the best tool to tackle it. Unfortunately, DL models need a large amount of training data that are often not available or difficult to collect. For this reason, in the last couple of years, several researchers are trying to find solutions for data augmentation for videos.

4.1. From Static Image to Video Data Augmentation

Data augmentation for static images is a well developed research field. Several models able to augment static image datasets are already available, and it is worth analysing more in detail the possibility of their use on video data augmentation. In order to apply standard image data augmentation methods to video sequences, the time domain needs to be taken in consideration. The changes applied to each frame must be coherent through time.

In video analysis, time series of images are usually organised in mini batches representing short clips. To guarantee a time coherence, geometric and color transformation must remain constant through the entire mini batch, and they can be randomised over different mini batches. In other scenarios, 2D motion models can be used to extend static image data augmentation methods to videos. Data augmentation systems based on random erasing, for example, generate the augmented images removing the pixel values from random patches of the original ones. For video data augmentation, instead of randomly selecting a patch for each image, the patch size and position can change based on a predefined 2D motion on the image plane. This approach simulates the occlusions created by a dynamic moving object in the scene.

DL based data augmentation models can perform better on videos if they are able to keep a memory of the previous frames to generate the new ones. The past frames contain information about temporal variations in the scene, like object motions, dynamic light changes, and weather evolution, among others. RNN are widely used for the analysis of text and time series due to their ability to retain a memory of the past inputs through their internal loops. Recently, 1D RNN models (i.e., LSTM and GRU) have been integrated to CNN (ConvLSTM [64], ConvGRU [65]) to perform video analysis and generation. Another approach used to analyse image temporal sequences is the use of 3D convolutions. In this case, the third dimension is used to stack several contiguous frames to obtain temporal information. Extending generator networks with a time series specific model like 3D convolutions or RNN is a promising solution.

Several simulator tools able to generate synthetic images for object detection dataset augmentation (e.g., the Unreal Engine 4 plugin "NVIDIA Deep learning Dataset Synthesizer (NDDS)" [66]) already exists, but their randomisation routines do not usually take into consideration time dependency for the creation of simulated video sequences. Some of the reviewed papers are starting to move in that direction (e.g., ElderSim, a synthetic data generation platform for human action recognition [49]).

4.2. Future Directions

From our review, it is possible to notice that basic transformations for video data augmentation are effective, but they are not as flexible as more powerful approaches like generative adversarial networks or 3D simulators. More recent papers tend to utilise GAN models to generate realistic synthetic data (see Figure 3), but the authors often use the same architectures both for problems of image and video analysis. Some authors are starting to

integrate, to generators, models used for sequence analysis like RNN or 3D convolutions, and we believe this will be the future direction.

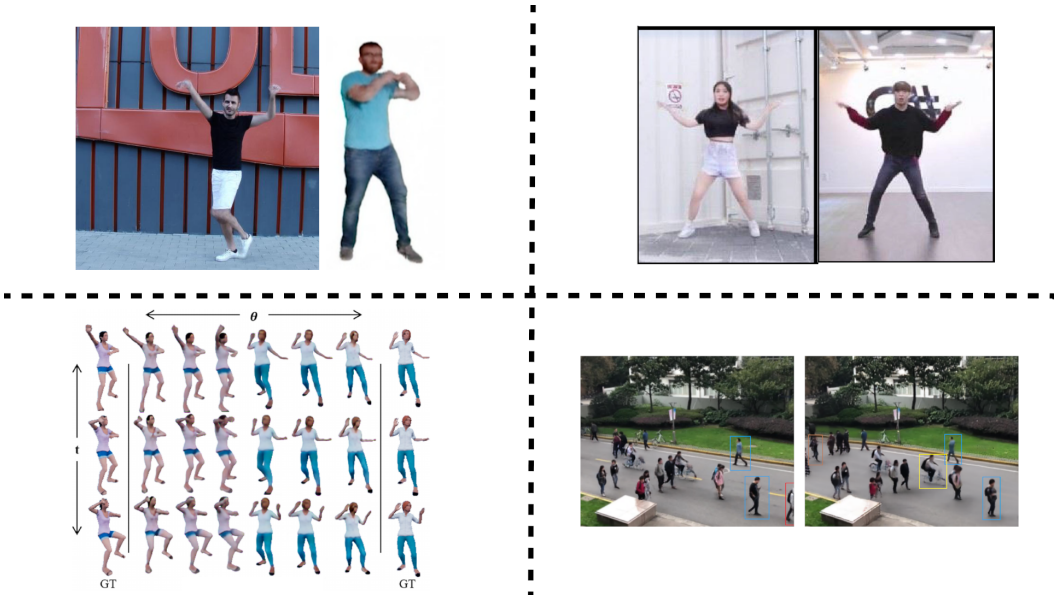

Figure 3. Examples of synthetic images generated by GAN models. Please refer to the original papers for more information. (**Upper left**) Aberman et al. [33]. (**Upper right**) Wei et al. [51]. (**Bottom left**) Kwon et al. [40]. (**Bottom right**) Chai et al. [41].

On the other side, 3D simulators and game engines are becoming more powerful and user-friendly. Game engine companies are getting involved in Computer Vision and Machine Learning problems and some specialized DL and AI plugins for the most popular engines (Unity and Unreal) are starting to appear. In the next few years, this collaboration between academic research and game industry will become even stronger. One of the biggest advantages of the use of simulators for data augmentation is the ease of automatically labelling the data and generating complex annotations like segmentation or depth maps. Moreover, next generation engines are able to render, at high framerate, images almost indistinguishable from real ones (see Figure 4). One drawback of simulated videos is the reality gap. Even if synthetic images generated by 3D game engines are getting close to be identical to real images, they are still not able to replicate all the visual and physical details of the real world. We have seen that, in order to reduce the reality gap, a combination between simulated videos and GAN models can be used. Researchers are starting to propose architectures where a simulator is used to generate the synthetic images, while a GAN model is used to refine them, resulting in generated images closer to real camera ones. This approach takes the best of the two worlds: high fidelity of GAN generated images and flexibility of 3D simulators.

Important aspects to take in consideration while training and running DL models are time and memory efficiency. Usually, the time needed by a data augmentation algorithm to generate the dataset is negligible compared to the one needed for collecting and labeling novel real-world data. Even in the eventuality that the augmented data generation is a time consuming process, once the augmented dataset is generated, various tests can be performed on it. Unfortunately, it is not always possible to generate the entire dataset before training. Sometimes, the dataset has to be generated online. The reasons for this can be many: limited memory to store the data, generation of the data guided by the training for reinforcement learning models, and so on. For data augmentation algorithms based on basic transformations, computational time usually is not an issue. Rotations, cropping, change in illumination, noise addition and image mixing are usually simple transformations that are quick to calculate. Complex simulations, on the other hand, can be computationally expensive. Fortunately, in the last few decades, game engines made

a significant improvement in speed and graphical fidelity (games usually run at 120/60 fps). Recent data augmentation methods that use game engines (Unity, Unreal Engine, GTA) to produce synthetic data are able to generate a frame in few milliseconds. Other data augmentation methods that are computationally demanding are the ones based on DL models (e.g., GAN networks). Big generator networks can require powerful GPUs to generate the augmented data and several hours to be trained. In case of online generation of the augmented data, sharing the GPU between the data augmentation model and the main DL model can be complicated. Computational efficiency of data augmentation methods is a crucial issue that needs to be carefully addressed in future papers on video data augmentation.

Figure 4. Examples of synthetic images generated by 3D simulators. Please refer to the original papers for more information. (**Upper left**) Kerim et al. [55]. (**Upper right**) De Souza et al. [42]. (**Bottom left**) Fonder et al. [35]. (**Bottom right**) Hwang et al. [49].

4.3. Conclusions

This paper presented a complete review of the state of the art in data augmentation specifically addressing the problem of video datasets. We analysed 35 papers published in the period between 2016 and the first months of 2022 pointing out the most common methodologies in use and future directions. Recently, video data augmentation has gained popularity, due to the rise of several applications based on video analysis. From our research, a review having its focus only on video data augmentation is missing and this survey has the target to fill that gap. We noticed how the problem of video data augmentation is having a big impact in the CV community, demonstrated by the exponential growth of papers on that topic in the last few years. This review shows that, in data augmentation, we are having a transition from methods based on basic image transformations to more complex generative and simulated models. The latter models are more powerful and flexible, but they also bring new challenges and open future research directions. We tried to address some of them in this review.

Author Contributions: Conceptualization, N.C.; methodology, N.C. and D.R.R.; investigation, N.C.; writing—original draft preparation, N.C.; writing—review and editing, N.C. and D.R.R.; funding acquisition, D.R.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union's Horizon 2020 Marie Skłodowska-Curie Actions Individual Fellowships under the Grant No. 101031646.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jiao, L.; Zhao, J. A survey on the new generation of deep learning in image processing. *IEEE Access* **2019**, *7*, 172231–172263. [CrossRef]
2. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
3. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
4. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
5. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
6. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2636–2645.
7. Guan, H.; Liu, M. Domain adaptation for medical image analysis: A survey. *IEEE Trans. Biomed. Eng.* **2021**, *69*, 1173–1185. [CrossRef] [PubMed]
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2012; Volume 25. Available online: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf (accessed on 14 February 2022).
9. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*. Available online: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf (accessed on 14 February 2022).
10. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
11. Tremblay, J.; To, T.; Sundaralingam, B.; Xiang, Y.; Fox, D.; Birchfield, S. Deep object pose estimation for semantic robotic grasping of household objects. *arXiv* **2018**, arXiv:1809.10790.
12. Technologies, U. Unity Homepage. Available online: https://unity.com/ (accessed on 14 February 2022).
13. Games, E. Unreal Engine Homepage. Available online: https://www.unrealengine.com/en-US/ (accessed on 14 February 2022).
14. Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 23–30.
15. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *arXiv* **2014**, arXiv:1406.2199.
16. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef] [PubMed]
17. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
18. Lee, N.; Choi, W.; Vernaza, P.; Choy, C.B.; Torr, P.H.; Chandraker, M. Desire: Distant future prediction in dynamic scenes with interacting agents. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 336–345.
19. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
20. Khalifa, N.E.; Loey, M.; Mirjalili, S. A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artif. Intell. Rev.* **2021**, 1–27. Available online: https://link.springer.com/article/10.1007/s10462-021-10066-4 (accessed on 14 February 2022). [CrossRef] [PubMed]
21. Wang, X.; Wang, K.; Lian, S. A survey on face data augmentation for the training of deep neural networks. *Neural Comput. Appl.* **2020**, *32*, 15503–15531. [CrossRef]
22. Chlap, P.; Min, H.; Vandenberg, N.; Dowling, J.; Holloway, L.; Haworth, A. A review of medical image data augmentation techniques for deep learning applications. *J. Med. Imaging Radiat. Oncol.* **2021**, *65*, 545–563. [CrossRef] [PubMed]
23. Naveed, H. Survey: Image mixing and deleting for data augmentation. *arXiv* **2021**, arXiv:2106.07085.
24. Scopus. Scopus Homepage. Available online: https://www.scopus.com/ (accessed on 14 February 2022).
25. Charalambous, C.; Bharath, A. A data augmentation methodology for training machine/deep learning gait recognition algorithms. In Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, 19–22 September 2016; pp. 110.1–110.12.
26. Wang, L.; Ge, L.; Li, R.; Fang, Y. Three-stream CNNs for action recognition. *Pattern Recognit. Lett.* **2017**, *92*, 33–40. [CrossRef]
27. De Souza, C.; Gaidon, A.; Cabon, Y.; López, A. Procedural generation of videos to train deep action recognition networks. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 2594–2604.

28. Wang, W.; Shen, J.; Shao, L. Video Salient Object Detection via Fully Convolutional Networks. *IEEE Trans. Image Process.* **2018**, *27*, 38–49. [CrossRef]
29. Griffith, E.; Mishra, C.; Ralph, J.; Maskell, S. A system for the generation of synthetic Wide Area Aerial surveillance imagery. *Simul. Model. Pract. Theory* **2018**, *84*, 286–308. [CrossRef]
30. Lu, S.P.; You, J.; Ceulemans, B.; Wang, M.; Munteanu, A. Synthesis of Shaking Video Using Motion Capture Data and Dynamic 3D Scene Modeling. In Proceedings of the International Conference on Image Processing, ICIP, Athens, Greece, 7–10 October 2018; pp. 1438–1442.
31. Dong, J.; Li, X.; Xu, C.; Yang, G.; Wang, X. Feature re-learning with data augmentation for content-based video recommendation. In Proceedings of the MM 2018—2018 ACM Multimedia Conference, Seoul, Korea, 22–26 October 2018; pp. 2058–2062.
32. Angus, M.; Elbalkini, M.; Khan, S.; Harakeh, A.; Andrienko, O.; Reading, C.; Waslander, S.; Czarnecki, K. Unlimited Road-scene Synthetic Annotation (URSA) Dataset. In Proceedings of the IEEE Conference on Intelligent Transportation Systems, ITSC, Maui, HI, USA, 4–7 November 2018; pp. 985–992.
33. Aberman, K.; Shi, M.; Liao, J.; Liscbinski, D.; Chen, B.; Cohen-Or, D. Deep Video-Based Performance Cloning. *Comput. Graph. Forum* **2019**, *38*, 219–233. [CrossRef]
34. Rimboux, A.; Dupre, R.; Daci, E.; Lagkas, T.; Sarigiannidis, P.; Remagnino, P.; Argyriou, V. Smart IoT cameras for crowd analysis based on augmentation for automatic pedestrian detection, simulation and annotation. In Proceedings of the 15th Annual International Conference on Distributed Computing in Sensor Systems, DCOSS 2019, Santorini Island, Greece, 29–31 May 2019; pp. 304–311.
35. Fonder, M.; Van Droogenbroeck, M. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 553–562.
36. Wu, D.; Chen, J.; Sharma, N.; Pan, S.; Long, G.; Blumenstein, M. Adversarial Action Data Augmentation for Similar Gesture Action Recognition. In Proceedings of the International Joint Conference on Neural Networks, Budapest, Hungary, 14–19 July 2019.
37. Sakkos, D.; Shum, H.; Ho, E. Illumination-based data augmentation for robust background subtraction. In Proceedings of the 2019 13th International Conference on Software, Knowledge, Information Management and Applications, SKIMA 2019, Island of Ulkulhas, Maldives, 26–28 August 2019.
38. Li, J.; Yang, M.; Liu, Y.; Wang, Y.; Zheng, Q.; Wang, D. Dynamic hand gesture recognition using multi-direction 3D convolutional neural networks. *Eng. Lett.* **2019**, *27*, 490–500.
39. Sakkos, D.; Ho, E.; Shum, H.; Elvin, G. Image editing-based data augmentation for illumination-insensitive background subtraction. *J. Enterp. Inf. Manag.* **2020**. Available online: https://www.emerald.com/insight/content/doi/10.1108/JEIM-02-2020-0042/full/html (accessed on 14 February 2022). [CrossRef]
40. Kwon, Y.; Petrangeli, S.; Kim, D.; Wang, H.; Park, E.; Swaminathan, V.; Fuchs, H. Rotationally-Temporally Consistent Novel View Synthesis of Human Performance Video. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 387–402.
41. Chai, L.; Liu, Y.; Liu, W.; Han, G.; He, S. CrowdGAN: Identity-free Interactive Crowd Video Generation and Beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. Available online: https://www.computer.org/csdl/journal/tp/5555/01/09286483/1por0TYwZvG (accessed on 14 February 2022). [CrossRef]
42. de Souza, C.; Gaidon, A.; Cabon, Y.; Murray, N.; López, A. Generating Human Action Videos by Coupling 3D Game Engines and Probabilistic Graphical Models. *Int. J. Comput. Vis.* **2020**, *128*, 1505–1536. [CrossRef]
43. Namitha, K.; Narayanan, A.; Geetha, M. A Synthetic Video Dataset Generation Toolbox for Surveillance Video Synopsis Applications. In Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing, ICCSP 2020, Nanjing, China, 10–12 January 2020; pp. 493–497.
44. Isobe, T.; Han, J.; Zhuz, F.; Liy, Y.; Wang, S. Intra-Clip Aggregation for Video Person Re-Identification. In Proceedings of the International Conference on Image Processing, ICIP, Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 2336–2340.
45. Zhang, Y.; Jia, G.; Chen, L.; Zhang, M.; Yong, J. Self-Paced Video Data Augmentation by Generative Adversarial Networks with Insufficient Samples. In Proceedings of the MM 2020—28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1652–1660.
46. Yun, S.; Oh, S.J.; Heo, B.; Han, D.; Kim, J. Videomix: Rethinking data augmentation for video classification. *arXiv* **2020**, arXiv:2012.03457.
47. Ye, Y.; Yang, K.; Xiang, K.; Wang, J.; Wang, K. Universal semantic segmentation for fisheye urban driving images. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 11–14 October 2020; pp. 648–655.
48. Wang, Q.; Gao, J.; Lin, W.; Yuan, Y. Pixel-Wise Crowd Understanding via Synthetic Data. *Int. J. Comput. Vis.* **2021**, *129*, 225–245. [CrossRef]
49. Hwang, H.; Jang, C.; Park, G.; Cho, J.; Kim, I. ElderSim: A Synthetic Data Generation Platform for Human Action Recognition in Eldercare Applications. *arXiv* **2020**, arXiv:2010.14742.
50. Tsou, Y.Y.; Lee, Y.A.; Hsu, C.T. Multi-task Learning for Simultaneous Video Generation and Remote Photoplethysmography Estimation. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020; Volume 12626, pp. 392–407.

51. Wei, D.; Xu, X.; Shen, H.; Huang, K. GAC-GAN: A General Method for Appearance-Controllable Human Video Motion Transfer. *IEEE Trans. Multimed.* **2021**, *23*, 2457–2470. [CrossRef]
52. Chen, Y.; Rong, F.; Duggal, S.; Wang, S.; Yan, X.; Manivasagam, S.; Xue, S.; Yumer, E.; Urtasun, R. GeoSim: Realistic Video Simulation via Geometry-Aware Composition for Self-Driving. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2021; pp. 7226–7236.
53. Dong, J.; Wang, X.; Zhang, L.; Xu, C.; Yang, G.; Li, X. Feature Re-Learning with Data Augmentation for Video Relevance Prediction. *IEEE Trans. Knowl. Data Eng.* **2021**, *33*, 1946–1959. [CrossRef]
54. Hu, L.; Huang, S.; Wang, S.; Liu, W.; Ning, J. Do We Really Need Frame-by-Frame Annotation Datasets for Object Tracking? In Proceedings of the MM 2021—29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 4949–4957.
55. Kerim, A.; Celikcan, U.; Erdem, E.; Erdem, A. Using synthetic data for person tracking under adverse weather conditions. *Image Vis. Comput.* **2021**, *111*, 104187. [CrossRef]
56. Varol, G.; Laptev, I.; Schmid, C.; Zisserman, A. Synthetic Humans for Action Recognition from Unseen Viewpoints. *Int. J. Comput. Vis.* **2021**, *129*, 2264–2287. [CrossRef]
57. Hu, Y.T.; Wang, J.; Yeh, R.; Schwing, A. SAIL-VOS 3D: A synthetic dataset and baselines for object detection and 3d mesh reconstruction from video data. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Nashville, TN, USA, 19–25 June 2021; pp. 3359–3369.
58. Bongini, F.; Berlincioni, L.; Bertini, M.; Del Bimbo, A. Partially Fake it Till you Make It: Mixing Real and Fake Thermal Images for Improved Object Detection. In Proceedings of the MM 2021—29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 5482–5490.
59. Otberdout, N.; Daoudi, M.; Kacem, A.; Ballihi, L.; Berretti, S. Dynamic Facial Expression Generation on Hilbert Hypersphere with Conditional Wasserstein Generative Adversarial Nets. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 848–863. [CrossRef]
60. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
61. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6023–6032.
62. Sadeghi, F.; Levine, S. Cad2rl: Real single-image flight without a single real image. *arXiv* **2016**, arXiv:1611.04201.
63. Blender. Blender Homepage. Available online: https://www.blender.org/ (accessed on 14 February 2022).
64. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. Available online: https://proceedings.neurips.cc/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf (accessed on 14 February 2022).
65. Siam, M.; Valipour, S.; Jagersand, M.; Ray, N. Convolutional gated recurrent networks for video segmentation. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3090–3094.
66. To, T.; Tremblay, J.; McKay, D.; Yamaguchi, Y.; Leung, K.; Balanon, A.; Cheng, J.; Hodge, W.; Birchfield, S. NDDS: NVIDIA Deep Learning Dataset Synthesizer. 2018. Available online: https://github.com/NVIDIA/Dataset_Synthesizer (accessed on 14 February 2022).