Synthetic Data for Object Detection with Neural Networks: State of the Art Survey of Domain Randomisation Techniques

ADAM WESTERSKI and FONG WEE TECK, Institute for Infocomm Research (I2R), A*STAR, Singapore

Machine learning relies heavily on access to large and well-maintained datasets. In this article, we focus on Computer Vision and object detection applications to survey past research on automatic generation of annotated datasets that does not require costly and time-consuming human labelling. In specific, we analyse research done in the area of Domain Randomisation applied to Neural Networks predominant in object detection since the last decade. We propose a set of criteria for comparison of previously published works, and utilise these criteria to make conclusions about various trends in the area, similarities/differences and key discoveries made since conception. The purpose of this work is to advise practitioner on leading solutions and help researchers gain better understanding of the landscape. The key takeaways from our analysis show the current state of the art solutions within the mid-quartile range allow object detection with typically about 1-25% performance decrease in comparison to manually annotated datasets; while the top performant approaches above the upper quartile gain about 2-32% lead over real data training in their specific application areas. Our survey shows the future outlook is more research into 3D generation techniques, with most innovative yet complex techniques related to end-to-end modifications of entire network architectures to suit synthetic data training.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; *Neural networks*; **Machine learning**; **Computer vision**; • **General and reference** → **Surveys and overviews**.

Additional Key Words and Phrases: synthetic data; domain randomisation; supervised learning; machine learning; object detection

ACM Reference Format:
Adam Westerski and Fong Wee Teck. yyyy. Synthetic Data for Object Detection with Neural Networks: State of the Art Survey of Domain Randomisation Techniques. *ACM Trans. Multimedia Comput. Commun. Appl.* xx, x (yyyy), 20 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In the second decade of 21st century there was a major shift in Computer Vision (CV) research. The classical CV approaches used hand-crafted features based on image descriptors such as Scale-Invariant Feature Transform (SIFT) [33] or Histogram of Oriented Gradients (HOG) [8]. Those features could be later used as input to regular machine learning models such SVM, Decision Trees etc., delivering predictions on object locations or object types in images. This methodology was disrupted by the progress made in Deep Learning which dominates CV research until today. In object detection task in specific, the accuracy gains in benchmarks such as ImageNet [9] and MSCOCO [31] to large degree can be attributed to multiple Neural Network architectural innovations (e.g. AlexNet [29], VGG [53], ResNet [19]). However, outside of such research competitions on fixed datasets, in real-life applications one of the key hurdles for deployment of award winning solutions

Authors' address: Adam Westerski, adam-westerski@i2r.a-star.edu.sg; Fong Wee Teck, Institute for Infocomm Research (I2R), A*STAR, 1 Fusionopolis Way, Singapore, 138632.

1551-6857/yyyy/0-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

is availability of annotated domain specific datasets. An asset obtained in a time-consuming manual human operated process, which is therefore very costly to maintain in a sustainable manner over time in a business scenario.

One of the answers of research community to the problem of dataset availability is Domain Randomisation (DR) - an algorithmic approach to generation of so-called synthetic data, ie. data with objects inserted into images in a randomised or semi-randomised manner, therefore coming with precise annotations regarding object position in the scene.

The theoretical basis for such approach is belief that there is some finite albeit large distribution of possible object placements, rotations etc. that covers all images in certain area of application for object detection. Therefore, the goal for Domain Randomisation is generating a new dataset to cover to an extent such distribution. In practice, those approaches have a number of shortcomings related to so-called 'domain gap', ie. the difficulty to accurately reflect this distribution. Those difficulties, as well as debate on their reasons and ways of addressing them have generated a number of innovations published across the several past years. This article is a review of such previously published work.

Our goal is to synthesise the key innovations and trends in area of Domain Randomisation, show what can be currently achieved and how does it compare to usage of traditional human crafted datasets. In order to do so, we propose a number of criteria to distinguish different approaches to Domain Randomisation as well as pick the relevant ones given domain of application.

We start the review with a simple dichotomy, splitting Domain Randomisation research into 2D or 3D image generation to point out the key contributions and trends in chronological order (see Sec. 3). Moving on, we compare works based on more detailed criteria pertaining to area of application, generated dataset characteristics and detection task specifics (see Sec. 4). We use this opportunity to discuss how authors evaluate their research and compare legitimacy of results based on evaluation scope and thoroughness of evaluation in each article. Further, for a more scrutinised and systematic approach, we establish a list of 20 characteristics of Domain Randomisation algorithms and classify published works based on those (see Sec. 5). By synthesising ablation studies from multiple articles, we compare which of those 20 techniques are most frequently addressed by researchers and which are attributed to achieve biggest gains in performance. We list all reported performances and compare relative gains made within each dataset/domain of application (see Sec. 6). Finally, taking advantage of earlier introduced DR categories, we plot all innovations on a timeline and review State of the Art from the perspective of progress done across the years (see Sec. 7). We conclude the study with remarks for practical use of Domain Randomisation techniques and extrapolate the earlier presented timeline perspective with an outlook for the future (see Sec. 8).

2 RELATED WORK

Although some researchers note the need of universal benchmarks for synthetic data generation [1, 45], so far there hasn't been any established and commonly followed means of comparing different approaches published in the literature. A proposal from [2] is to assess synthetic datasets via measurement of image or image feature distribution distance from the real dataset. This method omits the task dependant evaluation metrics and only requires images to calculate the metric. However, beyond the original publication the effectiveness of this approach has not been assessed or adapted in the state of the art. As can be seen in subsequent sections, published works that propose new original contributions to Domain Randomisation typically evaluate proposed algorithms only on their specific target domain dataset and using typical Computer Vision metrics. Frequently, for many publications the synthetic data, real images and implementation details are not shared. This makes direct comparison between different publications challenging. In our article, we establish

several means of comparing published works and intend to give guidelines as per effectiveness and scope of different methods.

The past surveys of state of the art in synthetic data generation [39, 61] focus on categorising broad spectrum of published articles and summarise the generation methods. In comparison to those, we aim at a more quantitive evaluation. Nikolenko [39] reviews all applications of synthetic data and charts the general landscape. He gathers a huge amount of references and shows a top level overview of major trends in multiple areas. We assume a narrower scope of Object Detection and Domain Randomisation only, which allows us to give a more in-depth and fine-grained comparison. Similarly to Nikolenko, we also leave out the review of data augmentation techniques as a separate topic and direct the reader towards plenty of other works already analysing this area [52, 64]. In comparison to [61] who delivers a qualitative comparison of synthetic data techniques, we propose to review published works in quantitive terms of the relative accuracy gains in comparison to their domain baselines and by systematically describing the scope of data generation against a proposed taxonomy.

All works analysed in aforementioned synthetic data surveys as well as in this article pertain to Neural Network research in specific. However, it should be noted that prior to Deep Learning boom in CV, scientists have considered using synthetic data for training classical machine learning algorithms as well, e.g. based on HOG. Attempts very similar to current Domain Randomisation were employed for goals of Object Detection (e.g. [56] or [55]). Many past ideas related to data generation keep coming back in recent research, however we omit those classical CV papers in our study as the gains in performance reported in those works are not evaluated in terms of Neural Networks and might give different results.

Likewise, studies on synthetic data generation go beyond application in vision into areas such as Natural Language Processing [14] or tabular data [15]. The parallels between those and vision have been already reviewed by Nikolenko is his study [39]; while the individual achievements for each domain have been reviewed in the highlighted surveys. We refer the reader to those works for more details, in our article we focus on the challenges ahead of the narrow area of Object Detection and Domain Randomisation.

3 OVERVIEW OF PUBLICATIONS AND CONTRIBUTION AREAS

The methodology used for listing all candidate papers for this survey is based on: i) reviewing all articles from aforementioned past state of the art surveys in synthetic data; ii) further following references from all obtained papers in step 1; iii) supplementing the list with papers found via keyword web search and academic aggregator services. For step 3, we used Google Scholar and analysed top 50 papers from three following keyword searches: "domain randomisation", "domain randomisation object detection", "synthetic data object detection". Out of 150 papers analysed in this final step, 96 were unrelated to this study, while 39 were re-occurrences found via previous methodology steps. For all steps of our methodology the criteria used for including articles in this study is as follows: i) perform object detection; ii) utilise Neural Networks; iii) generate new data without necessity for human annotation; iv) report performance measures, specifically mAP.

After following the described methodology we ended up with a list of 37 articles published between 2015 to 2022. We further divided those works into 4 main categories: core works that propose new innovations to the area of Domain Randomisation (23 articles); comparative studies that do not propose any innovation but evaluate different research directions for a selected aspect of Domain Randomisation (4 articles); evaluations of software frameworks (2 articles); applied research that only describes application of known methods to new areas (8 articles). All considered works are listed in Table 1 in chronological order.

Fig. 1. Examples of synthetic data: red bounding boxes mark objects added on top of 2D background. (a) 2D objects cut out from images and plotted onto new background [11]; (b) rendered 3D objects inserted using horizontal plane detection [17].

The process of data generation implemented in those publications is based on placement of foreground objects meant for detection onto some background scene. In terms of articles that offer algorithmic contributions to Domain Randomisation, typically authors choose one of two methods: i) use regular pinhole camera photos to create new 2D images via cut and paste technique; or ii) use of 3D models and rendering engines (see examples on Figure 1). In both scenarios, the final output are 2D images that can be supplied as training data for typical object detectors such as Faster RCNN [49], SSD [32] etc.

Across all published work, there is a visible distinction on how the aforementioned "domain gap" should be addressed: some authors try to create photo realistic images, while others experiment how far simplistic synthetic data can go until it impacts performance. In between those two extremes, a group of works shifts focus to other aspects such as adjusting Neural Network training [21], sourcing object models efficiently to ensure better coverage of target domain [62].

The non-realistic approach to Domain Randomisation was pioneered by [11] with 2D foreground objects cut from photos by detecting boundaries with a Neural Network and afterwards placing objects onto random photo backgrounds from a different dataset. In Dwibedi approach, all parameters regarding object placement are randomised within some predefined boundaries. Instead of focus on realism, Dwibedi is one of the first that applied blending to foreground objects on the premise that pasted artefacts onto backgrounds have superficial boundary pixel differences due to different lighting or imperfections in cutting images. Blending allows to avoid Neural Network detecting pixel boundaries and generalise to object features that are also present in real images. Those techniques were later used in multiple applied works as well as extended in research. [35] proposed pasting segments of objects to simulate occlusion, while [60] expanded on concept of accommodating for object boundaries by placing similarly cut distractor objects into background. Both [65] and [16] take advantage of those cut-paste techniques but additionally apply generative AI to improve the final detection performance. [16] uses Stable Diffusion [50] to generate contextually meaningful backgrounds; while [65] adds image post-processing with GAN [25] for foreground to increase distribution size and shrink the domain gap. Moving towards realism in 2D approach, several works propose different ways of realistic object placement and orientation. [10] attempt to

Table 1. Key contributions by article, chronological order*. Novelty Areas: 2D- 2D data generation; 3D- 3D data generation; NN- Neural Network modification; CS- comparative study; AA- area of application; FW- software framework

AUTHORS	YEAR	AREA	KEY CONTRIBUTION
Peng et al. [44]	2015	(3D)	Plain object texture/ plain background
Dwibedi et al. [11]	2017	(2D)	Object cuts; Blending objects and background
Mitash et al. [38]	2017	(3D)	Realistic object poses / physics engine to simulate gravity
Georgakis et al. [17]	2017	(3D)	Realistic object placement / detect horizontal surfaces
Hinterstoisser at al. [21]	2018	(NN)	Freezing backbone weights for synthetic training
Dvornik et al. [10]	2018	(2D)	Predict object placement using Neural Network
Borrego at al. [5]	2018	(3D)	Use of simple shapes / artificial textures (colour, gradient)
Tremblay et al. [59]	2018	(3D)	Use of Distractor objects
Nogues et. al. [40]	2018	(3D)	Post-processing synthetic images for realism
Wrenninge et. al. [63]	2018	(3D)	Physically based rendering (3D background / outdoor)
Hinterstoisser et al. [22]	2019	(3D)	Background generated from distractor 3D objects
Mahmood et. al. [35]	2019	(2D)	Simulate occlusion by segmenting object and paste parts
Prakash et. al. [47]	2019	(3D)	Procedural/parametric scene generation
Wong et. al. [62]	2019	(3D)	Photogrammetry to create 3D objects from photos
Tripathi et. al. [60]	2019	(NN)	Predict object placement/rotation with Neural Network
Kar et. al. [27]	2019	(NN)	Estimation of DR parameters with Neural Network
Hodaň et. al. [23]	2019	(3D)	Physically based rendering (3D backgrounds / indoor)
Chengsong et. al. [24]	2021	(NN)	Semi-supervised learning on top of synthetic training
Hughes et. al. [26]	2021	(3D)	Use of differential rendering / Adversarial samples
Eversberg et. al. [12]	2021	(3D)	Physically based rendering with 2D backgrounds
Clement et. al. [7]	2021	(NN)	Split detection and classification, GAN based blending
Yun et. al. [65]	2021	(2D)	Domain Adaptation to support synthetic data pipeline
Ge et. al. [16]	2022	(2D)	Generative AI used for synthetic backgrounds
Arcidiacono et. al. [3]	2018	(CS)	Comparison of 3D vs. 2D
Lidberg et. al. [30]	2018	(CS)	Comparison of 3D vs. 2D
Nowruzi et. al. [41]	2019	(CS)	Impact of synthetic data quantities
Fabbri et. al. [13]	2021	(CS)	Impact of synthetic data diversity
Heindl et. al. [20]	2021	(FW)	framework for data generation
Borkman et. al. [4]	2021	(FW)	framework for data generation
Param et. al. [48]	2017	(AA)	Fridge items detection
Buls et. al. [6]	2019	(AA)	Object piles / grappling
Shermeyer et. al. [51]	2021	(AA)	Arial detection
Zhang et. al. [66]	2021	(AA)	Weed detection in soil
Madan et. al. [34]	2021	(AA)	Icon detection in iconographic
Mao et. al. [36]	2021	(AA)	Bird detection
Khirodkar et. al. [28]	2021	(AA)	Car detection in parking lots
Margapuri et. al. [37]	2021	(AA)	Seed detection

*Articles below the dashed line do not contribute any significant new Domain Randomisation method applicable universally. Those works limit to new area of application for DR or evaluate some aspects of DR generation.

find the correct placement by teaching a model to detect object surroundings through constructing training data with target objects occluded. [60] include object detector into the data generation

process and train their model end-to-end in adversarial manner to estimate foreground object orientation and position.

Similar directions in research can be also observed in the data generation based on 3D models. The works that go towards simplicity are: [44] who experiments with variety of simple plain color textures and backgrounds; [5] furthers that study with use of simple geometric shapes as templates for detection; [59] proposes use of such primitives as distractors rather than objects for detection; [22] creates entire backgrounds from such irrelevant distractor objects. In the photo-realistic camp: [38] evaluate impact of realistic poses on detection accuracy; while [17] focus on realistic placement of objects in scenes. The extent to which photo-realism helps is also evaluated by some in later stages of scene generation via Physically Based Rendering (PBR) [12, 23, 63] or post-processing rendered images with GANs [7, 40]. Several works move towards simulators and try to create realistic 3D scenes with randomised elements constrained by rules [47] or scene graphs [27].

The comparison between realistic and non-realistic approaches in an ongoing debate and to our knowledge there has not been a comprehensive attempt at evaluation which could perform better and to what degree in similar conditions. There have been works that tried to put side by side 2D and 3D methodologies: [3] as well as [30] concluded 3D approach giving more promising results. However, both publications employ DR methods far more simplistic in comparison to state of the art techniques. The direct comparison of state of the art methods based on experimental results published in original publications is hard as frequently each publication uses different datasets and experiment settings. In the next section, we summarise per each paper those differences in more detail and propose to compare approaches via coverage of Domain Randomisation techniques.

4 DATASET SCOPES AND APPLICATION AREAS

The majority of reviewed articles evaluate their innovations via generation of training data for detection of small/medium shelf items (see Table 2), such as: kitchen items (e.g. [11]), toys (e.g. [21]), variety of supermarket items (e.g. [22]). Typically those are goods obtainable from regular stores, with some exceptions such as electronics/ manufacturing (e.g. [40]) or agriculture items (e.g. [24]). Among bigger objects, the most distinctive group is car detection related to studies on autonomous driving (e.g. [13, 59]). Remainder of reviewed articles operate on objects from niche application areas such as satellite imaginary detection [7, 51], animal detection [35, 36], document element detection [34].

Depending on the application area and its requirements, authors utilise their synthetic datasets to train machine learning model for detection of different amount of objects. Typically studies use models trained for less than 10 distinct objects (65% articles); more rarely between 10-50 objects (25%); with several exceptions that go beyond 50 objects [4, 22, 48] and up to 100 objects at most [51]. As noted by [18] the count of objects on the image, their distribution and visual differences affect performance of state of the art algorithms; making it necessary to adjust networks to better fit the particular goal.

Therefore, to show their performance gains for specific application areas, authors of Domain Randomisation typically obtain a manually labelled dataset for training and use it as baseline indication of what can be achieved in their domain of interest with regular data sourcing techniques. Subsequently, synthetic dataset is constructed and both methods are evaluated on manually labelled test set with state of the art detectors. About 45% of articles utilise public datasets, while the remainder create their own labelled data. Across the published works, the size of manually labelled train data typically is in range of several hundred (31% of articles) or several thousand samples (40%); the smallest reported train set is 175 samples [5], while the biggest is 20 thousand [21]. The test sets are in similar size range across studies; however typically per individual article smaller than train. The sizes of synthetic train datasets vary greatly from paper to paper, however most

Table 2. Detailed components of methods per article

	OBJECT CLASS		GENERATION		TRAIN SIZE	TEST
AUTHORS	TYPE	COUNT	FGRD	BGRD	(synth) / (real)	SIZE
Peng et al. [44]	VOC classes	20	3D (render)	2D	(s) 2k / (r) 2.5k	4.9k
Dwibedi et al. [11]	Kitchen Items	7	2D (360)	2D	(s) 6k / (r) 3-5k	1.5-3k
Mitash et al. [38]	Shelf items	11	3D (render)	3D	(s) n/a / (r) n/a	2.2k
Georgakis et al. [17]	Kitchen Items	11 / 5	2D (360)	2D	(s) 6.7k / (r) 4k	6.7k
Hinterstoisser at al. [21]	Toys/ kitchen	10	3D (scan)	2D	(s) n/a / (r) 20k	1k
Dvornik et al. [10]	VOC classes	1 / 20	2D (cut)	2D	(s) n/a / (r) 1.4k	4.9k
Borrego at al. [5]	Kitchen Items	3	3D (render)	2D	(s) 30k / (r) 175	49
Tremblay et al. [59]	Cars	36	3D (render)	2D	(s) 100k / (r) 6k	500
Nogues et. al. [40]	Electronics	n/a	3D (render)	2D	(s) 10k / (r) n/a	100
Wrenninge et. al. [63]	Cars	19	3D (render)	3D	(s) 25k / (r) 6k	500
Hinterstoisser et al. [22]	Market Items	64	3D (scan)	3D	(s) n/a / (r) 1.1k	250
Mahmood et. al. [35]	Lobster	1	2D (cut)	2D	(s) 1k / (r) 187	50
Prakash et. al. [47]	Cars	1	3D (render)	3D	(s) 25k / (r) 6k	7.5k
Wong et. al. [62]	Market Items	10	3D (render)	2D	(s) 100k / (r) 600	1k
Tripathi et. al. [60]	Kitchen Items	11	2D (cut)	2D	(s) n/a / (r) 3-5k	6.7k
Kar et. al. [27]	Cars	1	3D (render)	3D	(s) n/a / (r) n/a	7.5k
Hodaň et. al. [23]	Shelf objects	8 / 14	3D (render)	3D	(s) 48k / (r) n/a	1.3k
Chengsong et. al. [24]	Plants	4	2D (cut)	2D	(s) 1.2k / (r) 200	500
Hughes et. al. [26]	Kitchen Items	9	3D (render)	2D	(s) 3k / (r) 3-5k	1.5-3k
Eversberg et. al. [12]	Industry obj.	1	3D (scan)	2D	(s) 5k / (r) 200	650
Clement et. al. [7]	Airplanes	1	3D (render)	2D	(s) n/a / (r) n/a	n/a
Yun et. al. [65]	Kitchen Items	11	2D (360)	2D	(s) 6k / (r) 4-5k	1.5-3k
Ge et. al. [16]	VOC classes	20	2D (cut)	2D	(s) 60k / (r) 1.4k	4.9k
Arcidiacono et. al. [3]	Industrial obj.	6	2D / 3D	2D	(s) 10k (r) 500	500
Lidberg et. al. [30]	National flags	3	2D / 3D	2D	(s) 1.5k (r) 264	150
Nowruzi et. al. [41]	Persons	n/a	3D	3D	(s) 200k (r) 13k	2k
Fabbri et. al. [13]	Persons	n/a	3D	2D	(s) 1.3m (r) n/a	5.9k
Heindl et. al. [20]	Industrial obj.	6	3D (render)	3D	(s) 50k (r) 10k	10k
Borkman et. al. [4]	Market Items	63	3D (scan)	3D	(s) 400k (r) 760	254
Param et. al. [48]	Kitchen Items	55	3D (render)	3D	(s) 4k (r) 400	50
Buls et. al. [6]	Bottle/can	2	2D (360)	2D	(s) 6k (r) n/a	1.5k
Shermeyer et. al. [51]	Airplanes	100	3D (render)	2D	(s) 45k / (r) 253	n/a
Zhang et. al. [66]	Fruits	2	2D (cut)	2D	(s) 3.1k / (r) 1.3k	450
Madan et. al. [34]	Icons	1	2D (cut)	2D	(s) 10k / (r) n/a	1.4k
Mao et. al. [36]	Birds	n/a	3D (render)	2D	(s) 1k / (r) 900	110
Khirodkar et. al. [28]	Cars	1	3D (render)	3D	(s) 10k / (r) 5.6k	200-800
Margapuri et. al. [37]	Seeds	40	2D (cut)	2D	(s) 6.8k / (r) n/a	150

often are several times greater than manually labelled baselines; the smallest ones are around one thousand samples [24, 35, 36]; the biggest 1.3 million [13].

The generation process of synthetic images from those datasets consists of two key parts: foreground and background. In publications that use 3D objects for foreground (71% of all articles),

authors obtain their models from public repositories (e.g. [26] from BigBIRD [54]); prepare hand-crafted models (e.g. [5] using 3D primitives); or digitalise object models via 3D scanning (e.g. [22]) or other reconstruction techniques [62]. Subsequently such objects are plotted over either 2D background (56%) or 3D backgrounds (44%); with final output processed with rendering engine to add textures, lighting, shadows etc. For the 2D approach (37% of all articles) foreground objects are sourced through cutting them out from regular photos (e.g. [10, 23, 35]) or from specially crafted images with objects of interest photographed from different angles over uniform color backgrounds to ease the cutting process [6, 11, 17]. Later, in majority of publications those objects are pasted over 2D images taken from various thematically related datasets disjoint with object detection training or test sets (64%); in other cases backgrounds are obtained via photography done by publication authors of related scenes or environments (21%); or in some experimental cases uniform backgrounds [6] or AI generated backgrounds [16].

In the final step, the foreground and background objects are composed into the scene of synthetic image based on rules that very depending on the publication and approach taken by authors - we detail and classify such composition techniques in the next section.

5 CLASSIFICATION SCHEME FOR DOMAIN RANDOMISATION TECHNIQUES

Following the Domain Randomisation paradigm, authors of reviewed publications assume some set of parameters that describe properties of objects common for all generated scenes; and subsequently vary those parameters across synthetic images in order to obtain a broad distribution of samples in the dataset. In order to organise the contributions in a systematic way, we propose to list all those variables as they are reported in the reviewed publications. Subsequently, as shown on Figure 2, we group the DR parameters based on elements of the scene they affect:

- (F) foreground - objects to be detected
- (B) background - background photos/images or backgrounds composed of multiple objects
- (D) distractor objects / occlusion - similar as foreground objects but not being subject of detection
- (C) camera - modifications on how all elements of scene are captured
- (L) lights - lighting definition for entire scene
- (N) noise - image post-processing application to add distortion
- (R) blur - image post-processing used to smooth differences between background and all objects

Outside of the foreground and background already highlighted in previous section, distractor objects are another addition to the scene mentioned in multiple articles. Those are inserted either behind the foreground or in front to occlude it. The key role of distractors is to prevent the network from discriminating foreground objects based on deficiencies of insertion technique but on true object features. For example, in 2D cut-paste approaches without distractors the network can learn to detect inserted objects based on their lighting and colour shades standing out from background rather than low-level shape features (e.g. circular shape of ball) or high-level visual elements (e.g. car tires, pedestrian body parts). For techniques that relate to camera, lighting and noise; the main purpose is increasing the variance of scenes to better fit the target distribution as well as increasing realism. The blur techniques are mainly utilised as a tool to conceal transition from foreground to background.

For each of the parameters in those groups, there are different methods to achieving variability across dataset samples. Broadly, we propose to put them into three categories that appear in literature: (x) randomised value within some assumed preset range; (p) value asserted in algorithmic or procedural way based on context of the image, may or may not include random sampling; (f) fixed

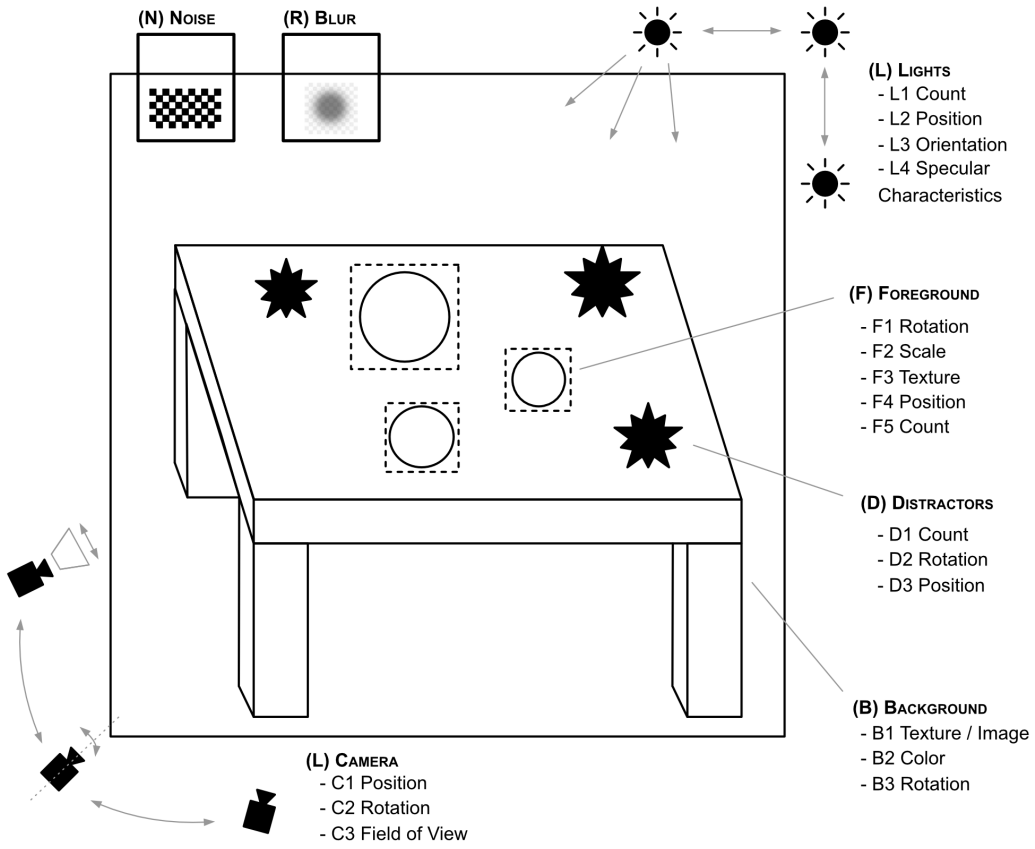

Fig. 2. Illustration of proposed categorisation for Domain Randomisation techniques and its relationship to the scene in the generated synthetic image.

value across all samples. The breakdown of particular generation techniques against parameter groups is listed in Table 3; contributions are marked as using a certain technique only if it's mentioned explicitly in the article.

The most frequently used technique is simple randomisation within a preset range (accounts for 66% of all reported DR usage). Typical examples are: placement of foreground objects positioned at random coordinates within the image resolution boundaries (12% of all simple randomisation usage), random rotation angle of foreground object within some range of degrees (11%), or random background from a set of predefined backgrounds (10%). Such techniques are core to DR and were proposed in early stages of research in the area (e.g. [11]). Currently, those are typically repeated by other authors in their implementations as a supplementary means to the main contribution. The more sophisticated approaches, denoted by "p" - procedural, require some additional logic to determine parameter value. In comparison to regular randomisation, those refer to 22.8% of total reported DR usage. Similarly, procedural techniques are most typically applied for calculation of foreground object position and rotation (respectively 17% and 15% of all procedural DR use). Some examples are: assuring that all rotation angles are equally represented in the dataset [22]; prediction of object position using neural network [10]; prediction of position based on horizontal

Table 3. Domain Randomisation techniques per article ("x" - randomised value per sample, "p" - procedural or algorithmic approach beyond simple randomisation,"f" - one fixed method applied to all samples)

ARTICLE	F1	F2	F3	F4	F5	B1	B2	B3	D1	D2	D3	C1	C2	C3	L1	L2	L3	L4	N1	R1
Peng et al. [44]	x		x			x														
Dwibedi et al. [11]	x	x		x	x	x			x		x									f
Mitash et al. [38]	x			x	x							f	f			x		x		
Georgakis et al. [17]	x	p		p	x	x														f
Hinterstoisser at al. [21]	p	p		x		x		x										x	f	f
Dvornik et al. [10]		p		p	p															x
Borrego at al. [5]			x	x	x							f	f			f	f			
Tremblay et al. [59]	x		x	x		x			x	x	x	x	x		x	x			x	
Nogues et. al. [40]	x		p	x		p						x	x					p		
Wrenninge et. al. [63]	p		p	p	p					p	p						p	p		
Hinterstoisser et al. [22]	p	p		x		p	x	p	x	x	x				x	x		x	x	x
Mahmood et. al. [35]	x	x		x		x														f
Prakash et. al. [47]	p		x	p	p	p	p	p	p	p	p	x	x	x		x	x	x		
Wong et. al. [62]	x		p	f		x						p	p		x	p		x		
Kar et. al. [27]	p		p	p						p	p						x	x		
Hodaň et. al. [23]	p		f	f								p	x	x						
Chengsong et. al. [24]	x			x	x	x	x													f
Hughes et. al. [26]		x		x		x														
Eversberg et. al. [12]	x	x	x	x	f				x	x	x	x	x		x	x	x	x		
Clement et. al. [7]	x			p		x														f
Yun et. al. [65]	x	x	p	p	x	p	p		x		x									f
Ge et. al. [16]	x	x		x	x	p														f
Arcidiacono et. al. [3]	x	x		x	x	x			x		x					f		x		
Fabbri et. al. [13]	p	p	p	p	p	f						f	f							
Heindl et. al. [20]	p		x	p	f				x		x									
Borkman et. al. [4]	x	x		x	x	x	x	x	x	x	x				f		x	x		
Param et. al. [48]			x	x	x	x						x	x			p				
Buls et. al. [6]	x			x																
Shermeyer et. al. [51]						p									x	x				
Zhang et. al. [66]	x	x		x	x	x			x	x	x								x	x
Madan et. al. [34]		f		p	f	x														
Mao et. al. [36]	x	x		x								x	x					x		f
Khirodkar et. al. [28]	x		x	x	x				x	x	x					x		x		

FOREGROUND - F1: rotation; F2:scale; F3: texture; F4: position; F5: count | BACKGROUND - B1: Texture/background image; B2: Color; B3: rotation
DISTRACTORS - D1: count; D2: rotation; D3: position | CAMERA: C1: position; C2: rotation; C3: field of view | LIGHTS - L1: count; L2: position; L3: orientation; L4: specular characteristics | NOISE - N1: type / count | BLUR - R1: type / location

planes detection for shelf/floor detection [17]. On the opposite side of spectrum, sometimes authors use same technique without any variance across all dataset samples (11.2% of all DR usage). This is frequently done in order to improve overall performance that relates to some deficiencies of neutral network design or issues with the overall DR methodology. An example of such use is application of blur to conceal pixel boundaries between foreground objects and background (e.g. Gaussian blur is used by [21] on all samples, while [10] uses random blur method from a pool of predefined ones).

Analysing the usage of those various techniques across the state of the art in Table 3 it can be observed that foreground generation is the key area where most scientists focus their efforts. 43% of all reported DR implementations are related to various forms of foreground object randomisation; followed by 15% lighting parameters; 13% distractor objects; 12% background; 10% camera; 4% blur and 1% noise. Likewise, looking at each individual article separately, the coverage of techniques in foreground group is greatest: on average 3 out of 5 technique types in foreground group are used in each reviewed article (STD = 0.9), while for all other groups 1 or less. This trend looks similar if we segment separately simple randomisation techniques as well as procedural randomisation.

Apart of the focus on foreground generation, several articles stand out by covering a particularly wide spectrum of techniques across all groups. [47] experiments with 16 out of 20 defined techniques as they propose a simulator like approach with the entire 3D scene constructed from scratch. This requires a lot of attention to detail and catering for many variables to achieve realism. As a result, the same article also leads in amount of procedural DR use with its key contribution of structured

randomisation, ie. usage of rules per each parameter for generation of scene. In comparison, the remaining articles that put emphasis on amount of techniques go in a different direction mostly opting for simple randomisation. [22] and [59] abandon effort for realistic scenes and put emphasis on scene variance in the background and distractor layer; [12] similar as [47] opt for realism but achieve it with improving rendering quality rather than scene organisation. Comparing the statistics for those works as summarised in previous sections, it can be seen that usage of big amount of DR techniques is not always followed by exhaustive evaluation or introduction of new innovations to DR. Notably, [4] provide one of the most comprehensive implementations in the state of the art but do not propose any novelty; [66] as well as [12] also cover a wide range of techniques but provide very small scale evaluations of their proposed innovations. Furthermore, as shown in next section, the amount of DR techniques used does not necessarily go hand in hand with performance gains.

6 PERFORMANCES COMPARISON

The reported performances in Domain Randomisation publications are frequently tested in different conditions. The spectrum of application areas in the state of the art is quite wide: detected objects have varying characteristics (e.g. size, visual complexity); captured scenes may have different lighting conditions; different backgrounds etc. Likewise, the evaluation datasets have different distributions, sizes; and are input into different state of the art object detection algorithms. The solutions described in this survey span more than 7 years between 2015 til present - a considerable time in terms of object detection neural network design and significant performance hikes as reported by [67]. In this context, multiple DR authors show that different networks vary in performance on same synthetic training and test sets [13, 59]. Furthermore, results from [17] show that combination of network architecture and detected object type plays a role and performance can very a lot more than one might expect extrapolating from a network trained on real data. All those factors make it hard to craft a scientifically credible comparison of subsequent DR contributions; and assess if they deliver performance gains universally or just in their own narrow context reported by authors. At the same time, most articles do not disclose implementation details and their code is not publicly available.

For those reasons, we abandon comparing individual solutions to each other in search for recommendation of the article with "best" synthetic data generation approach. Instead, we propose to compare relative performance gains achieved by articles within their own areas and use this to only derive high-level conclusions about the state of the art performances and observe emerging trends. Using our earlier classifications of articles and DR techniques, we attempt to connect repeating performance trends across articles with some broader categories of DR solutions that they utilise. In order to devise means for such analysis, we propose to look at the relationship between synthetic data performance to real dataset baseline that each article reports. In specific, we define two metrics that show relative gain made by new datasets expressed as fraction of baseline performance:

$$SYNTH_{DIFF} = \frac{mAP@0.5_{REAL} - mAP@0.5_{SYNTH}}{mAP@0.5_{REAL}} \tag{1}$$

$$SYNTH + REAL_{DIFF} = \frac{mAP@0.5_{REAL} - mAP@0.5_{SYNTH+REAL}}{mAP@0.5_{REAL}} \tag{2}$$

In above equations, all input performances are expressed using a popular metric for object detection that is reported in all analysed DR papers for all experiments - mean Average Precision (mAP) at Intersection over Union (IoU) threshold equal to 0.5 [43]. The newly defined metrics allow to partially abstract from the dataset context and understand the magnitude of gains made by

different algorithms expressed in a more universal manner than raw mAP metrics - ie. the starting baseline mAP for each article is different, therefore using relative measure we can abstract from that to tell which articles and what techniques make significant progress in their respective application areas. The *SYNTH* and *SYNTH+REAL* notations relate to experiments that DR authors typically conduct in order to asses the performance of their solution - measurement of object detection performance on real, human annotated test dataset using a model trained on: (1) exclusively synthetic data; (2) mix of real and synthetic data. The first experiment typically aims to show if synthetic data can fully replace real human annotated datasets; the second if synthetic data can boost the regular performance and if it can be used to decrease the amount of human annotation labour rather than fully eliminate it.

In our methodology, for each article we capture the single best result reported; we ignore all intermediate experiments and results on datasets that give lesser gain. (in case of articles that publish experiments on several datasets). E.g. [60] test on GMU Kitchen and Pascal VOC but their improvement for Pascal is only 0.75%, while for GMU 4% therefore we report experiments on GMU for that article. The complete set of all results can be found in Table 4.

Among the reviewed work, only 5 out of 20 articles report gain in mAP for synthetic trained model vs. real data trained model. However, out of those five cases, only [22] proposes innovation in terms of DR techniques and methodology, other top performant publications achieve their results through different means. All of those contributions use 3D models and just one advocates the realism approach [7]. Two articles try to solve market item detection [22, 62], two broadly understood industrial object detection [13, 20] and one relates to overhead imaginary analysis for detection of airplanes [7]. Out of those, [62] reports the biggest performance gain (31%) by recreating 3D objects from regular photos using photogrammetry. [22] follow with 17% increase of synthetic trained model over real data which they attribute to a method that distributes evenly the DR parameter values. According to [7] their performance gains (4.1%) can be achieved by splitting detection and classification, however contrary to predecessors the details of their test sets are not revealed. [13] achieve their 3.94% gain over baseline by experimenting with dataset diversity arguing that it's more important than scale. Among other works that do not report exceeding baseline performance, those between 1st and 3rd quantile score 1%-25% decrease in performance in comparison to real data trained models.

On the lowest performant side of the spectrum, the detection accuracy is reported to go as low as 60% or 90% decrease (respectively [44] and [35]). Looking at experiments with mix of real and synthetic data used for training, the results seem a lot more consistent across the state of the art with 77% articles reporting synthetic data giving edge over baseline. The magnitude of such increase is typically several percent (ie. less than 10% performance increase for 65% of the aforementioned cases); with some exceptions in 16-32% performance increase range.

In order to discover which techniques attribute to biggest mAP gain, DR articles perform an ablation study - re-train the same model on different synthetic data that was created with removal of one technique per experiment. This way it's possible to note the impact of parameter absence on the overall performance. The articles that deliver such studies and their results have been summarised in Table 5.

The analysis of results shows that majority of impactful techniques is located in foreground area (9 out of 12 studies). Within this group, technique giving highest relative gains in comparison to baseline is reported by [22]. In that study, a performance hit of 75% relates to replacing randomised amount of objects inserted into synthetic image (up to 20 objects); compared to a routine that generates all training data with single object per image. [48] confirm the usefulness of object count randomisation by testing synthetic data performance changes with 10-400 distinct objects, however they also show that there is a limit to performance increase albeit without explanation to it's

Table 4. Algorithm performance per article. Original performance measures (REAL, SYNTH, REAL+SYNTH) all refer to mAP@0.5 metric [43]; metric values are taken directly from corresponding articles as reported by their authors. Four last columns (marked with "DIFF") depict performance gain/hit connected to using synthetic data*. Performance gains are marked with bold.

ARTICLE	REAL	SYNTH	REAL+ SYNTH	SYNTH DIFF	SYNTH DIFF(%)	REAL+ SYNTH DIFF	REAL+ SYNTH DIFF(%)
Peng et al. [44]	0.58	0.19	0.31	-0.39	-67.24	-0.27	-46.55
Dwibedi et al. [11]	0.86	0.76	0.89	-0.10	-11.70	**+0.02**	**+2.89**
Mitash et al. [38]	0.75	0.70	n/a	-0.05	-6.66	-0.1	-13.51
Georgakis et al. [17]	0.83	0.52	0.85	-0.39	-37.40	**+0.02**	**+3.03**
Hinterstoisser at al. [21]	0.99	0.97	n/a	-0.01	-1.01	n/a	n/a
Dvornik et al. [10]	0.58	n/a	0.62	n/a	n/a	**+0.04**	**+6.89**
Borrego at al. [5]	0.66	0.58	0.83	-0.83	-12.03	**+0.17**	**+26.08**
Tremblay et al. [59]	0.96	0.78	0.98	-0.18	-18.98	**+0.02**	**+2.17**
Wrenninge et. al. [63]	0.54	0.34	0.90	-0.20	-36.64	-0.20	-36.64
Hinterstoisser et al. [22]	0.76	0.89	n/a	**+0.13**	**+17.10**	n/a	n/a
Mahmood et. al. [35]	0.21	0.02	0.49	-0.19	-90.40	**+0.28**	**+133**
Prakash et. al. [47]	0.85	0.77	0.89	-0.08	-9.41	**+0.04**	**+4.7**
Wong et. al. [62]	0.64	0.95	n/a	**+0.31**	**+32.67**	**+0.31**	**+32.67**
Tripathi et. al. [60]	0.86	n/a	0.89	n/a	n/a	**+0.03**	**+4.05**
Kar et. al. [27]	n/a	0.66	n/a	n/a	n/a	n/a	n/a
Chengsong et. al. [24]	0.51	0.46	n/a	-0.13	-25.50	n/a	n/a
Hughes et. al. [26]	0.89	0.42	0.90	-0.47	53.20	**+0.01**	**+1.57**
Eversberg et. al. [12]	0.99	0.99	1.00	0.00	00.00	**+0.01**	**+1.52**
Clement et. al. [7]	0.73	0.76	0.68	**+0.03**	**+4.10**	-0.05	-6.9
Yun et. al. [65]	0.86	0.81	n/a	-0.05	-5.81	n/a	n/a
Ge et. al. [16]	0.45	0.43	0.52	-0.02	-4.44	**+0.07**	**+15.5**
Arcidiacono et. al. [3]	0.85	0.80	0.91	-0.05	-5.98	**+0.06**	**+1.52**
Lidberg et. al. [30]	0.90	0.74	0.94	-0.16	-17.77	**+0.04**	**+4.47**
Nowruzi et. al. [41]	0.39	n/a	0.43	n/a	n/a	**+0.04**	**+8.97**
Fabbri et. al. [13]	0.76	0.79	n/a	**+0.03**	**+3.94**	n/a	n/a
Heindl et. al. [20]	0.72	0.74	n/a	**+0.02**	**+2.77**	n/a	n/a
Borkman et. al. [4]	0.72	0.54	0.85	-0.15	-25.00	**+0.13**	**+18.78**
Param et. al. [48]	0.28	0.24	0.36	-0.04	-14.20	**+0.08**	**+28.57**
Shermeyer et. al. [51]	0.97	0.87	0.95	-0.10	-10.12	-0.01	-1.54
Madan et. al. [34]	0.66	0.44	n/a	-0.22	-33.00	n/a	n/a
Mao et. al. [36]	0.34	n/a	0.40	n/a	n/a	**+0.05**	**+16.08**
Khirodkar et. al. [28]	0.99	0.99	n/a	0.00	00.00	n/a	n/a

* Difference between algorithm performance ran on real vs. synthetic dataset, expressed as percent of real dataset performance.

relationship with test set distribution. [24] also examines this technique and shows its benefits but concludes background colour adjustments as more useful. Amongst other studies, particularly interesting is foreground object positioning as it has been confirmed as key by multiple authors in various settings of realistic scene composition: [38] using physics engine, [17] contributing horizontal surface detection and [10] using neural networks to predict most likely object location. In

Table 5. Domain Randomisation ablation studies per article. Values indicate drop in mAP@0.5 performance when removing the selected technique (expressed as fraction of synthetic performance with all techniques; zero value indicates technique not used in study). Most impactful techniques per article marked with bold.

ARTICLE	F1	F2	F3	F4	F5	B1	B2	B3	D1	D3	L1	L2	L4	N1	R1
Peng et al. [44]	**.18**	0	**.18**	0	0	.07	0	0	0	0	0	0	0	0	0
Dwibedi et al. [11]	.10	0	0	**.16**	0	0	0	0	.03	0	0	0	0	0	.13
Mitash et al. [38]	**.55**	0	0	**.55**	0	0	0	0	0	0	0	.08	.08	0	0
Georgakis et al. [17]	0	**.15**	0	**.15**	0	0	0	0	0	0	0	0	0	0	.06
Dvornik et al. [10]	0	0	0	**.16**	0	0	0	0	0	0	0	0	0	0	0
Borrego at al. [5]	0	0	**.62**	0	0	0	0	0	0	0	0	0	0	0	0
Tremblay et al. [59]	0	0	.06	0	0	0	0	0	.01	0	**.08**	0	0	0	0
Hinterstoisser et al. [22]	.29	.29	0	0	**.75**	.50	.04	.49	0	0	.07	0	.15	0	.10
Prakash et. al. [47]	0	0	**.16**	0	0	.13	0	0	0	.05	0	0	.04	0	0
Chengsong et. al. [24]	0	0	0	0	.23	.02	**.47**	0	0	0	0	0	0	0	0
Param et. al. [48]	0	0	0	**.58**	0	0	0	0	0	0	0	0	0	0	0
Madan et. al. [34]	0	0	0	.26	0	**.76**	0	0	0	0	0	0	0	0	0

FOREGROUND - F1: rotation; F2:scale; F3: texture; F4: position; F5: count
BACKGROUND - B1: Texture/background image; B2: Color; B3: rotation
DISTRACTORS - D1: count; D2: rotation; D3: position | CAMERA: C1: position; C2: rotation; C3: field of view
LIGHTS - L1: count; L2: position; L3: orientation; L4: specular characteristics
NOISE - N1: type / count | BLUR - R1: type / location

the next section, we plot those and other discoveries of Domain Randomisation area onto timeline to show how progress was made by subsequent publications; followed by conclusions section where we try to extrapolate to give advice on domain future and best practices for current use of existing techniques.

7 DOMAIN RANDOMISATION RESEARCH ACTIVITY AND PROGRESS ACROSS YEARS

Within the subdomain of Domain Randomisation described in this review, the amount of published articles has been steadily growing across the years (see Figure 3). However, the amount of new DR techniques and categories pioneered by their inventors peak in year 2018 (as per categorisation introduced in Section 5). Up until that year, 70% of the total described techniques were proposed. Breaking this down by types of DR techniques, we can observe that the early years focused more on innovating simple value randomisation within a preset range, with more procedural contributions in 2018 and 2019. Throughout the analysed period, each year we can see a similar level of use for fixed value methodologies. As a whole, we see this trend reflecting the maturing complexity of proposed solutions: original contributions delivered simple to implement variable randomisations but pioneered use of all Domain Randomisation categories and techniques; the later innovators often propose more sophisticated ways of generating synthetic data but within those earlier defined technique categories and boundries. The stable presence of fixed value methods is related to the fact the often authors focus only on certain range of randomisation techniques and others are simplified.

This quantitive analysis of new DR techniques across the years does not however tell the full story of Domain Randomisation research progress. While initial publications have focused on technically simpler challenges, there were important publications throughout the analysed period with valuable contributions in the early days as well as later years, respective of the domain maturity

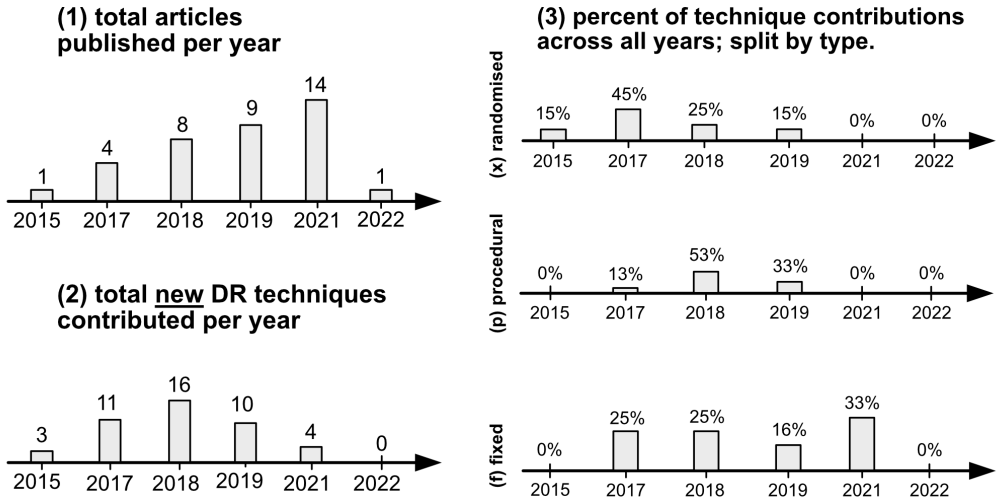

Fig. 3. Chart (1) shows total articles analysed in our review and grouped per publishing year; (2) depicts count of new techniques contributed in the State of the Art (SoA) for the first time, aggregated per calendar year; charts (3) present SoA percentage of original contributions broken down per technique type/complexity across years.

(see Figure 4). The first use of Domain Randomisation registered by us comes from Peng et al. [44] in 2015 with experiments involving simple 3D objects. The breakthrough in Peng work is related to usage of synthetic data for training of Convolutional Neural Network as they were just gaining popularity at that time; the prior works on synthetic data generation from Peng's perspective related to legacy approaches like HOG (outside of scope of our interest). Later, the next major milestone was work by Dwibedi [11] who established cut-paste technique for the first time and added blending (which has been reused by most publications ever since). Within the same year we can observe some first experiments with photo-realism and evaluations on how it helps with training on synthetic data. Those publications opened the debate between realism and non-realism in future research. In 2018 a major milestone was addition of distractor objects heavily utilised in most future works. Likewise work by Hinterstoisser [21] albeit simple in principle was first to take into consideration not only generation of data but also adjusting the neural network to achieve better results with such data. Later, in 2019 we can see two major milestones that add a lot more complexity into the data generation process: (1) structured DR and (2) end-to-end training joint with generation. The final years of state of the art analysed by us are more predominant of applied research work taking advantage of those early innovations rather than proposing some major new additions.

This shows that currently the Domain Randomisation area does not attract as big interest of breakthrough innovators as it used to at it's conception (see Figure 4). However, it is clear that key problems remain unsolved therefore there is still potential for new innovations and multiple future directions. Some of those we describe in the next section.

8 CONCLUSIONS AND FUTURE WORK DIRECTIONS

The common goal for all reviewed Domain Randomisation research is to construct synthetic datasets such that would match or even exceed performance that can be achieved with manually labelled data. The results of our analysis across the current state of the art show this is not yet the case.

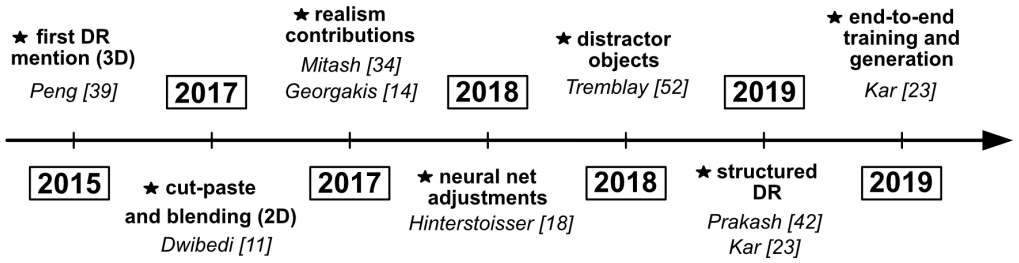

Fig. 4. Key milestones in Domain Randomisation research as observed across the past years.

Several articles demonstrate promising results but the majority of published work stays in the 1-25% performance decrease bracket in comparison to models trained on manually labelled datasets. We anticipate the future work to follow the current trend to increase model accuracy using minimal real data accompanied with synthetic samples rather than fully replacing real data (as depicted in performance results in Table 4).

Extrapolating from the publication count so far, we can observe that Domain Randomisation research seems to be shifting towards 3D models. Our study shows that those 3D approaches involve more variables in comparison to 2D (e.g. lighting, camera, textures as shown in Table 3). This increased complexity could potentially be the trigger for broader spectrum of current and future 3D-based research contributions. Additionally, we notice that there is a relationship between modelling approach and popularity of application areas. A strong representation of articles from manufacturing area has better access to CAD models (industry objects detection); likewise autonomous driving area that traditionally relies on simulators and has a long history of 3D modelling use. We also noted a substantial amount of robotics related articles that utilise synthetic data with 3D modelling for policy training [46], position estimation [57] or grasping [58]; however those frequently report different evaluation metrics, therefore were not included in this study.

Analysing all those articles and future work that they explicitly propose, we can see a consensus among authors: many mention expanding their innovative proposals to other areas of vision than object detection, especially towards image segmentation. Otherwise, other popular proposals for future work advocate increasing synthetic scene complexity, better photorealism and general effort towards Domain Adaptation [42] as supplementary to Domain Randomisation. Some of the work in Domain Adaptation as well as synthetic data for image segmentation has been already done, however its detailed analysis is outside of the scope of this review.

Regarding performance of competitive high level DR technics in object detection (such as 2D vs. 3D, realistic vs non-realistic etc.), our conclusions are in line with past state of the art overviews [39, 61] - the debate between specific key technical choices is still ongoing. While, some simple studies show the advantage of 3D based synthetic data over 2D, this does not find reflection in our analysis of performances from all articles across state of the art. We did not find any significant correlation between foreground/background modelling method and performance gains. Likewise, the same can be said about realism vs non-realism approaches as means to close the domain gap.

In terms of innovations proposed so far since the conception of DR domain; we can see that initial ideas of simple randomisation have evolved into more sophisticated proposals that cater not only for data generation but alter the network used for training and in some cases object detection. Albeit this has had mixed results, e.g. jointly generating data and training the object detection model proposed by [60] shows little performance boost. However, in academic terms, this opens

room for more innovation via research in alternative neural network architectures that would support synthetic data training better.

For a practitioner with a need to implement a synthetic data generation pipeline, our recommendation is to start with 2D cut-paste techniques to establish a baseline as this approach requires least effort and can deliver good results (as shown in [11]). For those practitioners that have more substantial resources, experimenting with 3D scanning, modelling and scene composition along with increased photorealism are good directions to follow. Applying the innovations in neural network architecture for increasing synthetic data performance is the most complex and risky task out of all reviewed. As shown in our article, specific neural network architectures evaluated in state of the art are usually limited to a single publication without much continuation in later work, therefore as this point it is hard to assess how such implementations would perform in different application domains and data scenarios.

REFERENCES

[1] Raghad Alghonaim and Edward Johns. 2021. Benchmarking Domain Randomisation for Visual Sim-to-Real Transfer. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. 12802–12808. https://doi.org/10.1109/ICRA48506.2021.9561134

[2] Mohammad Ani, Hector Basevi, and Aleš Leonardis. 2021. Quantifying the Use of Domain Randomization. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 6128–6135. https://doi.org/10.1109/ICPR48806.2021.9412118

[3] Claudio Salvatore Arcidiacono. 2018. *An Empirical Study on Synthetic Image Generation Techniques for Object Detectors*. Ph. D. Dissertation. KTH Royal Institute of Technology.

[4] Steve Borkman, Adam Crespi, Saurav Dhakad, Sujoy Ganguly, Jonathan Hogins, You-Cyuan Jhang, Mohsen Kamalzadeh, Bowen Li, Steven Leal, Pete Parisi, Cesar Romero, Wesley Smith, Alex Thaman, Samuel Warren, and Nupur Yadav. 2021. Unity Perception: Generate Synthetic Data for Computer Vision. *ArXiv* abs/2107.04259 (2021). https://doi.org/10.48550/arXiv.2107.04259

[5] João Borrego, Atabak Dehban, Rui Figueiredo, Plinio Moreno, Alexandre Bernardino, and José Santos-Victor. 2018. Applying Domain Randomization to Synthetic Data for Object Category Detection. *ArXiv* abs/1807.09834 (2018). https://doi.org/10.48550/arXiv.1807.09834

[6] Elvijs Buls, Roberts Kadikis, Ričards Cacurs, and Jānis Ārents. 2019. Generation of Synthetic Training Data for Object Detection in Piles. In *Eleventh International Conference on Machine Vision (ICMV 2018)*, Antanas Verikas, Dmitry P. Nikolaev, Petia Radeva, and Jianhong Zhou (Eds.), Vol. 11041. International Society for Optics and Photonics, SPIE, 533 – 540. https://doi.org/10.1117/12.2523203

[7] Nathan Clement, Alan Schoen, Arnold Boedihardjo, and Andrew Jenkins. 2021. Synthetic Data and Hierarchical Object Detection in Overhead Imagery. *ArXiv* abs/2102.00103 (2021). https://doi.org/10.48550/arXiv.2102.00103

[8] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. 886–893. https://doi.org/10.1109/CVPR.2005.177

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[10] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. 2018. Modeling Visual Context Is Key to Augmenting Object Detection Datasets. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 375–391. https://doi.org/10.1007/978-3-030-01258-8_23

[11] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. 2017. Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 1310–1319. https://doi.org/10.1109/ICCV.2017.146

[12] Leon Eversberg and Jens Lambrecht. 2021. Generating Images with Physics-Based Rendering for an Industrial Object Detection Task: Realism versus Domain Randomization. *Sensors* 21, 23 (2021). https://doi.org/10.3390/s21237901

[13] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Aljoša Ošep, Riccardo Gasparini, Orcun Cetintas, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. 2021. MOTSynth: How Can Synthetic Data Help Pedestrian Detection and Tracking?. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 10829–10839. https://doi.org/10.1109/ICCV48922.2021.01067

[14] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A Survey of Data Augmentation Approaches for NLP. In *Findings of the Association for Computational Linguistics:*

ACL-IJCNLP 2021. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.84

[15] Joao Fonseca and Fernando Bacao. 2023. Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data* 10, 115 (2023). https://doi.org/10.1186/s40537-023-00792-7

[16] Yunhao Ge, Jiashu Xu, Brian Nlong Zhao, Neel Joshi, Laurent Itti, and Vibhav Vineet. 2022. Beyond Generation: Harnessing Text to Image Models for Object Detection and Segmentation. *ArXiv* abs/1409.1556 (12 2022). https://doi.org/10.48550/arXiv.2206.09592

[17] Georgios Georgakis, Arsalan Mousavian, Alexander Berg, and Jana Kosecka. 2017. Synthesizing Training Data for Object Detection in Indoor Scenes. In *Proceedings of Robotics: Science and Systems*. Cambridge, Massachusetts. https://doi.org/10.15607/RSS.2017.XIII.043

[18] Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. 2019. Precise Detection in Densely Packed Scenes. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5222–5231. https://doi.org/10.1109/CVPR.2019.00537

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[20] Christoph Heindl, Lukas Brunner, Sebastian Zambal, and Josef Scharinger. 2021. BlendTorch: A Real-Time, Adaptive Domain Randomization Library. In *Pattern Recognition. ICPR International Workshops and Challenges*. Springer International Publishing, Cham, 538–551. https://doi.org/10.1007/978-3-030-68799-1_39

[21] Stefan Hinterstoisser, Vincent Lepetit, Paul Wohlhart, and Kurt Konolige. 2019. On Pre-Trained Image Features and Synthetic Images for Deep Learning. In *Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part I* (Munich, Germany). Springer-Verlag, Berlin, Heidelberg, 682–697. https://doi.org/10.1007/978-3-030-11009-3_42

[22] Stefan Hinterstoißer, Olivier Pauly, Tim Hauke Heibel, Martina Marek, and Martin Bokeloh. 2019. An Annotation Saved is an Annotation Earned: Using Fully Synthetic Training for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2787–2796. https://doi.org/10.1109/ICCVW.2019.00340

[23] Tomáš Hodaň, Vibhav Vineet, Ran Gal, Emanuel Shalev, Jon Hanzelka, Treb Connell, Pedro Urbina, Sudipta N. Sinha, and Brian Guenter. 2019. Photorealistic Image Synthesis for Object Instance Detection. In *2019 IEEE International Conference on Image Processing (ICIP)*. 66–70. https://doi.org/10.1109/ICIP.2019.8803821

[24] Chengsong Hu, J. Alex Thomasson, and Muthukumar V. Bagavathiannan. 2021. A powerful Image Synthesis and Semi-Supervised Learning Pipeline for Site-Specific Weed Detection. *Computers and Electronics in Agriculture* 190 (2021). https://doi.org/10.1016/j.compag.2021.106423

[25] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal Unsupervised Image-to-image Translation. In *ECCV*. 179–196.

[26] David Hughes and Hao Ji. 2021. Enhancing Object Detection Using Synthetic Examples. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. 1398–1402. https://doi.org/10.1109/CCWC51732.2021.9376062

[27] Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. 2019. Meta-Sim: Learning to Generate Synthetic Datasets. In *2019 IEEE International Conference on Computer Vision (ICCV)*. 4550–4559. https://doi.org/10.1109/ICCV.2019.00465

[28] Rawal Khirodkar, Donghyun Yoo, and Kris M. Kitani. 2019. Domain Randomization for Scene-Specific Car Detection and Pose Estimation. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019*. IEEE, Waikoloa Village, HI, USA, 1932–1940. https://doi.org/10.1109/WACV.2019.00210

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* 60, 6 (may 2017), 84–90. https://doi.org/10.1145/3065386

[30] Love Lidberg. 2018. *Object Detection using deep learning and synthetic data*. Master's thesis. Linkoping University.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*. Springer International Publishing, Cham, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[32] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016*. Springer International Publishing, Cham, 21–37. https://doi.org/0.1007/978-3-319-46448-0_2

[33] David G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 2. 1150–1157. https://doi.org/10.1109/ICCV.1999.790410

[34] Spandan Madan, Zoya Bylinskii, Carolina Nobre, Matthew Tancik, Adria Recasens, Kimberli Zhong, Sami Alsheikh, Aude Oliva, Fredo Durand, and Hanspeter Pfister. 2021. Parsing and Summarizing Infographics with Synthetically Trained Icon Detection. In *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*. 31–40. https://doi.org/10.1109/PacificVis52677.2021.00012

[35] Ammar Mahmood, Mohammed Bennamoun, Senjian An, Ferdous Sohel, Farid Boussaid, Renae Hovey, and Gary Kendrick. 2019. Automatic Detection of Western Rock Lobster Using Synthetic Data. *ICES Journal of Marine Science* 77, 4 (11 2019), 1308–1317. https://doi.org/10.1093/icesjms/fsz223

[36] Xin Mao, Jun Kang Chow, Pin Siang Tan, Kuan fu Liu, Jimmy Wu, Zhaoyu Su, Ye Hur Cheong, Ghee Leng Ooi, Chun Chiu Pang, and Yu-Hsing Wang. 2021. Domain Randomization-Enhanced Deep Learning Models for Bird Detection. *Scientific Reports* 11, 639 (Jan 2021). https://doi.org/10.1038/s41598-020-80101-x

[37] Venkat Margapuri and Mitchell Neilsen. 2021. Seed Phenotyping on Neural Networks using Domain Randomization and Transfer Learning. *ArXiv* abs/2012.13259 (2021). https://doi.org/10.48550/arXiv.2012.13259

[38] Chaitanya Mitash, Kostas E. Bekris, and Abdeslam Boularias. 2017. A Self-supervised Learning System for Object Detection using Physics Simulation and Multi-view Pose Estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vancouver, Canada, 545–551. https://doi.org/10.1109/IROS.2017.8202206

[39] Sergey I. Nikolenko. 2021. *Synthetic Data for Deep Learning*. Springer Cham. https://doi.org/10.1007/978-3-030-75178-4

[40] Fernando Camaro Nogues, Andrew Huie, and Sakyasingha Dasgupta. 2018. Object Detection using Domain Randomization and Generative Adversarial Refinement of Synthetic Images. *ArXiv* abs/1805.11778 (2018). https://doi.org/10.48550/arXiv.1805.11778

[41] Farzan Erlik Nowruzi, Prince Kapoor, Dhanvin Kolhatkar, Fahed Al Hassanat, Robert Laganiere, and Julien Rebut. 2019. How Much Real Data do we Actually Need: Analyzing Object Detection Performance Using Synthetic and Real Data. *ArXiv* abs/1907.07061 (2019). https://doi.org/10.48550/arXiv.1907.07061

[42] Poojan Oza, Vishwanath A. Sindagi, Vibashan Vishnukumar Sharmini, and Vishal M. Patel. 2023. Unsupervised Domain Adaptation of Object Detectors: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023), 1–24. https://doi.org/10.1109/TPAMI.2022.3217046

[43] Rafael Padilla, Sergio L. Netto, and Eduardo A. B. da Silva. 2020. A Survey on Performance Metrics for Object-Detection Algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. 237–242. https://doi.org/10.1109/IWSSIP48289.2020.9145130

[44] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. 2015. Learning Deep Object Detectors from 3D Models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 1278–1286. https://doi.org/10.1109/ICCV.2015.151

[45] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. 2018. VisDA: A Synthetic-to-Real Benchmark for Visual Domain Adaptation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2102–21025. https://doi.org/10.1109/CVPRW.2018.00271

[46] Riccardo Polvara, Massimiliano Patacchiola, Marc Hanheide, and Gerhard Neumann. 2020. Sim-to-Real Quadrotor Landing via Sequential Deep Q-Networks and Domain Randomization. *Robotics* 9, 1 (2020). https://doi.org/10.3390/robotics9010008

[47] Aayush Prakash, Shaad Boochoon, Mark Brophy, David Acuna, Eric Cameracci, Gavriel State, Omer Shapira, and Stan Birchfield. 2019. Structured Domain Randomization: Bridging the Reality Gap by Context-Aware Synthetic Data. In *2019 International Conference on Robotics and Automation (ICRA)*. 7249–7255. https://doi.org/10.1109/ICRA.2019.8794443

[48] Param S. Rajpura, Ravi S. Hegde, and Hristo Bojinov. 2017. Object Detection Using Deep CNNs Trained on Synthetic Images. *ArXiv* abs/1706.06782 (2017). https://doi.org/10.48550/ARXIV.1706.06782

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc., Cambridge, MA, USA, 91–99.

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

[51] Jacob Shermeyer, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, and Daeil Kim. 2021. RarePlanes: Synthetic Data Takes Flight. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE Computer Society, Los Alamitos, CA, USA, 207–217. https://doi.org/10.1109/WACV48630.2021.00025

[52] Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6, 60 (07 2019), 2196–1115. https://doi.org/10.1186/s40537-019-0197-0

[53] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv* abs/1409.1556 (09 2014). https://doi.org/10.48550/arXiv.1409.1556

[54] Arjun Singh, James Sha, Karthik S. Narayan, Tudor Achim, and P. Abbeel. 2014. BigBIRD: A Large-Scale 3D Database of Object Instances. *2014 IEEE International Conference on Robotics and Automation (ICRA)* (2014), 509–516. https://doi.org/10.1109/ICRA.2014.6906903

[55] Michael Stark, Michael Goesele, and Bernt Schiele. 2010. Back to the Future: Learning Shape Models from 3D CAD Data. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 106.1–106.11. https://doi.org/10.5244/C.24.106

[56] Baochen Sun and Kate Saenko. 2014. From Virtual to Reality: Fast Adaptation of Virtual Object Detectors to Real Domains. In *Proceedings of the British Machine Vision Conference*. BMVA Press. https://doi.org/10.5244/C.28.82

[57] Stefan Thalhammer, Timothy Patten, and Markus Vincze. 2019. SyDPose: Object Detection and Pose Estimation in Cluttered Real-World Depth Images Trained using Only Synthetic Data. In *2019 International Conference on 3D Vision (3DV)*. 106–115. https://doi.org/10.1109/3DV.2019.00021

[58] Josh Tobin, Lukas Biewald, Rocky Duan, Marcin Andrychowicz, Ankur Handa, Vikash Kumar, Bob McGrew, Alex Ray, Jonas Schneider, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. 2018. Domain Randomization and Generative Models for Robotic Grasping. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 3482–3489. https://doi.org/10.1109/IROS.2018.8593933

[59] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. 2018. Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE Computer Society, Los Alamitos, CA, USA, 1082–10828. https://doi.org/10.1109/CVPRW.2018.00143

[60] Shashank Tripathi, Siddhartha Chandra, Amit Agrawal, Ambrish Tyagi, James M. Rehg, and Visesh Chari. 2019. Learning to Generate Synthetic Data via Compositing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 461–470. https://doi.org/10.1109/CVPR.2019.00055

[61] Apostolia Tsirikoglou, Gabriel Eilertsen, and Jonas Unger. 2020. A Survey of Image Synthesis Methods for Visual Machine Learning. *Computer Graphics Forum* 39, 6 (2020), 426–451. https://doi.org/10.1111/cgf.14047

[62] Matthew Z. Wong, Kiyohito Kunii, Max Baylis, Wai Hong Ong, Pavel Kroupa, and Swen Koller. 2019. Synthetic Dataset Generation for Object-to-model Deep Learning in Industrial Applications. *PeerJ Computer Science* 5:e222 (Oct 2019). https://doi.org/10.7717/peerj-cs.222

[63] Magnus Wrenninge and Jonas Unger. 2018. Synscapes: A Photorealistic Synthetic Dataset for Street Scene Parsing. *ArXiv* abs/1810.08705 (2018). https://doi.org/10.48550/arXiv.1810.08705

[64] Mingle Xu, Sook Yoon, Alvaro Fuentes, and Dong Sun Park. 2023. A Survey of Image Augmentation Techniques for Deep Learning. *Pattern Recognition* 137 (2023), 109347. https://doi.org/10.1016/j.patcog.2023.109347

[65] Woo-han Yun, Taewoo Kim, Jaeyeon Lee, Jaehong Kim, and Junmo Kim. 2021. Cut-and-Paste Dataset Generation for Balancing Domain Gaps in Object Instance Detection. *ArXiv* abs/1409.1556 (01 2021). https://doi.org/10.48550/arXiv.1909.11972

[66] Li Zhang, YanZhao Ren, Sha Tao, Jingdun Jia, and Wanlin Gao. 2021. High-Quality Coarse-to-Fine Fruit Detector for Harvesting Robot in Open Environment. *KSII Transactions on Internet and Information Systems* 15, 2 (February 2021), 421–441. https://doi.org/10.3837/tiis.2021.02.003

[67] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2019. Object Detection in 20 Years: A Survey. *ArXiv* abs/1905.05055 (2019). https://doi.org/10.48550/arXiv.1905.05055