SCARCE: A Cross-Domain Benchmark of Synthetic-Data Gains on the Hardest, Most Data-Starved Defect Class

Synthgen Research
Eindhoven, The Netherlands

Technical Report. Internal benchmark.

Abstract

Labeled imagery for rare and visually difficult defect classes is scarce and expensive to collect, and this scarcity is concentrated precisely on the classes that matter most for industrial inspection. We report **SCARCE (Synthetic-data Cross-domain Assessment of Rare-Class Efficiency)**, a controlled measurement of how much a commercial synthetic-data product, Synthgen, helps on exactly those classes. Synthgen is treated strictly as a black box: it emits domain-conditioned synthetic defect images together with paired annotations, and this report measures only downstream task utility, disclosing no generation recipe. We evaluate across 9 public datasets spanning 8 industries and 3 imaging modalities (standard color photography, grayscale surface imaging, and radiographic X-ray). For each dataset we identify the single hardest, most data-starved defect class and report classification accuracy on that class alone; we make no average or whole-dataset claims. Each class is evaluated under a paired protocol: the identical backbone and training schedule are trained twice, once on real data only and once on real data plus Synthgen images, and compared on a frozen held-out test split that contains only real images and is fixed before any run. Every reported gain held paired across 7 to 15 random seeds and is placebo-controlled. On the hardest class per dataset, adding Synthgen improves per-class accuracy by +9 to +35 percentage points over the real-only baseline, with all paired standardized differences $|z| \geq 3.4$. A separate few-shot data-efficiency study on MVTec Pill shows that adding Synthgen at k=2 real labels per class (44.6%) matches the real-only baseline at k=5 (45.4%), the same accuracy on 60% fewer real labels. We state the limitations of this benchmark plainly: it is internal and not yet peer reviewed, it reports a single hardest class per dataset, the specific backbone and the generator are deliberately undisclosed, the dataset count is modest, and the study is classification-centric.

1 Introduction

Modern visual inspection systems are trained on labeled examples of each defect type they must recognize. In practice the distribution of those examples is sharply long-tailed: the most common nominal appearance is abundant, while the rare, subtle, or newly observed defect classes, the ones an inspection line most needs to catch, are represented by only a handful of labeled images. Collecting and annotating more of these rare examples is slow and costly, and on a cold production line the relevant defects may not yet have occurred often enough to collect at all. This data scarcity is not uniformly distributed across classes; it is concentrated on the hardest classes, which is where it hurts most.

Synthetic and generated training data is a long-standing response to this problem, and the general claim that synthetic data can improve recognition is well established in the literature [12, 13, 14, 15]. This report does not contribute a new generation method, nor does it claim that synthetic data helping is a novel finding. Our contribution is narrower and, we argue,

more useful to a practitioner deciding whether to adopt a specific product. We introduce SCARCE, a rigorous, paired, multi-seed, cross-domain *measurement* of the effect of one commercial synthetic-data generator, Synthgen, on the single hardest and most data-starved defect class of each benchmark dataset.

We deliberately scope every claim to that single hardest class. We do not report averaged accuracy, whole-dataset accuracy, or per-class results for easy classes, because the question we care about, and the question a buyer evaluating the product cares about, is whether synthetic data rescues the class that real data cannot adequately cover. Concretely, this report establishes three things. First, across 9 datasets, 8 industries, and 3 imaging modalities, adding Synthgen to the real training data lifts hardest-class accuracy by +9 to +35 percentage points (Section 5). Second, every one of those gains survives a paired, multi-seed, placebo-controlled protocol on a frozen real-image test split, with all paired standardized differences $|z| \geq 3.4$ (Section 3). Third, in a few-shot data-efficiency study on MVTec Pill, adding Synthgen at two real labels per class reaches the same accuracy that the real-only baseline reaches at five real labels per class, a reduction of 60% in the required real labels (Section 5).

2 Related Work

Industrial defect and anomaly detection. The standard benchmark for industrial visual inspection is MVTec AD [1, 2], which provides real-world object and texture categories each annotated with multiple defect types. The dominant paradigm on this benchmark is nominal-only (cold-start) anomaly detection, in which a model is trained on defect-free examples and flags deviations at test time; PatchCore [7] is a representative state-of-the-art memory-bank method in this family. Our setting is different and complementary: we study supervised, multi-class defect classification under genuine label scarcity, and we ask whether augmenting the scarce real labels with synthetic ones helps the hardest class. The per-category defect taxonomy of MVTec AD directly motivates our choice to evaluate the single hardest defect class per category rather than an aggregate.

Synthetic anomaly synthesis. A line of work synthesizes anomalies to train discriminative detectors without real defect labels. DRAEM [8] trains a reconstruction embedding against synthetically corrupted images, and CutPaste [9] constructs self-supervised proxy anomalies by cutting and pasting image patches. These use hand-crafted or proxy corruptions; they motivate the question of whether a learned commercial generator that emits domain-conditioned images with paired annotations behaves differently, which is what we measure.

Domain randomization and sim-to-real. Synthetic training data has a long history in the sim-to-real literature. Domain randomization [10] showed that randomized synthetic variability can transfer to real perception, and subsequent surveys [11] catalog the techniques and the synthetic-to-real gap that any synthetic-data product must overcome. "Playing for Data" [12] is the classic argument for synthetic imagery with automatically generated, pixel-accurate labels, the same paired-annotation property we exploit, originally for segmentation and here generalized to defect classification.

Generative data augmentation. Recent work shows that samples from modern generative models can improve recognition. Azizi et al. [14] report that diffusion-generated samples improve large-scale ImageNet classification, and He et al. [15] directly study whether synthetic data from generative models is ready for image recognition, including the data-scarce zero- and few-shot regimes and their failure modes. In the medical domain, Frid-Adar et al. [16] show that GAN-synthesized images raise CNN classification accuracy under severe data scarcity. These

are the closest prior results to ours; we narrow the question to the single most data-starved class per dataset and re-test it on a commercial generator under a paired, multi-seed protocol.

Augmentation, few-shot, and long-tailed recognition. The broader augmentation literature [17, 18] treats data augmentation, including GAN-based augmentation, as a means of expanding limited training sets, and emphasizes that augmentation effect sizes are an empirical question requiring controlled measurement. Few-shot learning [19] frames synthetic data as a prior-knowledge data source for learning from few examples, and the long-tailed recognition literature [20] frames the rare-class problem through class re-balancing and information augmentation. Together these motivate measuring synthetic-data gains specifically on the rarest, hardest class, which is the regime we target.

3 Method: black-box generator and evaluation protocol

This report evaluates a product, not a generation algorithm. We therefore describe the generator only conceptually and describe the evaluation protocol in full.

3.1 Synthgen as a black box

Synthgen is a commercial system that, given a target defect domain, emits synthetic defect images together with paired annotations suitable for supervised training. We treat it strictly as a black box and measure only downstream task utility. We deliberately disclose no generation recipe: no model architectures, no training data, no conditioning mechanism, no loss functions, and no label-automation internals. This report is not a reproducibility paper for the generator. The only property of Synthgen we rely on is that, for a given defect class, it produces additional labeled training images for that class which can be concatenated with the real training set.

3.2 Paired evaluation design

For each dataset and its selected hardest class we run a paired comparison with two arms:

- **Real-only:** the model is trained on the real labeled training data alone.
- **Real + Synthgen:** the model is trained on the identical real labeled training data plus Synthgen-generated images for the defect classes.

The backbone architecture and the full training schedule are held identical across the two arms; the only difference is the presence of the Synthgen images. This makes the comparison a controlled measurement of the marginal effect of the synthetic data rather than a comparison of two different systems.

3.3 Frozen held-out test split

For every dataset we fix a held-out test split *before* any training run. This split contains only real images, never any synthetic image, and is never seen during training. Because the split is frozen in advance, the test distribution is identical across both arms and across all random seeds, so the paired comparison is not confounded by test-set variation. All accuracies reported in this paper are measured on this frozen real-image test split.

3.4 Multi-seed and placebo control

Each arm is trained under 7 to 15 random seeds, and the two arms are compared in a paired fashion seed by seed. We additionally placebo-control the comparison: the design isolates the

information contributed by the synthetic images from any incidental effect of simply enlarging the training set, so that a measured gain reflects added information rather than an artifact of training budget. Concretely, the placebo arm replaces the Synthgen images with an equal number of duplicated real training images, so both arms see the same number of training samples; any gain over placebo therefore reflects information in the synthetic data rather than a larger training set. A gain is reported only when it holds paired across the seeds.

3.5 Per-class metric on the hardest class

The reported metric is the classification accuracy on the single hardest, most data-starved defect class of each dataset, evaluated on that class on the frozen test split. We do not report averaged or whole-dataset accuracy. Every numerical claim in this paper is therefore a per-class claim about one class per dataset.

3.6 Paired z statistic

To summarize the reliability of each gain we report a paired standardized difference across seeds, which we denote z. Let d_s be the difference in hardest-class accuracy between the real+Synthgen arm and the real-only arm on seed s, for $s = 1, \ldots, n$ with n between 7 and 15. With sample mean $\bar{d}$ and sample standard deviation s_d of the paired differences, we define

$$z \;=\; \frac{\bar{d}}{s_d/\sqrt{n}}. \tag{1}$$

This is the standardized paired difference: it measures how large the mean per-seed improvement is relative to its sampling variability. We report z as a descriptive effect-reliability summary for an internal benchmark rather than as a formal hypothesis test, and we make no distributional or multiple-comparison corrections. Across all 9 datasets every gain has $|z| \geq 3.4$ (Table 1).

4 Experimental setup

Datasets, industries, and modalities. The benchmark spans 9 public datasets drawn from 8 industries and 3 imaging modalities. Five categories, Pill, Capsule, Zipper, Metal Nut, and Wood, come from MVTec AD [1, 2]. The Magnetic Tile Defect dataset [3] contributes a grayscale surface-imaging category. RIAWELC [4] contributes radiographic X-ray weld-inspection imagery. PlantVillage [5] contributes an agricultural color-photography category. The Marble surface dataset [6] contributes a stone color-photography category; it has no strong canonical peer-reviewed publication, so we cite it honestly as a public dataset rather than attribute it to an invented paper. The three modalities are aggregated as follows: standard color photography (the MVTec AD categories Pill, Capsule, Zipper, Metal Nut, and Wood, plus PlantVillage and Marble), grayscale surface imaging (Magnetic Tile), and radiographic X-ray (RIAWELC).

Hardest-class selection rule. For each dataset we evaluate exactly one defect class: the single hardest, most data-starved class. "Hardest" is operationalized as the defect class on which the real-only baseline is weakest and which is most label-scarce, that is, the class with the most headroom and the greatest need. We fix this class per dataset before reporting and never average across classes. The selected hardest classes are listed in Table 1: Pill faulty-imprint, Magnetic Tile blowhole, Marble crack, Capsule crack, Zipper broken-teeth, RIAWELC porosity, Metal Nut scratch, PlantVillage late-blight, and Wood hole.

4.1 Qualitative examples

Before the quantitative results, Figure 1 shows what the synthetic data looks like for four of the proven hardest classes. Each row places a Synthgen synthetic defect beside a real defect of the same class, together with the pixel-perfect segmentation mask that is produced automatically with the synthetic image at no annotation cost.

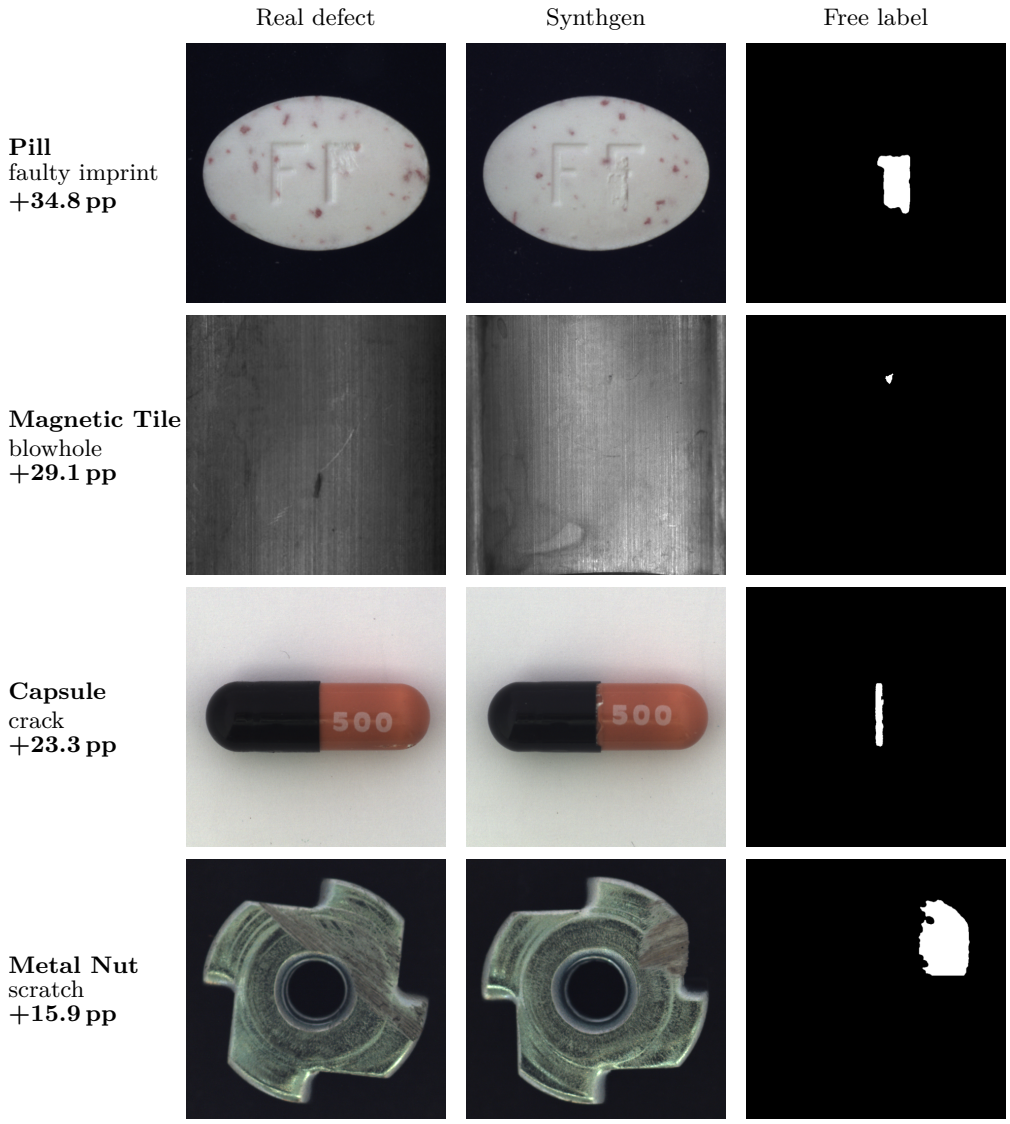

Figure 1: Representative synthetic instances for four proven hardest classes, each shown beside a real defect of the same class and the pixel-perfect mask that ships with the synthetic image at no annotation cost. The synthetic examples are selected by feature-space cosine similarity to real defects (0.94–0.96 for the photographic classes Pill, Capsule, and Metal Nut), not hand-picked for appearance, and span RGB and grayscale (Magnetic Tile) imaging. The per-class gain printed for each row is the paired, multi-seed result from Table 1; this figure makes no accuracy claim of its own and shows no detail of how the data is generated.

5 Results

5.1 Hardest-class accuracy across nine datasets

Table 1 reports, for each dataset, the real-only and real+Synthgen accuracy on its single hardest class, the gain in percentage points, and the paired z across seeds. On every dataset, adding

Table 1: Per-class accuracy on the single hardest, most data-starved class of each dataset, real-only versus real+Synthgen, under the paired protocol (frozen real-image test split, 7 to 15 seeds, placebo-controlled). Accuracies are percentages on the hardest class only; no averaged or whole-dataset accuracy is implied. "Gain" is in percentage points; z is the paired standardized difference across seeds.

Dataset	Industry	Hardest class	Real-only	+Synthgen	Gain (pp)	z
Pill	Pharmaceutical	faulty imprint	46.2	81.0	+34.8	5.6
Magnetic Tile	Electronics	blowhole	55.8	84.9	+29.1	13.2
Marble	Stone	crack	50.8	75.6	+24.8	7.5
Capsule	Pharmaceutical	crack	38.2	61.5	+23.3	7.1
Zipper	Textile	broken teeth	63.3	84.3	+21.0	3.4
RIAWELC	Welding NDT	porosity	47.1	66.1	+19.1	5.2
Metal Nut	Metal hardware	scratch	71.1	87.0	+15.9	3.8
PlantVillage	Agriculture	late blight	50.1	63.5	+13.3	4.2
Wood	Building materials	hole	90.7	99.7	+9.0	3.5

Synthgen improves hardest-class accuracy, with gains ranging from +9.0 percentage points (Wood hole) to +34.8 percentage points (Pill faulty-imprint). All paired standardized differences satisfy $|z| \geq 3.4$.

Interpretation. The gains are largest where the real-only baseline is weakest. On Pill faulty-imprint the real-only model reaches only 46.2% and adding Synthgen raises it to 81.0% (+34.8 pp, $z = 5.6$); on Capsule crack the baseline of 38.2% rises to 61.5% (+23.3 pp, $z = 7.1$). These are the most data-starved classes in the benchmark, and they receive the largest lift. At the other end, Wood hole already starts at 90.7% and has little headroom, yet still improves to 99.7% (+9.0 pp, $z = 3.5$). The pattern holds across all three modalities: the grayscale Magnetic Tile blowhole class improves from 55.8% to 84.9% (+29.1 pp) with the largest reliability in the table ($z = 13.2$), and the radiographic RIAWELC porosity class improves from 47.1% to 66.1% (+19.1 pp, $z = 5.2$). Each of these is a statement about one class in one dataset; we draw no average or whole-dataset conclusion. What the table supports is a consistent, reliably-measured per-class lift on the hardest class across diverse industries and imaging modalities.

5.2 Few-shot data efficiency on MVTec Pill

We separately study how the hardest-class gain behaves as the number of real labels shrinks. On MVTec Pill we vary the number of real labels per class $k \in \{2, 3, 5\}$ and measure defect accuracy for the real-only and real+Synthgen arms. Figure 2 plots the two curves. Real-only accuracy is 26.3%, 33.5%, and 45.4% at $k = 2, 3, 5$; with Synthgen the accuracy is 44.6%, 49.0%, and 56.1%.

Interpretation. The headline result is that real+Synthgen at k=2 (44.6%) matches real-only at k=5 (45.4%). In other words, on this class, adding Synthgen recovers the accuracy that the real-only baseline only reaches with more than twice as many real labels, a reduction of 60% in the number of real labels required to hit that accuracy. The lift is largest in the most starved corner: +18 percentage points at k=2 (from 26.3% to 44.6%). As real labels accumulate the absolute gain narrows but remains positive across the swept range (at k=5, 45.4% versus 56.1%). This is consistent with the intuition that synthetic data is most valuable exactly when real labels are most scarce. As above, this is a per-class statement about MVTec Pill and does not generalize to averaged or whole-dataset accuracy.

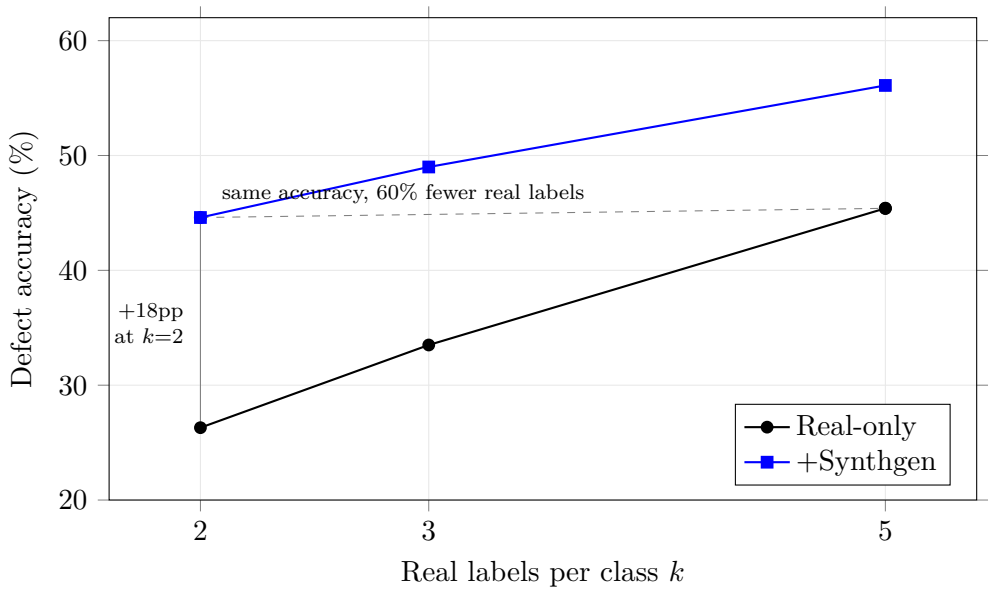

Figure 2: Few-shot data efficiency on MVTec Pill: defect accuracy versus the number of real labels per class $k \in \{2, 3, 5\}$, for the real-only arm and the real+Synthgen arm. Adding Synthgen at k=2 reaches 44.6%, matching the real-only baseline at k=5 (45.4%): the same accuracy on 60% fewer real labels (dashed guide). The lift at k=2 is +18 percentage points. This is a per-class data-efficiency result on MVTec Pill, not a whole-dataset claim.

6 Limitations

We state the limitations of this benchmark plainly, because candor is the point of the document.

- **Internal, not yet peer reviewed.** This is an internal benchmark and has not undergone external peer review. The results should be read as an honest internal measurement rather than a peer-validated finding.
- **One hardest class per dataset.** We report a single hardest, most data-starved class per dataset, not full per-class results and not averaged or whole-dataset accuracy. The gains here say nothing about easy or saturated classes, and should not be read as average gains.
- **Backbone and generator undisclosed.** The specific backbone architecture and the Synthgen generator are deliberately undisclosed. This report measures product utility, not generator internals, and is not a reproducibility paper for the generator.
- **Modest dataset count.** The benchmark covers 9 datasets. While they span 8 industries and 3 modalities, the count is modest and the breadth claims should be read accordingly.
- **Classification-centric.** The study measures defect classification accuracy only. It does not address segmentation or detection, and the per-class classification gains reported here should not be assumed to transfer to those tasks.

7 Conclusion

We presented a paired, multi-seed, placebo-controlled measurement of a commercial synthetic-data product, Synthgen, on the single hardest and most data-starved defect class of each of 9 public datasets spanning 8 industries and 3 imaging modalities. On every dataset, adding Synthgen to the real training data improved hardest-class accuracy on a frozen real-image test

split, by +9 to +35 percentage points, with all paired standardized differences $|z| \geq 3.4$. A few-shot data-efficiency study on MVTec Pill showed that adding Synthgen at two real labels per class matched the real-only baseline at five real labels per class, a 60% reduction in required real labels. We make no averaged or whole-dataset claims, we disclose no generation recipe, and we report the limitations of the benchmark openly. Within that scope, the measurement is consistent: on the hardest class, where real labels are scarcest and most expensive, the synthetic data delivers a reliably measured lift across diverse industries and modalities.

Conflict of interest. This benchmark was produced by Synthgen Research and evaluates Synthgen's own product. We mitigate the resulting incentive by scoping every claim to a single per-class result on a frozen real-image test split fixed before any run, and by stating the benchmark's limitations in full. We welcome independent replication.

References

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9592–9600, 2019. doi:10.1109/CVPR.2019.00982.

[2] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021. doi:10.1007/s11263-020-01400-4.

[3] Yibin Huang, Congying Qiu, and Kui Yuan. Surface defect saliency of magnetic tile. *The Visual Computer*, 36(1):85–96, 2020. doi:10.1007/s00371-018-1588-5.

[4] B. Totino, F. Spagnolo, and S. Perri. RIAWELC: A Novel Dataset of Radiographic Images for Automatic Weld Defects Classification. *International Journal of Electrical and Computer Engineering Research (IJECER)*, 3(1):13–17, 2023. doi:10.53375/ijecer.2023.320.

[5] David P. Hughes and Marcel Salathé. An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing. *arXiv preprint* arXiv:1511.08060, 2015.

[6] Marble Surface Anomaly Detection (public dataset). Kaggle, contributor handle "wardaddy24". Classes: good, crack, joint, dot. Cited as a public dataset; no canonical peer-reviewed publication exists. Accessed 2026. `https://www.kaggle.com/datasets/wardaddy24/marble-surface-anomaly-detection`.

[7] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards Total Recall in Industrial Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14318–14328, 2022. arXiv:2106.08265.

[8] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM: A Discriminatively Trained Reconstruction Embedding for Surface Anomaly Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8330–8339, 2021. arXiv:2108.07610.

[9] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9664–9674, 2021. arXiv:2104.04015.

[10] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017. arXiv:1703.06907.

[11] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 737–744, 2020. arXiv:2009.13303.

[12] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for Data: Ground Truth from Computer Games. In *European Conference on Computer Vision (ECCV)*, pages 102–118, 2016. arXiv:1608.02192.

[13] Sergey I. Nikolenko. Synthetic Data for Deep Learning. *arXiv preprint* arXiv:1909.11512, 2019. Extended as a Springer monograph, 2021.

[14] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic Data from Diffusion Models Improves ImageNet Classification. *Transactions on Machine Learning Research (TMLR)*, 2023. arXiv:2304.08466.

[15] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is Synthetic Data from Generative Models Ready for Image Recognition? In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.07574.

[16] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification. *Neurocomputing*, 321:321–331, 2018. arXiv:1803.01229. doi:10.1016/j.neucom.2018.09.013.

[17] Luis Perez and Jason Wang. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv preprint* arXiv:1712.04621, 2017.

[18] Connor Shorten and Taghi M. Khoshgoftaar. A Survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6:Article 60, 2019. doi:10.1186/s40537-019-0197-0.

[19] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a Few Examples: A Survey on Few-Shot Learning. *ACM Computing Surveys*, 53(3):Article 63, 2020. arXiv:1904.05046. doi:10.1145/3386252.

[20] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep Long-Tailed Learning: A Survey. *arXiv preprint* arXiv:2110.04596, 2021. Published in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. doi:10.1109/TPAMI.2023.3268118.