Synthetic Data Generation in Manufacturing

A State-of-the-Art Review

Summary

Synthetic data generation has emerged as a transformative solution for manufacturing, with estimated up to 60% of data currently used to develop AI and analytics projects will be synthetically generated [1]. The global synthetic data generation market, valued between $168.9 million and $576 million in 2024, is projected to reach $2.67-3.5 billion by 2030-2031, reflecting growth rates between 34-40% CAGR [1-4].

This paper examines the current state-of-the-art in synthetic data generation for manufacturing applications, with particular emphasis on quality control, defect detection, and process optimization. We analyze key technological approaches including Generative Adversarial Networks (GANs), Diffusion Models, and physics-based simulation, alongside industrial case studies from BMW, Siemens, and the steel industry.

1. Introduction

Modern manufacturing faces a critical paradox: AI-driven quality control and optimization systems require vast amounts of diverse, high-quality data, yet real-world manufacturing data is often scarce, imbalanced, and expensive to collect [5,6]. Challenges include data scarcity, privacy concerns, and the difficulty of capturing rare defect scenarios in production environments.

Traditional data collection methods present multiple constraints:

- Scarcity of defect samples: Defective products are rare in well-controlled processes [7]
- Privacy and confidentiality: Proprietary manufacturing data cannot be easily shared [8]
- Cost and time: Physical testing and data labeling are resource-intensive [5]
- Safety concerns: Testing failure modes can be dangerous in real environments [5]
- Variability: Capturing all operational scenarios requires extensive sampling [6]

Synthetic data generation addresses these challenges by creating artificial datasets that replicate statistical properties of real-world data without exposing sensitive information [8,9]. This enables manufacturers to leverage broader and more diverse machine learning models by incorporating artificial data points for training and evaluation [5].

2. Core Technologies and Methodologies

Generative Adversarial Networks (GANs) remain the dominant architecture for synthetic data generation in manufacturing, accounting for 38.2% of market share in 2024, though newer approaches are gaining ground [2].

Architecture Variants:

- **DCGAN (Deep Convolutional GAN)**: Widely used for visual inspection systems in aerospace applications including composite fibers, air rudder surfaces, and surface coatings, with reported improvements in defect detection performance [10]
- **StyleGAN**: Produces images with high perceptual quality and structural coherence, particularly effective for generating diverse defect patterns [6,11]
- **CycleGAN & Pix2Pix**: Enable image-to-image translation for generating defects from masks or transforming simulated to realistic images [12]

Industrial Applications: GANs have been successfully employed for Industrial “Internet of Things” data generation, infrastructure maintenance, and statistical quality control procedures [5]. Research demonstrates that GAN-based data augmentation achieved AUC ROC scores equal to or higher than 0.9898 in automated visual inspection tasks [10].

Diffusion models are experiencing rapid growth at 47.6% CAGR, outpacing GANs due to their ability to produce cleaner, more diverse outputs [2]. These models treat generation as incremental de-noising, gradually reconstructing data from noise [13].

Key Advantages:

- Superior image quality compared to GANs [11,13]
- More stable training without discriminator networks [13]
- Better handling of complex, multi-modal distributions [13]
- Can generate higher-quality images than GANs without requiring adversarial training, though they are more computationally intensive [11]

Diffusion models, particularly Stable Diffusion based on Latent Diffusion Models, have been applied to industrial surface defect generation for the first time, enabling new approaches to quality control training [12].

Beyond purely data-driven approaches, physics-informed AI integrates domain knowledge with synthetic generation.

NVIDIA Omniverse & Isaac Sim: BMW Group leveraged NVIDIA DGX systems to train deep-learning-based synthetic data generation models, creating SORDI, the largest open-source dataset for industrial environments with over 800,000 photorealistic images spanning 80 categories [14,15].

- Real-time physics simulation [16,17]
- Domain randomization for robustness [18]
- Integration with CAD and digital twin platforms [17,19]
- Generation of thousands of photorealistic training images with a single click using NVIDIA Omniverse Replicator and Isaac Sim [14,15]

Emerging hybrid GAN-mechanistic architectures integrate physics-informed constraints from first-principles models with adversarial training to ensure both statistical fidelity and physical consistency [20]. This addresses limitations in pure data-driven methods by incorporating domain knowledge while maintaining flexibility [20].

3. Manufacturing Applications (with examples)

Quality Control and Defect Detection

Visual Inspection Systems:

The steel and metal industries represent a primary use case for synthetic data in defect detection [21-23]. Machine learning algorithms including GANs, CNNs, and Vision Transformers have been employed for metal defect detection, with deep learning models offering enhanced decision-making capabilities even with limited data availability [22,23].

The application of these techniques resulted in the following Improvements:

- Synthetic data augmentation using GANs improved automated visual inspection performance, with best results showing AUC ROC scores of 0.9898 or higher [10]
- GAN-generated datasets for metal additive manufacturing streamlined data preparation by eliminating human intervention while maintaining high performance in defect detection [24]
- Steel surface defect classification using semi-supervised GANs achieved approximately 16% improvement over baseline methods [21]

Application Areas:

- Hot-rolled steel surface inspection [21,22]
- Aluminum casting defect detection [25]
- Composite fiber quality control [10]
- Additive manufacturing layer monitoring [24]
- Weld defect identification [23,326]

Process Optimization and Digital Twins

BMW Virtual Factory: BMW Group's Virtual Factory uses industrial 3D metaverse applications based on NVIDIA Omniverse to perform real-time simulations, enabling virtual optimization of layouts, robotics, and logistics systems [17,19]. The initiative projects up to 30% reduction in production planning costs [19].

Key Outcomes:

- Hundreds of thousands of images generated with the push of a button, reducing time for employees to develop and deploy AI models for QA tasks by two-thirds [14,27]
- Digital collision checks for new vehicle models performed automatically [17,19]
- Real-time collaboration across global facilities [17]
- Integration of building data, equipment data, logistics data, and vehicle data [28]

Synthetic Data for Robot Training:

BMW employs Isaac Sim to generate synthetic data with domain randomization, using millions of photorealistic images with infinite variations in textures, orientations, and lighting conditions to train delivery robots under any condition [18]. There techniques can be used for human-robot collaborative systems optimization, collision probability prediction, motion planning in variable environments [18], fleet coordination across facilities [17]

Examples (case studies)

1.BMW Group: Comprehensive AI Implementation

Challenge: Manufacturing 2.5 million vehicles annually with 99% customer customization requires continuous innovation in production efficiency and quality control [14,27].

Solution: BMW built a synthetic data pipeline that generates photorealistic training images for their AI-powered Presence Detection system, which automatically evaluates images of critical components in real time [14,27].

Technology Stack:

- NVIDIA DGX systems with Hopper architecture [14]
- NVIDIA Omniverse Replicator [15,27]
- NVIDIA Isaac Sim [15,18]
- SORDI.ai central platform for image data management [14,15]

Results:

- Created largest open-source industrial dataset (800,000+ images) [14,15]
- Reduced AI model development time by 66% [14,27]
- Achieved 30% efficiency gains in planning workflows [19]
- Enabled quality inspection at production line speeds [27]

2. Siemens: Integration of Simulation and AI

Challenge: Aerodynamic simulations for automotive applications required extensive computational resources and physical testing [29].

Solution: Siemens integrated NVIDIA Omniverse APIs into Simcenter STAR-CCM+ to create physics-based digital twins, connecting simulation tools to NVIDIA's generative AI capabilities within CFD workflows [29].

Results:

- GPU-accelerated simulations matching performance of 10,000+ CPU cores [29]
- Reduced energy consumption and hardware costs [29]
- Enhanced visualization of massive engineering datasets [29]
- Seamless integration of AI assistants into engineering workflows [29]

3. Steel Industry: Defect Detection at Scale

Challenge: Steel surface defects are rare, making it difficult to collect sufficient training data for automated inspection systems [21,22].

Solution: Multiple steel manufacturers implemented GAN-based synthetic data generation to augment limited defect datasets [21,30].

Technology stack:

- Data acquisition combining real defect images with GAN-generated synthetic samples, followed by preprocessing including resizing, normalization, and augmentation [21,30]
- Integration with Google Vision API and Manufacturing Data Engine
- Deployment of semi-supervised learning with Convolutional Autoencoders and GANs [21]

Results:

- Significant improvement in classification accuracy for hot-rolled plates [21,22]
- Ability to detect micro-level cracks and inclusions invisible to human inspectors [22,23]
- Real-time detection maintaining production throughput [26]
- Enhanced adaptability to emerging defect patterns [31]

4. Technical Framework and Best Practices

A validated framework for synthetic data generation in manufacturing entails four main stages: data collection, pre-processing, synthetic data generation, and evaluation [5].

Stage 1: Data Collection

- Gather representative samples from production [5]
- Document rare failure modes and edge cases [6]
- Establish baseline quality metrics [5]
- Ensure diverse operational conditions [5]

Stage 2: Pre-processing

- Standardize image resolution and formats [21,30]
- Normalize sensor data ranges [5]
- Apply domain-specific transformations [30]
- Create structured annotations [5,21]

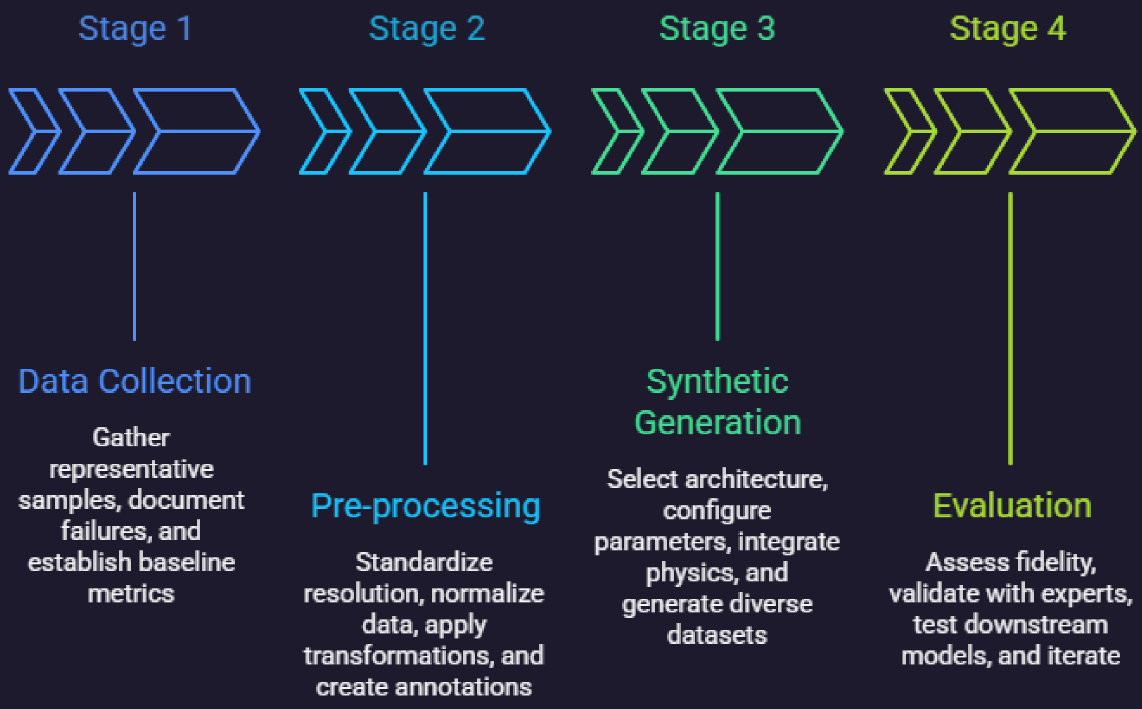
Stage 3: Synthetic Generation

- Select appropriate generative architecture [5,6]
- Configure domain randomization parameters [15,18]
- Integrate physics constraints where applicable [17,20]
- Generate diverse, balanced datasets [5,10]

Stage 4: Evaluation

- Assess statistical fidelity (SSIM, FID, LPIPS) [11,13]
- Validate with domain experts [5]

- Test downstream model performance [10,24]
- Iterate based on results [5]



Stage 1

Data Collection

Gather representative samples, document failures, and establish baseline metrics

Stage 2

Pre-processing

Standardize resolution, normalize data, apply transformations, and create annotations

Stage 3

Synthetic Generation

Select architecture, configure parameters, integrate physics, and generate diverse datasets

Stage 4

Evaluation

Assess fidelity, validate with experts, test downstream models, and iterate

Some Quality Assessment Metrics

Image Quality Metrics:

- SSIM (Structural Similarity Index)**:** Measures perceptual similarity [11,13]
- FID (Fréchet Inception Distance)**:** Evaluates distribution matching [11,13]
- LPIPS (Learned Perceptual Image Patch Similarity)**:** Assesses perceptual distance [11]
- CLIPScore: Evaluates semantic alignment [32]

How the performance is checked:

- Classification accuracy on real test data: train on synthetic, test on real, how accurate is the model? [10,21,22]
- Detection precision and recall: for object detection (e.g., finding defects), how many true positives vs. false alarms? [23,26]
- Generalization to unseen scenarios: does the model work on unseen scenarios not in training data? [5,24]
- Robustness to distribution shifts: does it handle distribution shifts (e.g., new cameras, lighting, machines)? [6]

Common Pitfalls and Solutions

Challenge 1: Mode Collapse in GANs

Solution: Use Wasserstein GAN with gradient penalty (WGAN-GP) [20]; Monitor diversity metrics throughout training [6]; Implement progressive training strategies [11]

Challenge 2: Insufficient Physical Realism

Solution: Integrate physics-based constraints [17,20]; use hybrid simulation-learning approaches [20,29]; validate with domain experts regularly [5]

Challenge 3: Data Bias Amplification

Solution: Ensure diverse seed data [5,8]; implement fairness constraints [8]; regular bias audits of generated data [8]

Challenge 4: Computational Requirements

Solution: Cloud deployment leveraging elastic GPU pools and integrated compliance tooling, accounting for 67.5% of 2024 deployments [42; optimize model architectures for efficiency [13]; use progressive generation strategies [11]

---

5. Market Landscape and Leading Solutions

Leading Commercial Platforms

1. **NVIDIA Omniverse**: Evolved into the de facto operating system for physical AI with over 300,000 downloads and 252+ enterprise deployments, delivering 30-70% efficiency gains across workflows [16,33]
2. **MOSTLY AI**: Specializes in privacy-preserving synthetic data with fairness controls, securing contracts including a $196,800 agreement with U.S. Department of Homeland Security [34]
3. **Gretel**: API-driven platform supporting tabular, text, and image data with strong developer integration
4. **Synthesis AI**: Focuses on photorealistic computer vision datasets for complex scenarios

Open-Source Solutions

Synthetic Data Vault (SDV): Python ecosystem for tabular, relational, and time-series data; Synthea: Healthcare-specific patient data generation; Community Tools: Growing ecosystem of specialized generators for manufacturing domains

---

6. Future Directions and Emerging Trends

Foundation Models and World Simulators

Integration with Cosmos World Foundation Models enables physics-based world generation and synthetic data creation at unprecedented scale, with neural networks that understand real-world physics properties [16]. This can provide more realistic multi-modal simulation, better handling of complex physical interactions [16], reduced need for extensive real-world data collection [8,32], enhanced digital twin capabilities [17,28]

Agentic AI and Autonomous Systems

Manufacturing is moving toward autonomous decision-making systems that require continuous learning from diverse scenarios [8,33]. Synthetic data generation will be critical for: safe training of autonomous factory systems [18,33], scenario planning for edge cases [5,18], continuous model updating without production disruption [8], integration with predictive maintenance systems [26]

Regulatory and Compliance Considerations

Regulations such as the EU AI Act require firms to test synthetic alternatives before processing personal data, making generation platforms a compliance necessity [8,9].This can be important for auditable data lineage, differential privacy guarantees [8], validation against regulatory standards [8,9], documentation of generation processes [5,8].

Quantum-Enhanced Generatio

Emerging research on quantum Wasserstein GANs with gradient penalty (QWGAN-GP) demonstrates potential advantages in capturing complex dynamics of industrial bioprocesses, offering a promising direction for addressing data scarcity [20].

---

7. Recommendations for Implementation

For Manufacturing Leaders

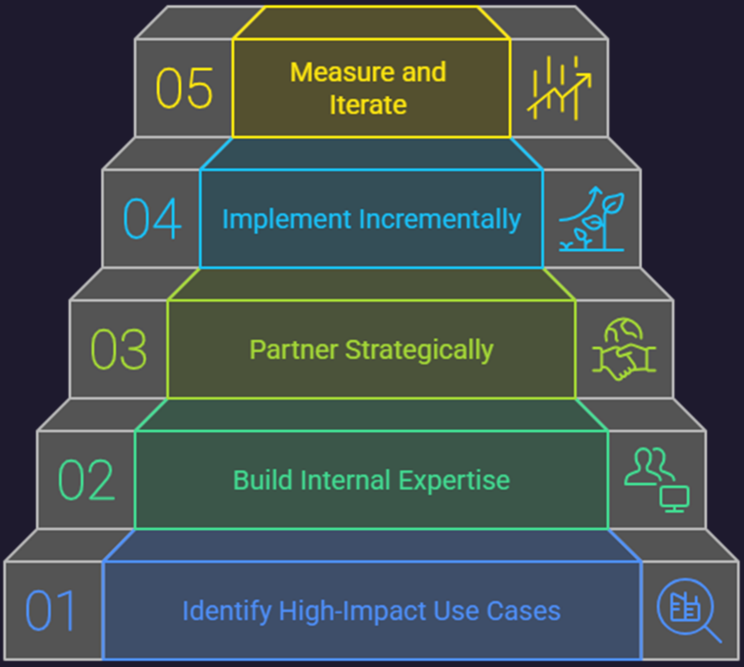
1. **Start with High-Impact Use Cases**: Focus on quality control and defect detection where ROI is clearest [10,14,27]
2. **Build Internal Expertise**: Invest in training teams on both AI and domain knowledge [5,33]
3. **Partner Strategically**: Leverage platforms like NVIDIA Omniverse or cloud providers [14,16,17]
4. **Implement Incrementally**: Begin with pilot projects before full-scale deployment [5]
5. **Measure and Iterate**: Establish clear metrics for success and continuously improve [5]

For Technical Teams

1. **Choose Architectures Wisely**: Match generation method to data type and use case [5,6]
2. **Prioritize Quality Over Quantity**: Better to have fewer high-quality synthetic samples [10,13]
3. **Validate Rigorously**: Never deploy models trained solely on synthetic data without real-world validation [5,10]
4. **Document Everything**: Maintain comprehensive records of generation processes and parameters [5,8]
5. **Stay Current**: Field is evolving rapidly; continuous learning is essential [6,32]

For Researchers

1. **Develop Standardized Benchmarks**: Industry needs consistent evaluation frameworks [6,32]
2. **Address Domain Gaps**: More work needed on physics-constrained generation [20]
3. **Focus on Interpretability**: Make generation processes more transparent and controllable [6,32]
4. **Collaborate with Industry**: Ensure research addresses real manufacturing challenges [5,25]
5. **Open-Source Contributions**: Share tools and datasets to accelerate field development [14]



8. Conclusions

Synthetic data generation has transitioned from experimental technique to production-ready technology in manufacturing. The global market growth from $168.9 million to projected $3.5 billion by 2031 reflects the transformative potential and increasing adoption across industries [1-4].

GANs and diffusion models have demonstrated proven effectiveness in manufacturing quality control, with documented improvements in defect detection accuracy and operational efficiency. Case studies from BMW, Siemens, and steel manufacturers show measurable ROI, including 30-70% efficiency gains and 66% reduction in AI development time. Beyond defect detection, synthetic data enables diverse applications such as digital twins, robotic training, process optimization, and predictive maintenance.

The landscape is rapidly evolving, with diffusion models growing at 47.6% CAGR, cloud deployment holding 67.5% market share, and integration with foundation models signaling continued advancement. For manufacturers to remain competitive, synthetic data generation is becoming essential infrastructure rather than an optional enhancement.

The convergence of generative AI, physics-based simulation, and industrial domain expertise is creating unprecedented opportunities for manufacturing optimization. Organizations that strategically invest in synthetic data capabilities now will be positioned to lead in the AI-driven future of manufacturing.

References

[1] Fortune Business Insights. (2024). Synthetic Data Generation Market Forecast Analysis [2030]. https://www.fortunebusinessinsights.com/synthetic-data-generation-market-108433

[2] Mordor Intelligence. (2025). Synthetic Data Market Size, Share, Trends & Research Report, 2030. https://www.mordorintelligence.com/industry-reports/synthetic-data-market

[3] Precedence Research. (2024). Synthetic Data Generation Market Size, Report By 2034. https://www.precedenceresearch.com/synthetic-data-generation-market

[4] Research Nester. (2025). Synthetic Data Generation Market Size, Share & Growth Forecast 2035. https://www.researchnester.com/reports/synthetic-data-generation-market/5711

[5] Buggineni, V., Chen, C., & Camelio, J. (2024). Enhancing manufacturing operations with synthetic data: a systematic framework for data generation, accuracy, and utility. *Frontiers in Manufacturing Technology*, 4. https://doi.org/10.3389/fmtec.2024.1320166

[6] Journal of Intelligent Manufacturing. (2025). Generative AI in industrial machine vision: a review. Springer. https://doi.org/10.1007/s10845-025-02604-6

[7] ClearBox AI. (2024). Synthetic data for the manufacturing industry: the new NGA4M project. https://www.clearbox.ai/blog/2024-06-10-synthetic-data-for-the-manufacturing-industry-the-new-NGA4M-project

[8] TechnoStacks. (2025). Generative AI in Business: How Synthetic Data Transforms Industries. https://technostacks.com/blog/generative-ai-in-business-transforming-industries-with-synthetic-data/

[9] Globe Newswire. (2024). Synthetic Data Generation Market Research 2024 - Global Industry Size, Share, Trends, Opportunity, and Forecast, 2019-2029. https://www.globenewswire.com/news-release/2024/10/10/2961354/28124/en/Synthetic-Data-Generation-Market-Research-2024-Global-Industry-Size-Share-Trends-Opportunity-and-Forecast-2019-2029-Rising-Demand-for-Diverse-Sources-Advancements-in-GANs.html ; https://www.grandviewresearch.com/industry-analysis/synthetic-data-generation-market-report

[10] ScienceDirect. (2023). Synthetic Data Augmentation Using GAN For Improved Automated Visual Inspection. https://www.globenewswire.com/news-release/2024/10/10/2961354/28124/en/

[11] Z.Sordo, et.al, Synthetic Scientific Image Generation with VAE, GAN, and Diffusion Model Architectures. J. Imaging 2025, 11(8), 252; https://doi.org/10.3390/jimaging11080252

[12] Xiaopin Zhong et.al, An Overview of Image Generation of Industrial Surface Defects.Sensors (Basel). 19:8160 (2023); doi: 10.3390/s23198160

[13] Jun Zhu, Synthetic data generation by diffusion models. *National Science Review*, 11(8), (2024); https://doi.org/10.1093/nsr/nwae276

[14] NVIDIA. (2024). Case Study: NVIDIA Boosts BMW Group's Production Efficiency with AI. https://www.nvidia.com/en-us/case-studies/bmw-optimizes-production-with-ai-and-dgx-systems/

[15] NVIDIA. (2024). BMW Group Develop Custom Application on NVIDIA Omniverse. https://www.nvidia.com/en-us/case-studies/bmw-group-develop/

[16] Magniative.(2025). How NVIDIA Is Building the Operating System for Physical AI https://www.maginative.com/article/how-nvidia-is-building-the-operating-system-for-physical-ai/

[17] BMW Press Release. (2025). BMW Group scales virtual factory. https://www.press.bmwgroup.com/global/article/detail/T0450699EN/bmw-group-scales-virtual-factory

[18] NVIDIA Customer Stories. (2024). BMW employs NVIDIA Isaac Sim to train delivery robots. https://www.nvidia.com/en-us/case-studies/paving-the-future-of-factories-with-nvidia-omniverse-enterprise/

[19] BMW Press Release. (2021). BMW Group and NVIDIA take virtual factory planning to the next level. https://www.press.bmwgroup.com/global/article/detail/T0329569EN/bmw-group-and-nvidia-take-virtual-factory-planning-to-the-next-level

[20] arXiv. (2025). Quantum-enhanced generative adversarial networks for industrial bioprocess data. arXiv:2510.17688v1, https://arxiv.org/abs/2508.09209v2

[21] A. Boikov, et al. Synthetic Data Generation for Steel Defect Detection and Classification Using Deep Learning. *Symmetry*, 13(7), 1176 (2021) https://www.mdpi.com/2073-8994/13/7/1176

[22] V. Vasan et,al, Detection and classification of surface defects on hot-rolled steel using vision transformers. *Heliyon*, 10(19), e38498 (2024)

[23] Y.Feng et. Al, Research on metal surface defect detection method based on deep learning, Scientific Reports 16, 1436 (2026). https://www.nature.com/articles/s41598-025-31235-3

[24] arXiv. (2024). Scalable AI Framework for Defect Detection in Metal Additive Manufacturing. arXiv:2411.00960. https://arxiv.org/abs/2411.00960

[25] arXiv. (2024). A Comprehensive Survey on Machine Learning Driven Material Defect Detection. arXiv:2406.07880v1. https://arxiv.org/abs/2406.07880

[26] AkriData. (2025). AI Surface Defect Detection in Steel & Metal Industries. https://akridata.ai/blog/surface-defect-detection-ai-steel-metal-industries/

[27] NVIDIA. (2024). Discover how BMW Group is Transforming Manufacturing. https://www.nvidia.com/en-us/lp/omniverse/how-bmw-group-transforming-manufacturing/

[28] NVIDIA. (2024). Industrial Facility Digital Twins—Use Case. https://www.nvidia.com/en-us/use-cases/industrial-facility-digital-twins/

[29] NVIDIA. (2026). Siemens and NVIDIA Expand Partnership to Build the Industrial AI Operating System https://nvidianews.nvidia.com/news/siemens-and-nvidia-expand-partnership-industrial-ai-operating-system

[30] E.Guclu, i.Aydin, E. Akin, Enhanced defect detection on steel surfaces using integrated residual refinement module with synthetic data augmentation, Measurement 250, 117136 (2025) https://www.sciencedirect.com/science/article/abs/pii/S0263224125004956

[31] S.Hu et al, Application of self-supervised learning in steel surface defect detection. *Journal of Materials Informatics*, 4, 21 (2025) https://www.oaepublish.com/articles/jmi.2025.21

[32] arXiv. (2025). Generative Models for Synthetic Data: Transforming Data Mining in the GenAI Era. arXiv:2508.19570. https://arxiv.org/abs/2508.19570

[33] A3 Automation. (2025). What is NVIDIA Omniverse and How Will it Affect U.S. Manufacturing. https://www.automate.org/blogs/what-is-nvidia-omniverse-and-how-will-it-affect-u-s-manufacturing

[34] Emergen Research. (2024). Top 10 Companies in Synthetic Data Generation Market in 2025 Shaping Industry Trends.