Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2025, 12(3):14-20**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

AI Driven Data Synthesis and Augmentation

Dr. Amardeep

Department of Computer Science, Gopichand Arya Mahila College, Abohar, Punjab, India

Abstract: Generative Artificial Intelligence (AI) has emerged as a transformative force in machine learning, particularly in the realm of data synthesis and augmentation. Data augmentation plays a critical role in expanding datasets, thus enhancing model performance across various complex tasks, such as image recognition, natural language processing, and speech recognition. The demand for diverse and extensive datasets continues to rise, and generative AI provides an innovative solution by producing high-quality synthetic data to complement real-world datasets. Techniques such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models are central to this process, enabling the creation of synthetic data that mirrors the original dataset and introduces variety. These generative models have significantly improved model accuracy, generalization, and robustness, especially in areas with limited labelled data. However, the integration of generative AI also introduces challenges, including potential biases in the generated data and ethical considerations related to its use. Despite these challenges, the potential applications of generative AI in data augmentation are vast, offering new solutions to data scarcity, bias, and generalization problems. This paper reviews the lifecycle of data generation techniques for AI Models, from data preparation to application, highlighting the constraints and potential pathways for future development. By providing a comprehensive understanding of these methodologies, this research aims to guide researchers in selecting appropriate data generation strategies for constructing accurate and explore future advancements in the field.

Keywords: Data Augmentation, Data Synthesis, Synthetic data, Data Scarcity Solutions

1. Introduction

In modern machine learning, the technology has become crucial for solving complex machine vision challenges. The intricate design of state-of-the-art machine learning models lies in their vast array of parameters, which need to be fine-tuned to capture a wide range of visual phenomena. This complexity is further heightened by the diverse appearance variations of real-world objects and scenes, making it essential to include different data variations during training [1]. Consequently, the process of training machine learning models requires an extensive amount of annotated data to ensure they generalize well and avoid overfitting. However, collecting and annotating such large datasets is often extremely time-consuming and expensive [2]. To address these challenges, data augmentation has emerged as an effective solution. It involves the artificial creation of new data samples to expand the training set. This process typically includes transformations that change the visual characteristics of the original data while maintaining their labels, thereby simulating real-world conditions such as different view angles, pose variations, and other visual distortions.

Data augmentation is vital in various machine learning scenarios, particularly when the available training data is insufficient, of poor quality, or not representative of the target data. Common problems include limited training data, poor perceptual quality, lack of adequate appearance variations, skewed class proportions, and data available under a single condition [3]. While the first four issues can be addressed by manipulating existing data to produce additional training samples, the last two challenges often require the creation of entirely new training data. In such cases, synthetic data augmentation becomes essential. By generating new data from scratch,

synthetic data augmentation can meet the specific needs of various machine vision tasks, providing task-specific data formats and annotation schemes that traditional transformation-based methods might not fulfill [4].

The importance of synthetic data augmentation is especially clear in application areas where obtaining real-world data is impractical or excessively costly. For example, in autonomous driving and navigation, pose estimation, affordance learning, object grasping, and manipulation, data synthesis methods are invaluable [5]. These methods can simulate a broad range of task-specific, real-world variabilities in the synthesized data, supporting non-standard image modalities such as point clouds and voxels. Approaches based on 3D modeling offer scalable resolutions and flexible content and labeling schemes, tailored to specific use cases. This versatility highlights the importance of synthetic data augmentation in providing high-quality training data for machine learning applications, especially in emerging fields where traditional data collection methods fall short [6].

Despite the growing significance of synthetic data augmentation, the existing literature lacks comprehensive surveys on this topic. While many studies focus on traditional data augmentation techniques, few address the unique challenges and methods associated with synthetic data generation [7]. This survey aims to fill this gap by offering an in-depth analysis of synthetic data augmentation methods, discussing their principles, use cases, and limitations. By enriching the current literature, this work seeks to emphasize the critical role of synthetic data augmentation in advancing machine learning applications, particularly in scenarios characterized by severe data scarcity [8].

2. Overview of Synthetic Data Augmentation

Synthetic data augmentation techniques provide useful methods for enhancing and expanding training datasets in machine learning. These techniques help to improve model performance and generalization across various applications. Geometric data augmentation methods such as affine transformations [9], projective transformations [10], and nonlinear deformations [11] are aimed at creating various transformations of the original images to handle spatial variations resulting from changes in object size, orientation, or view angles. Common geometric transformations include rotation, shearing, scaling or resizing, nonlinear deformation, cropping, and flipping. On the other hand, photometric techniques, such as color jittering [12], lighting perturbation [13-14], and image denoising [15], manipulate the qualitative properties of images, including contrast, brightness, color, hue, saturation, and noise levels. These techniques make the resulting machine learning models invariant to changes in these properties. To ensure good generalization performance in different scenarios, it is often necessary to apply many of these procedures simultaneously.

Recently, more advanced data augmentation methods have gained popularity. One of the important classes of techniques [16-19] involves transforming different image regions discretely instead of uniformly manipulating the entire input space. These methods have proven effective in simulating complex visual effects such as non-uniform noise, non-uniform illumination, partial occlusion, and out-of-plane rotations. Another major approach to data augmentation exploits feature space transformation to introduce variability in training data. These regularization techniques manipulate learned feature representations within deep neural networks to alter the visual appearance of underlying images. Examples include feature mixing [16], feature interpolation [20], feature dropping [21], and selective augmentation of useful features. Although these methods might not always lead to semantically meaningful alterations, they have proven valuable in enhancing the performance of machine learning models.

Across different modalities, data augmentation techniques often exhibit similarities. For instance, in image data, augmentation operations encompass mosaic [22], flipping, copy-pasting, adding noise, and pairing. Similarly, in text data, augmentation operations involve synonym replacement, copy-pasting, and other techniques. To cater to the demands of multimodal learning, existing research has addressed cross-modal information alignment during data augmentation. MixGen [23] generates new training samples by linearly interpolating images and concatenating text sequences from two existing image-text pairs. The semantic relationship within the newly generated image-text pair remains consistent and matched. In the rapidly advancing landscape of large language models (LLMs), data augmentation has emerged as a cornerstone for enhancing model performance through the diversification of training exemplars, circumventing the need for extensive additional data gathering.

Data labeling leverages the comprehensive language understanding capabilities of LLMs to annotate vast unlabeled datasets [24]. This methodology is particularly beneficial in fields with a substantial corpus of unlabeled data, such as cross-lingual processing and multimodal learning [25], where automation can significantly expedite the data preparation process. Recent research explores the zero-shot annotation ability of LLMs, such as GPT-4, for labeling political Twitter [22]. Additionally, Khan et al. focus on visual question answering (VQA) tasks by generating pseudo-label data from unlabeled images using the SelTDA framework. These advancements highlight the significant potential of data augmentation and labeling in improving machine learning models' robustness and adaptability [23].

Data synthesis aims to create entirely new data from scratch or based on generative models that mimic the distribution of real data. With advancements in generative AI, there have been significant improvements in the quality and efficiency of generating synthetic data. The paper categorizes data synthesis methods into three main types: general model distillation, domain model distillation, and model self-improvement. General model distillation involves leveraging powerful general models, such as StableVicuna, ChatGPT, and GPT-4, to generate datasets that enhance the capabilities of weaker models. Techniques include using predefined templates to generate tiny stories [26] and employing large language models (LLMs) to evaluate the quality of generated data. Studies have shown that high-quality data can train a powerful model, exemplified by the comprehensive generation of textbooks and exercises from GPT-3.5 [27]. Other methods have achieved performance improvements by generating instruction datasets and fine-tuning models [28-30].

Domain model distillation focuses on using models tailored to generate data within a specific domain. This approach is necessary when general models do not meet the specific needs of industry applications. For example, in code programming, domain model distillation generates instructional data tailored to specific coding tasks [31-32]. In mathematics, methods like Minerva [33] and DeepSeekMath [34] generate solutions to mathematical problems, ensuring accuracy and diversity. Additionally, industry data often presents barriers such as limited data scales and inaccessibility within specific enterprises. These challenges necessitate domain-specific models to address the unique requirements effectively. Model self-improvement refers to the process where a model generates higher-quality data to enhance its capabilities. For instance, leveraging existing instructions to adjust the model and prompting it to paraphrase documents in specific styles can improve performance with minimal human intervention [35-36].

3. Assessing The Efficiency of Synthetic Data Augmentation

Currently, there are numerous large-scale synthetic datasets available for the training and evaluation of machine vision models. Table 1 provides a summary of some of the most significant synthetic datasets.

Table 1: Publicly available large scale synthetic datasets

Dataset Name	Type	Size/Volume	Description	Source
fineweb-edu-score-2	Text	5.4 trillion tokens	A dataset focused on educational content for training large language models (LLMs).	ProjectPro
Cosmopedia	Text	25 billion tokens	The largest open-source synthetic dataset, covering diverse topics in various text formats.	Hugging Face
OpenMathInstruct-1	Text-Code	1.8 million problem-solution pairs	Combines natural language instructions with Python code for math problem-solving.	Hugging Face
The Pile	Text	800 GB	A corpus from 22 datasets aimed at enhancing model generalization across diverse contexts.	Kili Technology

C4 (Colossal Clean Crawled Corpus)	Text	750 GB	Derived from Common Crawl, focusing on natural language data with heavy deduplication.	Kili Technology
Starcoder Data	Code	783 GB	Programming-centric dataset containing code from GitHub and Jupyter Notebooks.	Kili Technology
ROOTS	Multilingual Text	1.6 TB	Curated from multiple sources to train multilingual LLMs, including deduplicated data.	Kili Technology
HAPNEST	Genotype/Phenotype Data	1 million individuals	Generates synthetic genetic data for polygenic risk scoring with diverse traits.	Nature
GPR+	Image	808 identities, 475,104 bounding boxes	Upgraded dataset for person re-identification with detailed attribute annotations.	GPR+ Project
SyntheWorld	Image	40,000 images	A synthetic dataset designed for land cover mapping with high-resolution images.	IEEE Explore

Numerous studies have shown the effectiveness of synthetic data augmentation techniques in various machine vision applications. In some instances, synthetic data has even outperformed real data in enhancing model generalization. For example, Wang et al. [37] reported that models trained on synthetic data achieved better results in face recognition tasks compared to those trained on real data. Similarly, Rogez and Schmid [38] consistently found that synthetic data yielded higher performance than real data in pose estimation tasks. These findings suggest that synthetic images, which are often cleaner and free from irrelevant artifacts, can be particularly advantageous in settings that do not require high levels of photorealism, such as depth perception [39] and pose estimation [38, 40]. However, despite these promising results, it is important to note that synthetic data alone does not always guarantee optimal performance. Richter et al. [41] demonstrated that while synthetic data can significantly reduce the amount of real training data needed, it does not always achieve satisfactory results on its own.

Combining synthetic and real data has proven to be a more effective approach in many cases. Rajpura and Bojinov [42] compared the performance of deep learning-based object detectors trained on synthetic, real, and hybrid datasets. They found that while models trained solely on synthetic data performed worse (24 mAP) than those trained on real data (28 mAP), the inclusion of both synthetic and real images improved performance by up to 12% (36 mAP). Similarly, Alhajia et al. [43] observed that training in an augmented reality environment that integrates both real and synthetic objects resulted in significantly higher performance compared to using either type of data alone. Additionally, Zhang et al. [44] found that increasing the proportion of synthetic data in the training set does not always lead to a linear increase in model performance. In some tasks, performance gains plateaued at around 25% synthetic data composition. These findings highlight the importance of a balanced approach, leveraging both synthetic and real data to achieve the best results.

4. Conclusion

Exploring AI-driven data synthesis and augmentation reveals a significant impact on enhancing machine learning models. These synthetic data augmentation methods have proven invaluable, especially in situations where real data is scarce or insufficient. High-fidelity synthetic data has shown particular promise in addressing challenges like missing data and biases, especially in healthcare. These techniques help alleviate data scarcity

and enable more accurate predictions, aiding advancements in disease prediction, drug discovery, and personalized medicine.

Generative AI technologies, such as GANs and VAEs, have greatly improved the quality of synthetic data, making it closely resemble real-world data. Additionally, augmentation techniques have expanded datasets, providing diverse samples for training machine learning models. As this field progresses, evaluation metrics for synthetic datasets have also advanced, including measures of utility, privacy, and domain-specific characteristics. While synthetic data offers great potential, it also brings challenges related to biases, representativeness, and privacy concerns. Addressing these issues requires careful evaluation, transparent documentation, and rigorous assessment to ensure the effective and ethical use of synthetic data in AI-driven applications. The ongoing developments in this field hold the potential to revolutionize data-driven research and applications across various domains.

References

[1]. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). Carla: An open urban driving simulator. In Conference on robot learning (pp. 1-16). PMLR.

[2]. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., & Farhadi, A. (2017). Ai2-thor: An interactive 3d environment for visual ai. arXiv preprint arXiv:1712.05474.

[3]. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I., & Schmid, C. (2017). Learning from synthetic humans. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 109-117).

[4]. Rogez, G., & Schmid, C. (2016). Mocap-guided data augmentation for 3d pose estimation in the wild. Advances in neural information processing systems, 29.

[5]. Mo, K., Qin, Y., Xiang, F., Su, H., & Guibas, L. (2022). O2o-afford: Annotation-free large-scale object-object affordance learning. In Conference on Robot Learning (pp. 1666-1677). PMLR.

[6]. Chu, F.-J., Xu, R., & Vela, P. A. (2019). Learning affordance segmentation for real-world robotic manipulation via synthetic images. IEEE Robotics and Automation Letters, 4(2), 1140-1147.

[7]. Lin, Y., Tang, C., Chu, F.-J., & Vela, P. A. (2020). Using synthetic data and deep networks to recognize primitive shapes for object grasping. In 2020 IEEE International Conference on Robotics and Automation (ICRA) (pp. 10494-10501). IEEE.

[8]. Ummadisingu, A., Takahashi, K., & Fukaya, N. (2022). Cluttered food grasping with adaptive fingers and synthetic-data trained object detection. arXiv preprint arXiv:2203.05187.

[9]. A. H. Ornek and M. Ceylan, "Comparison of traditional transformations for data augmentation in deep learning of medical thermography," in 2019 42nd International Conference on Telecommunications and Signal Processing (TSP). IEEE, 2019, pp. 191–194.

[10]. K. Wang, B. Fang, J. Qian, S. Yang, X. Zhou, and J. Zhou, "Perspective transformation data augmentation for object detection," IEEE Access, vol. 8, pp. 4935–4943, 2019.

[11]. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in 2016 fourth international conference on 3D vision (3DV). IEEE, 2016, pp. 565–571.

[12]. E. K. Kim, H. Lee, J. Y. Kim, and S. Kim, "Data augmentation method by applying color perturbation of inverse PSNR and geometric transformations for object recognition based on deep learning," Applied Sciences, vol. 10, no. 11, p. 3755, 2020.

[13]. D. Sakkos, H. P. Shum, and E. S. Ho, "Illumination-based data augmentation for robust background subtraction," in 2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA). IEEE, 2019, pp. 1–8.

[14]. O. Mazhar and J. Kober, "Random shadows and highlights: A new data augmentation method for extreme lighting conditions," arXiv preprint arXiv:2101.05361, 2021.

[15]. A. Kotwal, R. Bhalodia, and S. P. Awate, "Joint desmoking and denoising of laparoscopy images," in 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). IEEE, 2016, pp. 1050–1054.

[16]. H. Li, X. Zhang, Q. Tian, and H. Xiong, “Attribute mix: semantic data augmentation for fine grained recognition,” in 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP). IEEE, 2020, pp. 243–246.
[17]. S. Feng, S. Yang, Z. Niu, J. Xie, M. Wei, and P. Li, “Grid cut and mix: flexible and efficient data augmentation,” in Twelfth International Conference on Graphics and Image Processing (ICGIP 2020), vol. 11720. International Society for Optics and Photonics, 2021, p. 1172028.
[18]. S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, “Cutmix: Regularization strategy to train strong classifiers with localizable features,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6023–6032.
[19]. J. Yoo, N. Ahn, and K.-A. Sohn, “Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8375–8384.
[20]. J. Lemley, S. Bazrafkan, and P. Corcoran, “Smart augmentation learning an optimal data augmentation strategy,” IEEE Access, vol. 5, pp. 5858–5869, 2017.
[21]. X. Li, Y. Dai, Y. Ge, J. Liu, Y. Shan, and L.-Y. Duan, “Uncertainty modeling for out-of-distribution generalization,” arXiv preprint arXiv:2202.03958, 2022.
[22]. Hao, Xiaoshuai, et al. "Mixgen: A New Multi-Modal Data Augmentation." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 379-389.
[23]. Khan, Zaid, et al. "Q: How to Specialize Large Vision-Language Models to Data-Scarce VQA Tasks? A: Self-Train on Unlabeled Images!" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15005-15015.
[24]. Zhu, Yiming, et al. "Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks." arXiv preprint arXiv:2304.10145 (2023).
[25]. Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." Proceedings of the National Academy of Sciences, vol. 120, no. 30, 2023, e2305016120.
[26]. Ronen Eldan and Yuanzhi Li. "TinyStories: How Small Can Language Models Be and Still Speak Coherent English?" arXiv preprint arXiv:2305.07759 (2023).
[27]. Suriya Gunasekar et al. "Textbooks Are All You Need." arXiv preprint arXiv:2306.11644 (2023).
[28]. Lichang Chen et al. "Alpagasus: Training a Better Alpaca with Fewer Data." arXiv preprint arXiv:2307.08701 (2023).
[29]. Or Honovich et al. "Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor." arXiv preprint arXiv:2212.09689 (2022).
[30]. Taori, Rohan, et al. "Stanford Alpaca: An Instruction-Following Llama Model." 2023, https://github.com/tatsu-lab/stanford_alpaca.
[31]. Ziyang Luo et al. "WizardCoder: Empowering Code Large Language Models with Evol-Instruct." International Conference on Learning Representations (ICLR), 2024.
[32]. Yuxiang Wei et al. "Magicoder: Empowering Code Generation with OSS-Instruct." Forty-First International Conference on Machine Learning, 2024.
[33]. Aitor Lewkowycz et al. "Solving Quantitative Reasoning Problems with Language Models." Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 3843–3857.
[34]. Huajian Xin et al. "DeepSeek-Prover: Advancing Theorem Proving in LLMs through Large-Scale Synthetic Data." arXiv abs/2405.14333 (2024).
[35]. Pratyush Maini et al. "Rephrasing the Web: A Recipe for Compute and Data-Efficient Language Modeling." arXiv preprint arXiv:2401.16380 (2024).
[36]. Yizhong Wang et al. "Self-Instruct: Aligning Language Models with Self-Generated Instructions." The 61st Annual Meeting of the Association for Computational Linguistics, 2023.
[37]. X. Wang, K. Wang, and S. Lian, “A survey on face data augmentation for the training of deep neural networks,” Neural computing and applications, vol. 32, no. 19, pp. 15 503–15 531, 2020.
[38]. G. Rogez and C. Schmid, “Mocap-guided data augmentation for 3d pose estimation in the wild,” Advances in neural information processing systems, vol. 29, 2016.

[39]. G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 109–117.
[40]. A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2107–2116.
[41]. S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in European conference on computer vision. Springer, 2016, pp. 102–118.
[42]. P. S. Rajpura, H. Bojinov, and R. S. Hegde, "Object detection using deep cnns trained on synthetic images," arXiv preprint arXiv:1706.06782, 2017.
[43]. H. A. Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets deep learning for car instance segmentation in urban scenes," in British machine vision conference, vol. 1, 2017, p. 2.
[44]. Z. Zhang, L. Yang, and Y. Zheng, "Multimodal medical volumes translation and segmentation with generative adversarial network," Handbook of Medical Image Computing and Computer Assisted Intervention, pp. 183–204, 2020.