From Scarcity to Abundance: Expansion Manufacturing Data through Limited Defect Images

Junhyung Moon*
KT
junhyung.moon@kt.com

Minyeol Yang*
Sungkyunkwan University
yangget@g.skku.edu

Songmi Park
Sungkyunkwan University
thdal1232@gmail.com

Jongpil Jeong†
Sungkyunkwan University
jpjeong@skku.edu

Abstract

The increasing adoption of IoT sensors, communication capabilities, and software applications in manufacturing environments has led to a growing demand for handling diverse large-scale manufacturing data. This trend indicates that AI is being researched and developed as an essential tool for improving cost-effectiveness and efficiency. Recently, there has been a significant increase in demand for process improvement using deep learning technology in smart manufacturing processes. However, obtaining a sufficient amount of training data in real industrial environments is challenging due to security and cost concerns for companies. Therefore, we propose utilizing generative artificial intelligence to efficiently expand manufacturing datasets. For data augmentation, we use a model that combines Stable Diffusion and LoRA fine-tuning, and apply the text generation approach of BLIP. We anticipate that these data augmentation will help to improve the performance of artificial intelligence in the manufacturing field while reducing the cost of data collection.

Keywords: Smart Manufacturing, Data Scarcity, Data Augmentation, Generative AI

1. Introduction

The growing adoption of smart manufacturing technologies has shown that there is significant potential to further improve the state recognition of manufacturing machines and processes, reduce downtime, and increase the automation level and product quality (Tercan and Meisen, 2022; Wang et al., 2018). Monitoring, evaluation, and analysis based on Artificial Intelligence (AI) algorithms are highly valuable for improving system productivity, product quality, and sustainability. The adoption of such technologies can reduce errors in the production process and improve overall efficiency, providing significant benefits to companies (Ahmad and Rahimi, 2022). However, despite the development of AI technology as an essential tool for improving efficiency and cost effectiveness, there is a challenge in acquiring labeled and well-organized data in the actual industrial field.

In general, sufficient training data for each label in a deep learning model leads to better classification performance by the neural network. When the training dataset is small, the sparsity of data can limit sample variability, and in most cases, it can result in a decrease in network performance (X. Li et al., 2020). According to Mehmood et al. (2020), when the available image data is limited, deep learning models can suffer from overfitting. They suggest that data augmentation can be used to extend imbalanced data and achieve excellent accuracy in multi-class classification. one of the biggest challenges in the development phase of modern application AI models such as Vision Transformers (ViT) and deep learning-based classification models is the difficulty of obtaining labeled manufacturing data from the industrial environment (J. P. Yun et al., 2020). Generally, data generated in real world industrial environments is often not shared with researchers due to security concerns and various constraints imposed by companies. Additionally, outsourcing the data labeling process can be costly, making it challenging to obtain satisfactory datasets (Imoto et al., 2019). We aim to use generative AI technology to efficiently augment and generate data for training AI models in the manufacturing sector. This will contribute

* Both authors contributed equally to this research.
† Corresponding Author

URI: https://hdl.handle.net/10125/106501
978-0-9981331-7-1

to improving the performance of data-driven services and, furthermore, have a positive impact on customer satisfaction with AI-based manufacturing industry services.

The main objective of this research is to construct a dataset by expanding the sample images assuming that a very small amount of data samples is obtained through manufacturing companies. More specifically, the following research tasks are performed.

- We undertake a novel approach to address the problem, which consists of three steps. The first step is to extract text from images using Bootstrapping Language-Image Pre-training (BLIP), a Vision-Language Pre-training (VLP) model that has recently gained traction with advances in generative AI. The second step involves performing slight fine-tuning using the Low Rank Adaptation of Large Language model (LoRA) approach on the Stable Diffusion model, which supports text-to-image synthesis using the extracted text and the original image as input (Hu et al., 2021; Rombach et al., 2022). The third step is data augmentation, which involves generating an image from the previously derived model.
- To ensure the usability of the augmented data in real world manufacturing scenarios, we evaluate its effectiveness through binary and multi-class classification using various outlier detection models. Ultimately, we find that the dataset constructed through these processes is highly valuable for binary and multi-class classification in manufacturing outlier detection models.

The structure of the paper is as follows. In the second section, we explain the fundamental theories, review prior research, and formulate hypotheses. In the third section, we describe our research process and propose a methodology. In the fourth section, we provide a detailed explanation of the empirical results along with performance analysis. In the final section, we discuss the research findings, propose implications, and explore potential avenues for contributing to the advancements in the relevant field.

2. Related work

2.1. Data imbalance challenges in manufacturing environment

Innovation in product and service development within the manufacturing industry is accelerating. Increasing production volume without compromising the existing quality is one of the key challenges in manufacturing production (Sharma et al., 2022). To enhance the accuracy and efficiency of processes, deep learning-based manufacturing defect classification can be effectively utilized. However, collecting a large quantity of images in real world environments can be challenging and time-consuming (Meister et al., 2021). To utilize a reliably performing neural network for welding defect detection in the metal manufacturing industry, it is essential to gather a sufficient amount of training data. However, due to the industry-specific nature where defects occur infrequently, obtaining an appropriate volume of visually confirmable data for specific defects is challenging. Furthermore, concerns related to a company's profitability or the exposure of core technologies can make data sharing difficult. Such data imbalance issues have the potential to degrade algorithm performance (J. P. Yun et al., 2020). To ensure this, it is crucial to collect a sufficient amount of training data.

2.2. Image data augmentation techniques

The performance of an AI model relies heavily on a large amount of training data available for training. Efforts are underway to address this issue by expanding the quantity of training data through data augmentation techniques (Chlap et al., 2021). Image data augmentation techniques can enhance both the quantity and quality of the training dataset, enabling the construction of deep learning models with higher performance. In S. Yun et al. (2019), it is stated that CutMix, using local dropout, allows for better generalization and localization by maximizing the utilization of removed regions. In particular, CutMix employs a method where the removed region is replaced with a patch from another image, rather than simply removing pixels. Xie et al. (2020) utilized rendering techniques to generate augmented object poses for data augmentation in 3D image datasets. This approach has demonstrated the capability to enrich training data substantially. This technique has gained attention, particularly in the field of autonomous driving, where the detection of 3D objects using LiDAR is essential. However, this method faces challenges in maintaining a balance between preserving the semantically meaningful information of the original image and generating diverse images (B. Li et al., 2023). To address this, Shorten and Khoshgoftaar (2019) applied GAN (Generative Adversarial Networks) as a basis and incorporated geometric transformations, color space augmentation, kernel filters, random erasing, image blending, and meta-learning methods to generate diverse images

while maintaining the balance between preserving semantically meaningful information. Furthermore, Neural Style Transfer introduced in Zheng et al. (2019) presented a method where the style of one image is applied to another image while preserving the content of the latter image. While this approach allows for preserving the semantically meaningful information of the original image, it still suffers from limitations in terms of diversity due to the constrained application of a limited set of styles.

2.3. Method for generating images with semantic preservation

Recently, there have been ongoing studies aiming to utilize the Stable Diffusion model for data augmentation while preserving semantic information. Stable Diffusion is a methodology that utilizes text to generate images (Sonkar et al., 2022). Stöckl (2023) demonstrated that images generated using stable diffusion not only encompass a wide range of concepts but also produce accurate images for a highly diverse set of expressions. Text is used as input in the training and generation processes to convert it into images. During the training process, the training data comprises image-caption pairs. The discriminator is used to train the generated images, which are transformed into numerical representations, to match the corresponding captions (Somepalli et al., 2023). In the generation process, the text is first transformed into word embeddings using Natural Language Processing (NLP) techniques, utilizing the Contrastive Language-Image Pre-Training (CLIP) text encoder (Wei et al., 2023). The text is converted into a noisy latent vector in the latent space using the diffusion method. Then, it is restored back to the image domain using the noise-image operations learned during the previous training. In this context, the latent space refers to a space where data is represented and distinguished based on its features. Afterward, the latent vector goes through a Variational Auto-Encoder (VAE) to transform it into a new image (Croitoru et al., 2023). Through the decoder of the VAE, the latent vector is decoded into the desired size, yielding the generated image as the output.

2.4. Text extraction from manufacturing images using BLIP

To train the Stable Diffusion model, text that describes the images is required (Kawar et al., 2023). If you only have image data and need to extract text, there are several methods available for extracting text from images. In this paper, BLIP was adopted as the method for extracting text from manufacturing image data. BLIP is a new VLP framework that enables the joint learning of understanding and generation tasks. By integrating multitask learning, which combines two tasks that could not be improved simultaneously, BLIP can perform a wide range of vision-language tasks and enhance the performance of each task. In the generation task, the caption synthesizer is used as a decoder, while in the understanding task, the caption filter serves as an encoder. Through this approach, the noise in text-based image datasets obtained from the web is reduced, ensuring performance across multiple tasks. BLIP is specialized in solving image captioning problems and possesses strong visual understanding capabilities. Therefore, it can effectively recognize the visual features and structure of manufacturing image data. For instance, BLIP can recognize and extract textual information from images, such as their shape, size, color, position, and texture. These diverse advantages make BLIP highly useful for extracting and transforming text from manufacturing images, thereby enhancing its performance in this context (J. Li et al., 2022).

2.5. LoRA fine tuning

Modern image generation models like Stable Diffusion can experience significant resource consumption due to their large model sizes. Therefore, there is a need for methods that conserve resources while achieving optimal efficiency in Large Language Models (LLM) and multimodal models. Hu et al. (2021) discusses methods for efficiently fine-tuning large-scale language models. This method involves compressing existing language models into smaller ones and rapidly fine-tuning them for new tasks. For instance, it adjusts a large model approximated with low-rank matrices using a language model trained for the manufacturing domain. This task involves additional training data and task-specific loss functions. Fine-tuning is only applied to these smaller models, and the underlying structure of the original LLM remains unchanged.

LoRA facilitates parameter transfer from large to small models, efficiently copying parameters from the large model to the smaller one, aiding in setting initial weights when fine-tuning begins. Additionally, it employs regularization techniques during the construction and fine-tuning of small models to prevent overfitting and improve generalization performance. By combining these ideas and techniques, LoRA offers an efficient way to fine tune LLM, particularly reducing computational and memory requirements. This simplifies and accelerates the deployment and improvement of language models for new tasks.

3. Methodology

3.1. Manufacturing image generation process

In this section, we describe the experimental process of expanding a very small amount of data obtained in the manufacturing domain into an AI training dataset. The data generation process proposed in this paper can be seen in Figure 1. Our idea lies in extracting text from a small amount of manufacturing image data and generating multiple images again through the extracted text. First, we secure a small-scale manufacturing image dataset and undergo data augmentation processes using it. Then, the augmented data is utilized in various functions and tasks of smart manufacturing. This may include tasks such as anomaly detection.

The data augmentation process can be divided into three steps. In the first step, text is extracted using existing image data. In the second step, the image and text are used as input to train the Stable Diffusion model, which is a large model. In this process, we follow the LoRA fine-tuning approach, where only the attention areas are trained without retraining the entire set of parameters. Finally, in the third step, data augmentation is performed by generating images using the fine tuned model.

3.2. Extracting text from images

The Original images in Figure 1 are datasets produced in the manufacturing process, consisting of leather, carpet, and wood image data, as described in more detail in section 4.2 of the paper. As explained in section 2.4, BLIP takes an image as input, extracts its features, and converts those features into text. The training structure of BLIP consists of a Multimodal Mixture of Encoder-Decoder (MED) framework, which combines three encoder-decoder structures with different roles. To filter meaningful captions based on the image, an understanding module is used, employing an image-grounded text encoder. The image-grounded text encoder passes through a cross-attention layer, utilizing image embedding information to output embeddings that indicate the degree of matching between the image and the text in the text encoder. In the final step, a text decoder designed for utilizing the image through dataset bootstrapping is located, aiming at image generation (J. Li et al., 2022). Through these stages, the necessary text for image generation is extracted.

Using the proposed model, we extracted the text from the leather image as shown in Figure 1, which states ”a close-up of a brown leather surface with a small piece of a small object”. Preprocessing the folded characteristic of the leather data as ”a small piece of a small object” contributes to generating more accurate images in the subsequent image generation model.

3.3. Efficient fine-tuning with LoRA for Stable Diffusion

In this paper, we adopt the Stable Diffusion as mentioned in section 2.3 as our image generation model. This model possesses the ability to learn and comprehend the relationship between text descriptions and images, while maintaining semantic coherence according to the given descriptions during the image generation process (Rombach et al., 2022). Furthermore, during the generation of images, the images are manipulated according to textual descriptions, and when the textual descriptions are modified, the images respond accordingly. This enables interactive and flexible image generation in various application domains. This model models the interaction between images and text, effectively leveraging both types of information to generate more diverse and rich results. For these reasons, we have chosen Stable Diffusion for image generation tasks based on manufacturing image data and textual descriptions.

The Stable Diffusion model is difficult to use in manufacturing environments with limited resources and data due to its large size. Therefore, we employ LoRA, which facilitates low-rank adaptation of LLM. This technique enables efficient fine-tuning of existing large models and allows them to adapt to specific tasks or domains. More detailed information about LoRA can be found in the existing research (Hu et al., 2021), and it is discussed in section 2.5. Step 2 of Figure 1 takes the extracted text and original image as input through the section 3.2 process, and performs LoRA fine-tuning on the Stable Diffusion model. The ’Fine tuned model’ in Figure 1 becomes a data augmentation model that helps address the problem of data scarcity in the manufacturing ecosystem.

3.4. Data augmentation using Stable Diffusion model

The third step in the data generation process involves generating augmented data using a fine tuned Stable Diffusion model on existing images. To generate images using the Stable Diffusion model, a prompt that describes the image is required. The prompt used in this case is obtained from the caption text extracted through BLIP from the original image in step 1, as shown in Figure 1. The reason for adopting the given caption as a prompt is that it was used together with the Stable Diffusion model during the training process, and it is the

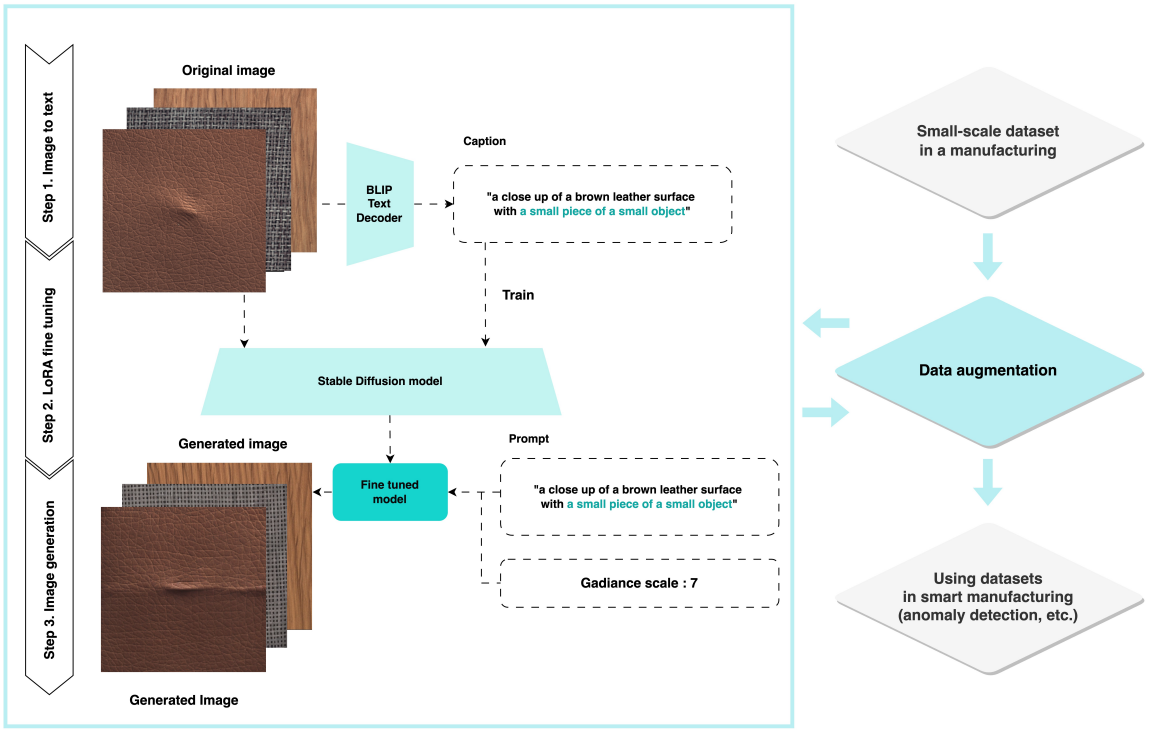

Figure 1. Manufacturing data augmentation framework.

text that best describes the manufacturing image data to be generated. The guidance scale in step 3 affects the stability of the images, and a smaller guidance scale leads to more consistent quality. A higher value will have a greater impact on the text. In this paper, based on previous research, several values were experimentally tested through hypotheses, and a guidance scale of 7, which yielded the best results, was chosen for the image generation process.

The data augmentation technique proposed in this thesis aims to construct a model that understands specific data well and augment images through that model. The generated model through this approach can generate data with distinct features and extract images in a more natural form. In the following section, we classify real data using the images generated through the data augmentation method proposed in this paper. The proposed approach is quantitatively demonstrated to be meaningful.

4. Performance analysis

In this section, we expand the dataset through our approach. We demonstrate the effectiveness of our approach through experiments on detecting defects in manufacturing data using an image multi-classifier.

4.1. Experiment environment

In this study, the experimental setup was as follows. An Intel(R) Xeon(R) CPU was used for the CPU, and the RAM was configured to 32GB. The Graphics Card used was V100. Additionally, for this experiment, Convolutional Neural Network (CNN) and ViT were implemented through the Keras Python library.

4.2. Datasets

We utilized the MVTec-AD Dataset, which consists of real world data generated within manufacturing processes. This dataset was publicly released for training manufacturing AI in Bergmann et al. (2019), and numerous studies have recently been conducted using it. The dataset contains diverse data categorized into texture and object categories. Among the texture data, we used leather, carpet, and wood, while for the object data, we attempted data augmentation using screws. These data were acquired using high resolution industrial RGB sensors, capturing images of 2048 $*$ 2048 size , which were then cropped to 1024 $*$ 1024 size. Most of the image data was obtained under controlled environmental conditions, particularly in controlled lighting environments.

Table 1 to 4 provide detailed descriptions of the original dataset and the augmented dataset used in this paper. Information regarding each category and class can be inferred from the table names. The image length is consistently 1024 across all categories, and the augmented images have been expanded to 50 times the size of the original dataset. In this study, we experimented with data augmentation using minimal data, assuming the case of collecting data in small-scale businesses where obtaining data is difficult.

Table 1. Description of screw dataset: (a) manipulated front, (b) scratch head, (c) scratch neck, (d) good, (e) Image length.

	(a)	(b)	(c)	(d)	(e)
Original data	24	24	26	41	1024
Augmented data	1200	1200	1300	2050	1024

Table 2. Description of leather dataset: (a) cut, (b) fold, (c) glue, (d) good, (e) Image length.

	(a)	(b)	(c)	(d)	(e)
Original data	19	17	19	32	1024
Augmented data	950	850	950	1600	1024

Table 3. Description of wood dataset: (a) color, (b) hole, (c) liquid, (d) good, (e) Image length.

	(a)	(b)	(c)	(d)	(e)
Original data	8	10	10	19	1024
Augmented data	400	500	500	950	1024

Table 4. Description of carpet dataset: (a) color, (b) cut, (c) hole, (d) good, (e) Image length.

	(a)	(b)	(c)	(d)	(e)
Original data	19	17	17	28	1024
Augmented data	950	850	850	1400	1024

4.3. Defective image data augmentation

In this paper, we propose a method for augmenting data using Stable Diffusion and LoRA fine-tuning approaches to address the issues of data scarcity and imbalance. We employ the stable-diffusion-v1-5 pretrained model for data augmentation, and generate captions for fine-tuning the Stable Diffusion model using the BLIP model. In BLIP, we set the batch size to 8, and the minimum and maximum lengths to 15 and 50, respectively. In order to generate data in the tuning model, VAE utilized the Denoising Diffusion Implicit Model (DDIM), with a batch size set to 4.

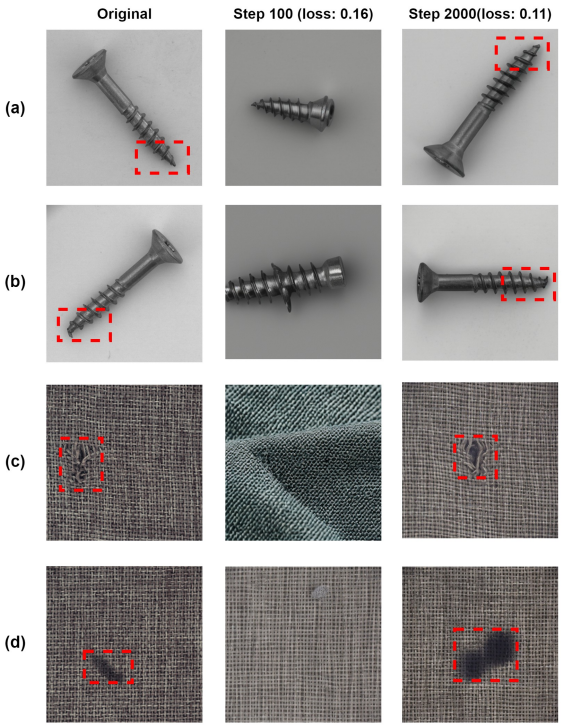

Figure 2. Examples of the data generator outputs as the number of steps changes, (a) good, (b) manipulated front, (c) cut, (d) color

Alongside the original images, we trained a pretrained model with captions extracted using BLIP. During this process, the epoch was set to 30. Figures reveal the generation of images that closely resemble the trained images, depending on the number of steps for fine-tuning the Stable Diffusion pretrained model. In Figure 2, (a) represents normal data, while (b) depicts a modified frontal image with slight defects (such as foreign objects and minor bending at the front of the screw). Figure (c) displays cut defect data in a carpet, and (d) showcases defect images with stains. When the step count is 100, all four cases (a), (b), (c), and (d) fail to faithfully reproduce the distinctive features of the original images. The shapes of screws and carpets can be identified in the generated images.

However, due to significant differences from the original images, there might be issues with utilizing these images as augmented data. Particularly in the case of (a), the length of the screw object has anomalously shortened, which could adversely impact results when employing such augmented data. In the case of (b), where the front part of the screw should be bent or distorted to reflect reality, the red area of the

original image did not generate correctly at the 100-step. Similarly, in cases (c) and (d), the normal generation of the red area from the original image did not occur. On the other hand, as the training steps were iterated, when the number of steps exceeded 2000, we could observe that the red dashed rectangle region of the original images was well captured. In the case of (a) data, no bending was found in the front part, whereas for (b) data, upon inspecting the red dashed box, we can notice a subtle deformation of the front part to one side. In (c), the cut defect area in the carpet is accurately represented, and similarly in (d), it can be confirmed that stains are properly generated within the red dashed rectangular area. Therefore, with an increasing number of steps, it becomes evident that the model can generate data that faithfully reflects the characteristics of the original data. Additionally, the decreasing loss values from 0.16 to 0.11 indicate that the model is progressively fine tuned effectively with the increasing steps, ensuring a suitable fit to the training data.

Figure 3 presents the results of data augmentation through fine-tuning using the approach proposed in this paper. The 'Original' area in Figure 3 represents the original image data, while the 'Generated images' consist of 24 data points generated through the fine-tuning model for screw, leather, and wood. The alphabets in Figure 3 denote the characteristics of each data point, and the explanations for these characteristics can be found in the title of Figure 3. It is evident that the images have been generated according to their respective features. In (a), scratches are observed on the neck part of the screw, and in (b), the end part of the screw appears bent. In the case of (c), the folded characteristics of leather are well reflected, while in (d), glue can be seen on top of the leather. Furthermore, in (e), water droplets are visible on the surface of the wood. When examining the hole images, one can observe that holes are naturally generated in various forms within the wood. In the case of (g), the normal fabric structure of the carpet is observed. Furthermore, in (h), it can be confirmed that the defect area, a hole within the carpet, has been appropriately formed. Therefore, the proper augmentation of data can be visually confirmed through Figure 3. When examining the augmented images, it can be observed through Figure 3 that the Texture Data set generates augmented images that are relatively more natural compared to the Object Data set.

4.4. Experimental results on data classification

We implemented binary classification and multi-classification using CNN and ViT for the performance evaluation of augmented data. CNN extracts features while preserving spatial information of the image through convolution and pooling layers and performs multi-classification (Krizhevsky et al., 2017). Additionally, the ViT model is an application of transformer's self-attention, which has shown excellent performance in the NLP field, to vision tasks (Dosovitskiy et al., 2020). ViT model achieves better performance when pretrained with a sufficient dataset and fine tuned with specific domain data. However, it has the disadvantage of not guaranteeing performance when sufficient dataset is not available. We utilized the ViT-B/16 model as the pretrained model and trained it using augmented data to demonstrate the performance of the proposed approach. For evaluating the proposed approach, we used accuracy, f1 Score, recall, and precision as evaluation metrics for the multi-classification model. Table 5 describe how these metrics are calculated, where TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false negative.

Table 5. Performance metrics.

Metrics	Formula
Accuracy	$\frac{TN+TP}{TN+TP+FN+FP}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1 score	$2 \times \left(\frac{Precision \times Recall}{Precision+Recall}\right)$

The detailed information of the experiment can be found below. Initially, the experiments were conducted with over 50 epochs, but it showed a tendency to converge within 20 iterations, so we set the number of epochs to 20 for all the data. We used the Adam optimizer (Kingma and Ba, 2014). Additionally, the target image size was set to $256 * 256$, and the learning rate was set to 0.0001. The training was performed with a batch size of 128. For the ViT model, a patch size of 16 was used. The train and validation data ratio of augmented image data has been composed using the commonly employed 8:2 ratio. The entire original dataset has been used for the test data.

Table 6 below presents the experimental results of training and testing ViT and CNN classifiers using augmented data. As explained in Table 1 to 4, each data group used in the experiments is categorized into three types of defective data and good data. The leather, wood, and screw data were used to train a multi-classification model, while the carpet data was used to train a binary classification model for evaluation and verification of the results.

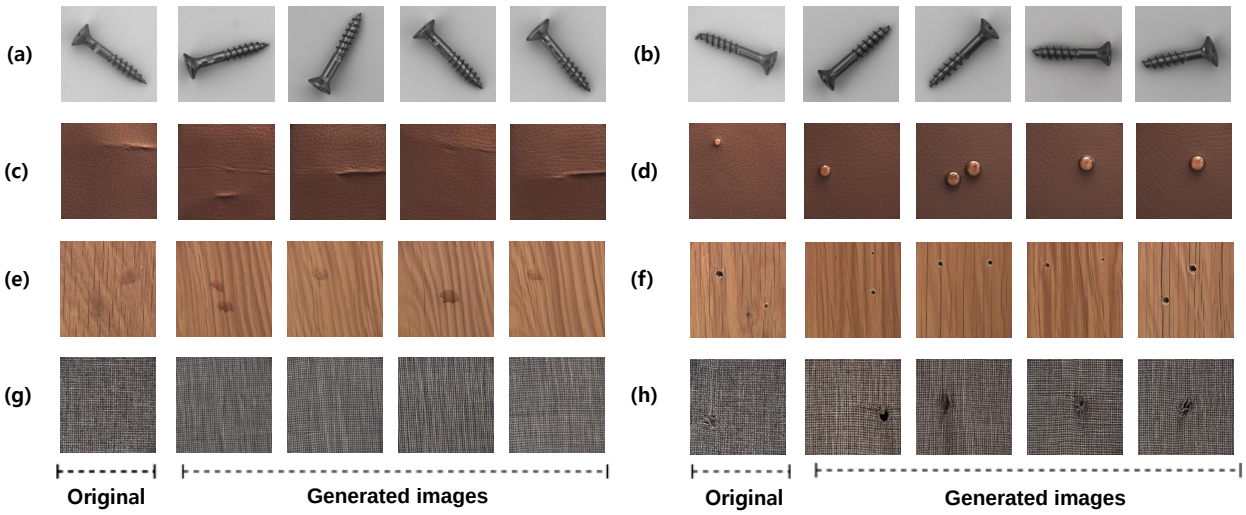

Figure 3. Samples of real data and generated dataset: (a) scratch neck, (b) manipulated front, (c) fold, (d) glue, (e) liquid, (f) hole, (g) good, (h) hole.

Table 6. Segmentation result(Accuracy/F1 score/Recall/Precision) of four image classification datasets.

		Leather	Wood	Screw	Carpet
	Type	Multi-class	Multi-class	Multi-class	Binary-class
CNN	Accuracy	0.67	0.61	0.54	0.6
	F1 score	0.64	0.57	0.51	0.57
	Recall	0.59	0.89	0.52	0.6
	Precision	0.69	0.42	0.61	0.56
ViT	Accuracy	0.71	0.78	0.75	0.79
	F1 score	0.68	0.73	0.73	0.74
	Recall	0.66	0.77	0.64	0.74
	Precision	0.81	0.73	0.84	0.75

Table 6 presents the results of training multi-class and binary classification models using augmented data and testing with original field-real data. For Leather, ViT achieves an accuracy of 0.71, while CNN achieves 0.67. The F1 Score for Leather is 0.68 for ViT and 0.64 for CNN. Similarly, for Wood, ViT obtains a superior result of 0.78 compared to CNN's 0.61. In terms of F1 Score, ViT obtains 0.73. In the case of Screw, ViT's accuracy of 0.75 is 0.21 higher than CNN's accuracy of 0.54 for this data. The F1 Score for the Screw data is 0.73 for ViT and 0.51 for CNN, indicating that ViT achieved a superior outcome. For the Carpet data, a binary classification was conducted, and while CNN scored 0.6, ViT attained a higher accuracy of 0.79. The F1 Score favored ViT with a value of 0.74 over CNN's 0.57. ViT exhibited a Recall of 0.74 and a Precision of 0.75, while CNN achieved a Recall of 0.6 and a Precision of 0.56. As evident from the following figures, it can be observed that ViT exhibits superior performance to CNN in terms of both Recall and Precision. Through various metrics displayed in Table 6, it can be confirmed that ViT outperforms CNN overall.

5. Discussion and conclusions

This paper proposes a data augmentation method to address the issue of data scarcity in the manufacturing ecosystem. The existing methods have been limited by significant resource consumption and challenges in collecting high-quality data. To overcome these limitations, our research presents a method for generating an image dataset using a small number of defective data, which can be utilized for training and fine-tuning AI models. The benefits our approach offers in various manufacturing service aspects are as follows. **Quality enhancement:** Employing generative AI for data augmentation, similar to our

approach, enhances the performance of models and systems utilized in manufacturing processes. This data augmentation leads to improved product quality and reduced defects, ultimately assisting manufacturing companies in delivering high-quality products to customers. **Production process optimization:** Another valuable outcome of our approach is the ability to monitor and optimize the production process using generative AI. For instance, augmenting data generated on the production line provides more real-time insights, enabling optimization of production volumes or reduction in energy consumption. **Automation and agility:** Data augmentation through generative AI enhances automation and agility within the production process. Real-time data generation and analysis enable continuous monitoring of production line status and swift responses to issues. This increases flexibility in production schedules and aids in agile management of manufacturing processes. **Cost savings:** Our proposed method for data augmentation allows for the acquisition of higher-quality data, enhancing resource efficiency and cost savings. This minimizes unnecessary experiment and data collection costs while reducing waste in the production process. **Enhanced customer satisfaction:** Leveraging generative AI to improve product quality, service customization, and reliability leads to enhanced customer satisfaction. Strengthened customer relationships and long-term customer loyalty are achievable outcomes. In summary, utilizing generative AI for data augmentation is a powerful tool for enhancing various aspects of smart manufacturing services. Through this approach, manufacturing companies can maintain competitiveness and succeed in the evolving landscape of smart manufacturing.

The limitations of this thesis can be summarized as follows. First, the performance, biases, strengths, and weaknesses of text-to-image models such as Stable Diffusion can significantly influence the quality and characteristics of the augmented data. Second, the performance of the image to text model needs further improvement. Finally, the complete elimination of human intervention, such as the developer's involvement in the data preprocessing stage, has not been achieved during the text extraction process. Accordingly, as a future research direction, it is necessary to explore methods to improve the performance of image generation models or minimize the impact of the models on the augmented data. Therefore, as a future research direction, we need to explore ways to improve the performance of the image generation model or minimize its influence. And we should conduct research on improving the performance of models that extract text from images, such as BLIP. Based on this, we can develop automated software methodologies and frameworks to minimize human intervention. Research involving the introduction of eXplainable Artificial Intelligence (XAI) to enhance the reliability of augmented data or exploring improvements in various tasks beyond anomaly detection using augmented data can be conducted. The data augmentation method proposed in this paper is expected to serve as a valuable tool for addressing data scarcity issues within the manufacturing ecosystem.

References

Ahmad, H. M., & Rahimi, A. (2022). Deep learning methods for object detection in smart manufacturing: A survey. *Journal of Manufacturing Systems*, *64*, 181–196.

Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9592–9600.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, *65*(5), 545–563.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Imoto, K., Nakai, T., Ike, T., Haruki, K., & Sato, Y. (2019). A cnn-based transfer learning method for defect classification in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, *32*(4), 455–459.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., & Irani, M. (2023). Imagic: Text-based real image editing

with diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Li, B., Wang, X., Xu, X., Hou, Y., Feng, Y., Wang, F., & Che, W. (2023). Semantic-guided image augmentation with pre-trained models. *arXiv preprint arXiv:2302.02070*.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International Conference on Machine Learning*, 12888–12900.

Li, X., Zhang, W., Ding, Q., & Sun, J.-Q. (2020). Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation. *Journal of Intelligent Manufacturing*, *31*, 433–452.

Mehmood, A., Maqsood, M., Bashir, M., & Shuyuan, Y. (2020). A deep siamese convolution neural network for multi-class classification of alzheimer disease. *Brain sciences*, *10*(2), 84.

Meister, S., Möller, N., Stüve, J., & Groves, R. M. (2021). Synthetic image data augmentation for fibre layup inspection processes: Techniques to enhance the data set. *Journal of Intelligent Manufacturing*, *32*, 1767–1789.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Sharma, M., Luthra, S., Joshi, S., & Kumar, A. (2022). Implementing challenges of artificial intelligence: Evidence from public manufacturing sector of an emerging economy. *Government Information Quarterly*, *39*(4), 101624.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1–48.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., & Goldstein, T. (2023). Diffusion art or digital forgery? investigating data replication in diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6048–6058.

Sonkar, S., Liu, N., & Baraniuk, R. G. (2022). A visual tour of current challenges in multimodal language models. *arXiv preprint arXiv:2210.12565*.

Stöckl, A. (2023). Evaluating a synthetic image dataset generated with stable diffusion. *International Congress on Information and Communication Technology*, 805–818.

Tercan, H., & Meisen, T. (2022). Machine learning and deep learning based predictive quality in manufacturing: A systematic review. *Journal of Intelligent Manufacturing*, *33*(7), 1879–1905.

Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. *Journal of manufacturing systems*, *48*, 144–156.

Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., & Zuo, W. (2023). Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*.

Xie, L., Xiang, C., Yu, Z., Xu, G., Yang, Z., Cai, D., & He, X. (2020). Pi-rcnn: An efficient multi-sensor 3d object detector with point-based attentive cont-conv fusion module. *Proceedings of the AAAI conference on artificial intelligence*, *34*(07), 12460–12467.

Yun, J. P., Shin, W. C., Koo, G., Kim, M. S., Lee, C., & Lee, S. J. (2020). Automated defect inspection system for metal surfaces based on deep learning and data augmentation. *Journal of Manufacturing Systems*, *55*, 317–324.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zheng, X., Chalasani, T., Ghosal, K., Lutz, S., & Smolic, A. (2019). Stada: Style transfer as data augmentation. *arXiv preprint arXiv:1909.01056*.