Multimodal sensor-to-machined surface image diffusion for defect detection in industrial processes

Jae Gyeong Choi[1], Yun Seok Kang[2], Hyung Wook Park[3], Sunghoon Lim[4]

[1,2,3,4] *50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan 44919, Republic of Korea*
choil6043@unist.ac.kr
yskang@unist.ac.kr
hwpark@unist.ac.kr
sunghoonlim@unist.ac.kr

Abstract

Generative models, particularly diffusion-based approaches, have gained significant attention recently due to their ability to create realistic outputs. Despite their potential, the application of these models in manufacturing remains largely unexplored. This work presents a framework that addresses this gap by generating machined surface images guided by multiple sensor inputs in manufacturing. The proposed model integrates information from multiple sensors with varying sampling rates using multimodal embedding and employs a latent diffusion model to translate the fused sensor embedding into an image embedding, which is then converted into a machined surface image. The effectiveness of the framework is validated using real-world time-series data, including force, torque, acceleration, sound, collected from various industrial processes, such as a carbon-fiber-reinforced plastic drilling process. The results demonstrate the model's ability to predict defects from the generated machined surface images. The proposed approach can potentially revolutionize prognostics and health management (PHM) in smart manufacturing by enabling sensor-guided visual inspection, defect detection, process monitoring, and predictive maintenance.

1. Introduction

Prognostics and health management (PHM) have emerged as a critical aspect of modern manufacturing to improve system reliability, reduce maintenance costs, and minimize unplanned downtime (Lei et al., 2018). Accurately detecting defects is crucial for effective PHM, enabling proactive maintenance and preventing potential failures. However, relying solely on a single sensor or limited modalities for defect detection can be challenging due to the complex nature of manufacturing systems (Choi et al., 2024). Moreover, traditional approaches that provide binary predictions (i.e., defect present or absent) or numerical estimates of defect severity may not offer sufficient insight into the specific nature and location of the defects. These limitations hinder the effectiveness of PHM, as they do not provide a comprehensive understanding of the system's health status and may not enable targeted maintenance actions.

Recent advancements in generative models, particularly diffusion-based approaches, have led to significant breakthroughs in various domains, including text-to-image generation (Ramesh et al., 2022; Rombach et al., 2022). These models have demonstrated remarkable performance in capturing complex data distributions and generating high-fidelity samples. While they have shown outstanding performance in creating realistic images from textual descriptions, their potential in manufacturing applications has not been fully explored. In manufacturing sites, the ability to generate accurate visual representations of machined surfaces based on sensor data can significantly benefit process monitoring and predictive maintenance, highlighting potential defects and facilitating effective PHM.

This work proposes Sensor2Image++, a framework for generating machined surface images guided by multimodal sensor inputs. Building upon the success of the previous work, Sensor2Image (Choi et al., 2023), which translates single sensor data into images, Sensor2Image++ excels in synthesizing high-fidelity machined surface images while effectively addressing the challenge of integrating information from multiple sensors with varying sampling rates. This enhancement ensures that our model comprehensively captures the intricacies of the manufacturing process, thereby advancing the state-of-the-art in sensor-guided image synthesis. By providing a powerful tool for PHM, Sensor2Image++ has the potential to significantly enable defect detection, process monitoring, and predictive maintenance.

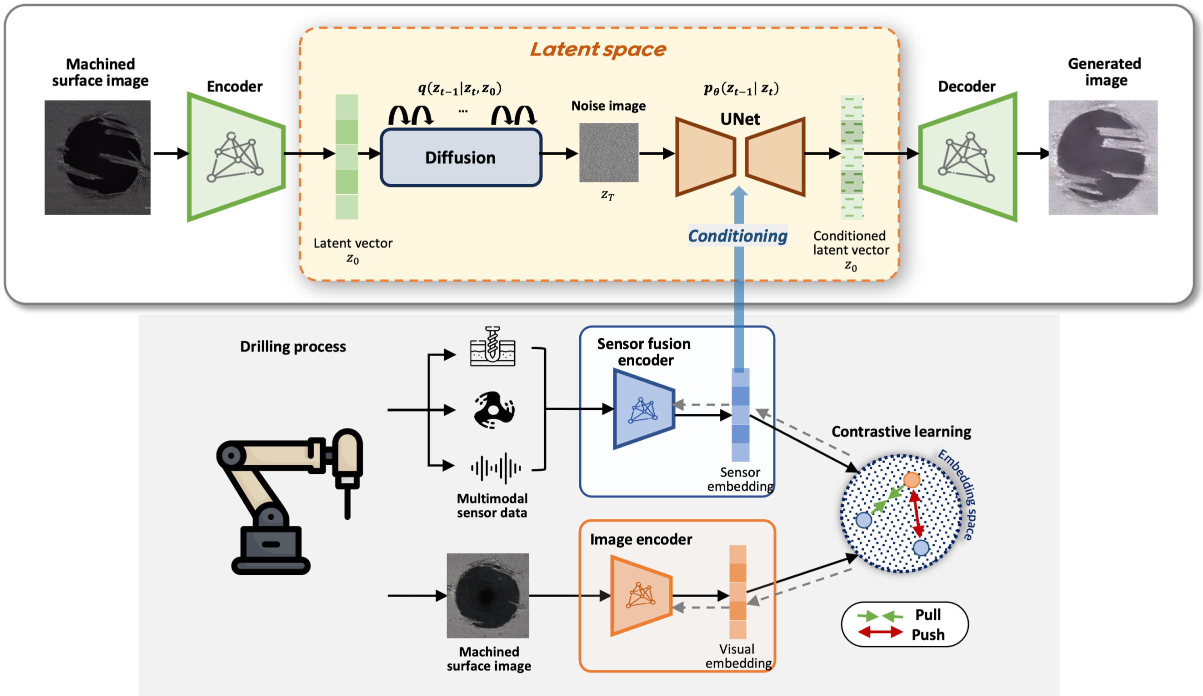

Figure 1. Sensor2Image++ for machined surface image synthesis using multimodal sensor data with varying sampling rates

2. PROBLEM STATEMENT

The accurate and timely detection of defects is paramount for the effective PHM of manufacturing processes. However, relying solely on sensor data for defect detection can be challenging due to the complex nature of manufacturing systems and the varying sampling rates of different sensors. Integrating information from multiple sensors and generating visual representations of machined surfaces can provide valuable insights for defect detection and predictive maintenance. The challenge lies in developing a framework that can effectively fuse multimodal sensor data and generate realistic machined surface images highlighting defects.

3. EXPECTED NOVEL CONTRIBUTIONS

The proposed research aims to make the following novel contributions to the field:

- Introduction of Sensor2Image++, a novel framework that generates machined surface images guided by multimodal sensor inputs, enabling enhanced process monitoring, predictive maintenance, and PHM in smart manufacturing through sensor-guided visual inspection and defect detection
- A versatile framework that enables the generation of high-fidelity surface images from real-world time series sensor data, spanning various industrial processes, such as carbon-fiber-reinforced plastic drilling, milling, and trimming. The proposed model's capability to accurately capture process-induced defects, including delamination, tool wear, and surface roughness, highlights its potential for broad applicability across diverse manufacturing domains.

4. METHOD

4.1. A latent diffusion model for machined surface image synthesis

The proposed approach employs a latent diffusion model architecture for generating images of machined surfaces as shown in Fig. 1. The model is constructed upon a variational autoencoder (VAE) framework comprising an encoder and a decoder, a diffusion process, and a U-Net denoiser. The encoder maps the input machined surface image x_0 to a latent representation z_0 in the latent space. The input image of the machined hole surface, x_0, is fed into the encoder $E_\theta(\cdot)$, which maps the input to a latent representation z_0 in the latent space $\mathcal{Z}$. The diffusion process gradually adds noise to z_0 over T steps to obtain a noisy latent representation z_T. The U-Net architecture is employed as the denoising function $p_\theta(z_t|z_{t-1}, z_0)$ to predict and remove the noise at each step.

To generate a single machined surface image corresponding to multiple sensor inputs collected from a process, it is essential that the generative model is capable of producing deterministic outputs. To achieve this, the denoising diffusion implicit models (DDIM) sampling process (Song et al., 2022) is employed. DDIM introduces a deterministic schedule for the

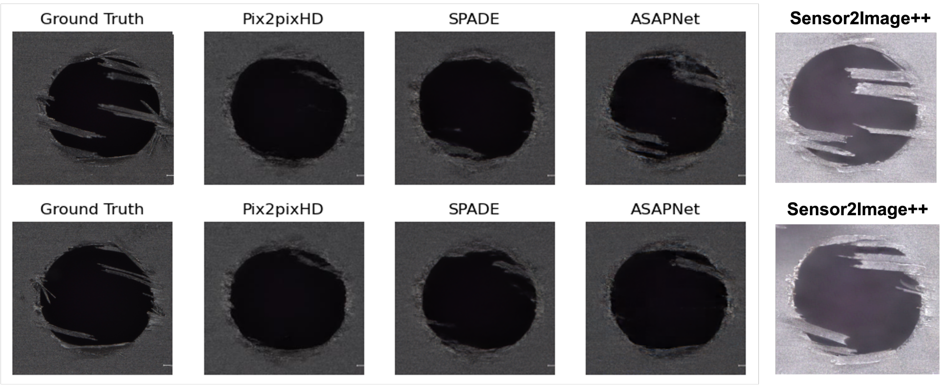

Figure 2. Qualitative comparison of various baselines in sensor-to-image tasks.

denoising process, allowing consistent and reproducible images to be generated. The DDIM sampling process is defined by a sequence of latent variables denoted by $z_1, z_2, \ldots, z_T$, where T is the total number of diffusion steps. The latent variable z_t is obtained by iteratively applying the DDIM update rule:

$$\begin{aligned} z_{t-1} =& \sqrt{\bar{\alpha}_{t-1}} \left(\frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t)}{\sqrt{\bar{\alpha}_t}} \right) \\ &+ \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(z_t, t) \end{aligned} \tag{1}$$

where $\bar{\alpha}_t$ is a deterministic variance schedule, and $\epsilon_\theta(z_t, t)$ is a learned denoising function that predicts the noise added at each step.

The decoder $D_\phi(\cdot)$ takes the noisy latent representation z_T and reconstructs the generated image $\hat{x}_0 = D_\phi(z_T)$. By training the model to denoise the latent space, the underlying structure and characteristics of the machined surface images are effectively captured, enabling the generation of realistic and diverse samples.

4.2. Conditional multimodal sensor embedding

The conditioning of multimodal sensor embedding to a latent diffusion model involves integrating information extracted from various sensors into the generative process, thereby influencing the generation of machined surface images. To achieve this, sensor and image encoders that extract features from sensor data and machined surface images are introduced. The sensor fusion encoder processes time-series data from various sensors, such as force, torque, acceleration, and sound, with different sampling rates. In contrast, the image encoder extracts visual features from the machined surface images, thereby capturing the surface quality and potential defects.

Let $D = \{(S_i, I_i)\}_{i=1}^N$ be a dataset consisting of N pairs of sensor data frames S_i and their corresponding images I_i. The primary objective is to train a sensor fusion encoder $f_S(\cdot)$ that extracts informative features z_S from the sensor data, such that they are well-aligned with the features z_I extracted from the images using an image encoder $f_I(\cdot)$. Given the dataset, D, the sensor features are computed as $z_S = f_S(S)$, and the image features as $z_I = f_I(I)$, where both z_I and z_S are vectors in the same dimensional space. This approach enables learning aligned features across different modalities, resulting in a shared sensor-to-image embedding space.

To ensure that the extracted sensor and image embedding are closely aligned for the same experimental set, contrastive learning is employed to train the sensor fusion encoder and image encoder. The objective is to minimize the distance between the embeddings in the embedding space for corresponding sensor-image pairs while maximizing the distance for non-corresponding pairs. This is achieved by employing a contrastive loss function, such as InfoNCE (Chen et al., 2020), which prompts the model to learn a representation space in which similar samples (i.e., sensor-image pairs from the same experimental set) are proximal while dissimilar samples are distal.

5. PROPOSED RESEARCH PLAN

Data collection and preprocessing. The research aims to collect various industrial process data from various manufacturing domains. The dataset will include time-series sensor data and surface images. The data collection process will focus on acquiring a comprehensive and representative dataset covering various manufacturing processes and materials. This extensive dataset will enable the development of a robust and generalized framework for sensor-guided visual inspection and defect detection. The collected data will undergo preprocessing and synchronization to ensure compatibility with the proposed model, facilitating seamless integration and analysis.

Multimodal sensor embedding. A sensor fusion and image encoder will be developed to extract features from sensor data and machined surface images. For the multimodal sensor data

S, which includes various sensor inputs such as force, torque, acceleration, and sound, we use a sensor fusion encoder $f_S(\cdot)$ to extract a sensor embedding:

$$\boldsymbol{z_S} = f_S(W \cdot [s_1, s_2, ..., s_m]) \tag{2}$$

Given a machined surface image I, the image encoder $f_I(\cdot)$ maps it to a latent vector $\boldsymbol{z_0}$ in the latent space:

$$\boldsymbol{z_I} = f_I(I) \tag{3}$$

The encoders will be trained using contrastive learning to ensure that the extracted sensor and image embeddings are closely aligned.

$$\begin{aligned} \mathcal{L}^{\text{SupCon}} &= \sum_{i=1}^{2N} \ell_i^{\text{SupCon}} \\ \ell_i^{\text{SupCon}} &= -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{\sum_{a=1}^{2N} \mathbf{1}_{i \neq a} \cdot \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau)} \end{aligned} \tag{4}$$

Here, N denotes the batch size, and $2N$ represents the total number of original and augmented samples within a batch. The feature vector z_i concatenates the multimodal sensor and image embedding. The scalar temperature parameter τ belongs to the set of positive real numbers, denoted as $\tau \in R^+$. The index $i \in \{1, \ldots, 2N\}$ represents an anchor. $P(i)$ denotes the set of all samples belonging to the same process trial as the anchor i. The index $a \in \{1, \ldots, 2N\} \setminus \{i\}$ represents all possible indices in the batch, excluding the anchor i.

Latent diffusion model development. A latent diffusion model will be built on a VAE framework with an encoder and a decoder. DDIM will create consistent and deterministic machined surface images corresponding to time series sensor data. Forward process:

$$\begin{aligned} &q_\sigma\left(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{z}_0\right) \\ &= \mathcal{N}\left(\sqrt{\alpha_{t-1}} \boldsymbol{z}_0 + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \frac{\boldsymbol{z}_t - \sqrt{\alpha_t} \boldsymbol{z}_0}{\sqrt{1 - \alpha_t}}, \sigma_t^2 \boldsymbol{I}\right) \end{aligned}$$

where $\boldsymbol{z_0} = f_\theta^{(t)}\left(\boldsymbol{z_t}\right) := \left(\boldsymbol{z_t} - \sqrt{1 - \alpha_t} \cdot \epsilon_\theta^{(t)}\left(\boldsymbol{z_t}\right)\right) / \sqrt{\alpha_t}$, $\alpha_t = \prod_{s=1}^t (1 - \beta_s)$, and $\epsilon_\theta(\boldsymbol{z_t}, t, \boldsymbol{z_S})$ is a noise estimation network that predicts the noise added at each timestep, conditioned on the sensor embedding z_S.

Generative process:

$$\boldsymbol{p_\theta}^{(t)}\left(\boldsymbol{z_{t-1}} \mid \boldsymbol{z_t}\right) = \begin{cases} \mathcal{N}\left(f_\theta^{(1)}\left(\boldsymbol{z}_1\right), \sigma_1^2 \boldsymbol{I}\right) & \text{if } t = 1 \\ q_\sigma\left(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, f_\theta^{(t)}\left(\boldsymbol{z}_t\right)\right) & \text{otherwise,} \end{cases} \tag{5}$$

The noise estimation network $\epsilon_\theta(z_t, t, z_S)$ is trained to minimize the following loss function:

$$L := E_{\mathcal{E}(x), y, \epsilon, t}\left[\left\|\epsilon - \epsilon_\theta\left(\boldsymbol{z_t}, t, \boldsymbol{z_s}\right)\right\|_2^2\right] \tag{6}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is a standard Gaussian noise vector.

During inference, the DDIM sampler starts from a random Gaussian noise vector $z_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises it using the learned noise estimation network, conditioned on the sensor embedding:

$$\begin{aligned} \boldsymbol{z_{t-1}} = \sqrt{\alpha_{t-1}} &\left(\frac{\boldsymbol{z_t} - \sqrt{1 - \alpha_t} \epsilon_\theta(\boldsymbol{z_t}, t, \boldsymbol{z_s})}{\sqrt{\alpha_t}} \right) \\ &+ \sqrt{1 - \alpha_{t-1}} \epsilon_\theta(\boldsymbol{z_t}, t, \boldsymbol{z_s}) \end{aligned}$$

Finally, the denoised latent vector $\hat{z_0}$ is passed through the decoder (generator) $D_\phi(\cdot)$ to obtain the reconstructed machined surface image:

$$\hat{I} = D_\phi(\hat{z_0}) \tag{7}$$

Model training and evaluation. The proposed Sensor2Image++ framework will be trained using the collected dataset. The model's performance will be evaluated using qualitative metrics such as peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and its ability to predict defects such as delamination from the generated images.

Comparison with existing methods. The performance of Sensor2Image++ will be evaluated in comparison with existing methods in terms of image quality and sensor data utilization to demonstrate its superiority.

6. CONCLUSION AND FUTURE WORKS

The proposed Sensor2Image++ framework has the potential to significantly advance PHM in manufacturing by enabling sensor-guided visual inspection, defect detection, and predictive maintenance. By leveraging multimodal learning and diffusion models, the framework can generate high-fidelity machined surface images highlighting potential defects, providing valuable insights for PHM. Furthermore, the proposed model combines a latent diffusion model conditioned on multimodal sensor embedding to effectively integrate information from heterogeneous sensor data with different sampling rates and generate realistic machined surface images. The successful completion of this research will provide a powerful tool for improving system reliability, reducing maintenance costs, and minimizing unplanned downtime in smart manufacturing.

Future work will extend the framework to accommodate a broader range of sensor modalities and industrial processes, further enhancing its applicability and impact in PHM. Furthermore, integrating the proposed framework with existing

PHM methods, such as remaining practical life estimation and fault diagnosis, will be investigated to develop a comprehensive and robust PHM solution for smart manufacturing.

REFERENCES

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A simple framework for contrastive learning of visual representations.*

Choi, J. G., Kim, D., Chung, M., Park, H. W., & Lim, S. (2023). Sensor to machined surface image generation in cfrp drilling. In *Iise annual conference and expo.*

Choi, J. G., Kim, D. C., Chung, M., Lim, S., & Park, H. W. (2024). Multimodal 1d cnn for delamination prediction in cfrp drilling process with industrial robots. *Computers & Industrial Engineering*, *190*, 110074.

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to rul prediction. *Mechanical Systems and Signal Processing*, *104*, 799-834.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical text-conditional image generation with clip latents.*

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-resolution image synthesis with latent diffusion models.*

Song, J., Meng, C., & Ermon, S. (2022). *Denoising diffusion implicit models.*