Multimodal sensor-guided diffusion model for machined surface image synthesis

Jae Gyeong Choi
Ulsan National Institute of Science and Technology
Ulsan, Republic of Korea
choil6043@unist.ac.kr

Sunghoon Lim
Ulsan National Institute of Science and Technology
Ulsan, Republic of Korea
sunghoonlim@unist.ac.kr

ABSTRACT

Generative models, particularly diffusion-based approaches, have gained significant attention in recent years 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, voltage, and current, collected from a carbon-fiber-reinforced plastic drilling process. The results demonstrate the model's ability to predict delamination from the generated machined surface images. The proposed approach has the potential to enhance process monitoring, quality control, and predictive maintenance in smart manufacturing by enabling sensor-guided visual inspection and defect detection.

CCS CONCEPTS

• **Applied computing** → Industry and manufacturing; • **Computing methodologies** → Machined surface image generation.

KEYWORDS

Image synthesis; sensor-to-image; generative model; latent diffusion model

ACM Reference Format:
Jae Gyeong Choi and Sunghoon Lim. 2018. Multimodal sensor-guided diffusion model for machined surface image synthesis. In *Proceedings of (Conference acronym 'XX).* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recent advancements in generative models, particularly diffusion-based approaches, have led to significant breakthroughs in various domains, including text-to-image generation [3, 4]. While these models have demonstrated remarkable 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 greatly benefit process monitoring, and predictive maintenance.

In this work, we propose Sensor2Image++, a framework for generating machined surface images guided by multimodal sensor inputs. Building upon the success of our previous work, Sensor2Image [2], 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.

2 METHOD

2.1 Latent diffusion model for machined surface image synthesis

The proposed approach employs a latent diffusion model architecture for generating images of machined surfaces. The model is constructed upon a variational autoencoder (VAE) framework, which comprises an encoder and a decoder. 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 $\mathcal{Z}$. 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, we employ the dnoising diffusion implicit models (DDIM) sampling process [5]. DDIM introduces a deterministic schedule for the denoising process, allowing for the generation of consistent and reproducible images. 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:

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

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.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

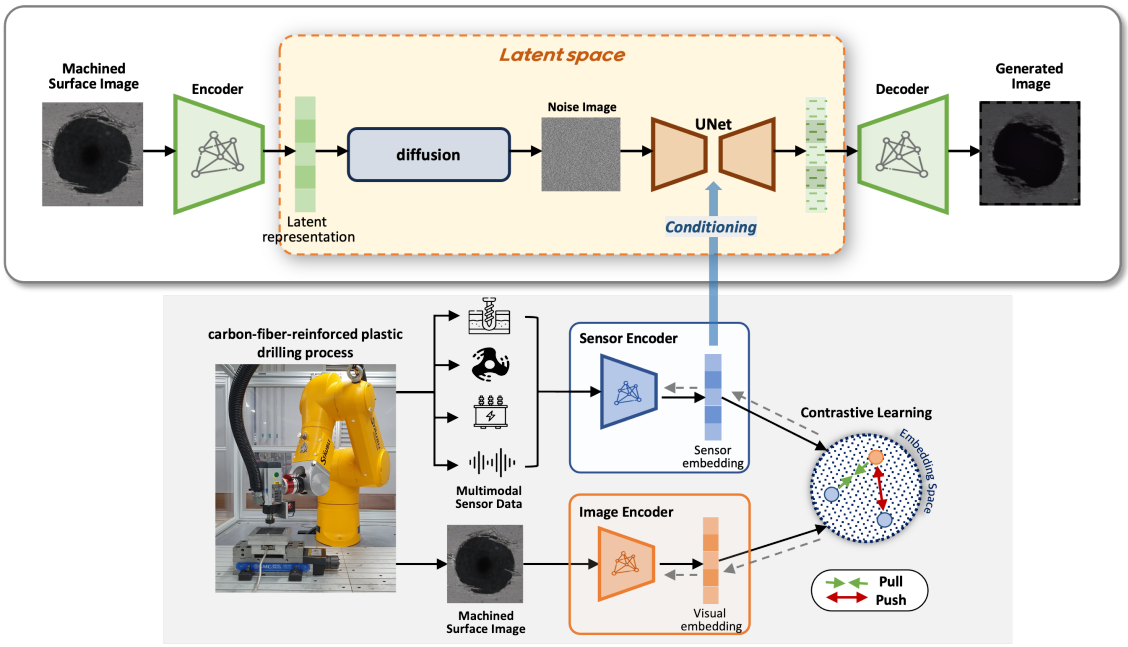

Figure 1: Sensor2Image++ for machined surface image synthesis using multimodal sensor data with varying sampling rate

2.2 Conditional multimodal sensor embedding

The conditioning of multimodal sensor embedding to a latent diffusion model involves the integration of information extracted from various sensors into the generative process, thereby influencing the generation of machined surface images. To achieve this, we introduce sensor and image encoders that extract features from sensor data and machined surface images, respectively. The sensor encoder processes time-series data from various sensors, such as force, torque, acceleration, sound, voltage, and current with different sampling rates. In contrast, the image encoder is tasked with extracting 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 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 the learning of aligned features across different modalities, resulting in a shared sensor-to-image embedding space.

In order to ensure that the extracted sensor embedding and image embedding are closely aligned for the same experimental set, we employ contrastive learning to train the sensor encoder and image encoder. The objective is to minimize the distance between the embedding 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 [1], 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.

3 CONCLUSION AND FUTURE WORK

We presented Sensor2Image++, a framework for generating machined surface images guided by multimodal sensor inputs in manufacturing. 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 experimental results on a real-world carbon-fiber-reinforced plastic drilling dataset will be used to demonstrate the superiority of Sensor2Image++ over existing methods in terms of image quality and sensor data utilization. Future work will be directed towards the extension of the framework to accommodate a wider range of sensor modalities and manufacturing processes.

REFERENCES

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. (2020). arXiv: 2002.05709 [cs.LG].

[2] Jae Gyeong Choi, Dongchan Kim, Miyoung Chung, Hyung Wook Park, and Sunghoon Lim. 2023. Sensor to machined surface image generation in cfrp drilling. In *IISE Annual Conference and Expo*. IISE.

[3] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. (2022). arXiv: 2204.06125 [cs.CV].

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. (2022). arXiv: 2112.10752 [cs.CV].

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2022. Denoising diffusion implicit models. (2022). arXiv: 2010.02502 [cs.LG].