Enhancing Fault Tolerance in Multimodal Learning: A VAE-Based Approach with Probabilistic Fusion

1st Diyar Altinses
Department of Automation Technology and learning systems
South Westphalia University of Applied Sciences
Soest, Germany
altinses.diyar@fh-swf.de

2nd Andreas Schwung
Department of Automation Technology and learning systems
South Westphalia University of Applied Sciences
Soest, Germany
schwung.andreas@fh-swf.de

***Abstract*—Multimodal learning is critical for robust perception in complex systems, yet integrating diverse modalities while ensuring fault tolerance remains a significant challenge. This paper presents a novel approach for fusing multimodal latent representations using a denoising variational autoencoder framework, where the fusion is achieved through the multiplication of probability density functions corresponding to each modality. By modeling each modality as a Gaussian distribution in the latent space, we derive a fused representation that optimally combines information from all modalities while preserving their probabilistic structure. We introduce a failure injection mechanism during training, where non-linear transformations simulate realistic faults in individual modalities. Experiments on industrial datasets demonstrate that our approach achieves superior reconstruction accuracy and robustness compared to existing methods, even in the presence of corrupted or missing modalities.**

***Index Terms*—Neural data fusion, variational autoencoder, multimodal industrial data, failure correction**

I. Introduction

Unplanned downtimes in industrial settings can lead to substantial financial losses [1]. These costs stem from labor costs, emergency repairs, lost revenue, or overhead costs. A significant contributor to such downtimes is sensor failure, which impairs the monitoring and control of critical industrial processes [2]. In particular, harsh environmental conditions, such as extreme temperatures, vibrations, or electromagnetic interference, can accelerate sensor degradation and increase failure rates [3]. As a result, developing robust fault-tolerant strategies is crucial to ensuring operational continuity and minimizing economic losses.

Traditionally, industries have implemented redundancy strategies to enhance fault tolerance, deploying multiple sensors to monitor the same parameter. However, redundancy is costly because it involves duplicating resources and growing complexity, which increases the initial investment and ongoing expenses [4]. Additionally, this approach has limitations since environmental disturbances affecting one sensor are likely to impact its redundant counterparts similarly, thereby compromising the effectiveness of redundancy [4].

An emerging and promising alternative involves the application of machine learning techniques, particularly multimodal learning. Multimodal machine learning integrates data from various sensor modalities, allowing the system to mitigate the weaknesses of one modality by leveraging the strengths of another correlated one [5]. For instance, combining vibration and acoustic data has been shown to improve the accuracy of equipment fault predictions [6]. Recent advancements have demonstrated that multimodal learning approaches not only improve fault tolerance accuracy but also enhance efficiency and cybersecurity in industrial applications [7].

In this study, we introduce a novel Multimodal Variational Denoising Autoencoder framework designed to address a wide range of realistic sensor failures. These failures encompass minor degradations, such as blurring and noise corruption, as well as complete sensor malfunctions. To effectively integrate multiple sensor modalities, we develop a latent distributional fusion strategy, which enables robust information exchange across modalities by leveraging their probabilistic representations. To evaluate the effectiveness of our approach, we conduct a comparative analysis against a standard autoencoder architecture that employs summation-based aggregation. The evaluation is performed both with and without failure injection, allowing us to assess the resilience of different architectures under varying conditions. Our key contributions can be summarized as follows:

(i) We propose a multimodal variational denoising autoencoder framework capable of handling partial and complete sensor failures through multimodal latent-space fusion.
(ii) Unlike conventional approaches that rely on simple feature concatenation or summation, our method exploits the underlying probabilistic structure of multimodal sensor data.
(iii) We systematically inject failures into multimodal sensor streams to evaluate the resilience and generalization ability of our proposed framework.
(iv) We benchmark our approach against a classical autoencoder architecture using summation-based fusion and analyze its robustness under different failure scenarios.

The organization of this paper is as follows: We begin with a review of existing research in Section 2. Following this, we outline our proposed methodology in Section 3, detailing the design of the multimodal variational denoising autoencoder and the fusion strategy employed. In Section 4, we present

our experimental results, examining the performance of our approach and assessing its reconstruction capabilities in terms of effectiveness and real-world applicability. Finally, in Section 5, we summarize our key findings and discuss potential avenues for future research.

II. Related Work

This section reviews the relevant literature on multimodal fusion for fault-tolerant systems and unimodal and multimodal variational autoencoders (VAEs). These areas form the foundation for this study, which aims to enhance the robustness of multimodal representations for failure correction by leveraging a variational autoencoder-based fusion approach.

A. Multimodal Fusion

Multimodal fusion has emerged as a pivotal area of research, aiming to integrate information from diverse data sources to enhance system performance and robustness. Recent advancements have led to a variety of methodologies and applications across multiple domains.

Deep learning techniques have significantly contributed to multimodal data fusion. Ramachandram and Taylor [8] present a survey on deep learning methods for multimodal data fusion, emphasizing the integration of heterogeneous data sources to improve model performance. Similarly, Li and Tang [9] provide a comprehensive survey of recent advancements in multimodal alignment and fusion within machine learning, highlighting the integration of diverse data types such as text, images, audio, and video [9]. A study by Feng et al. reviews multimodal sensor fusion techniques for autonomous vehicles, focusing on the integration of LiDAR, radar, and camera data to improve perception accuracy [10].

The medical field has also benefited from multimodal fusion techniques. A study by Huang et al. proposes a multimodal fusion framework combining electronic health records and medical images to improve disease diagnosis accuracy [11]. In the realm of brain imaging, the fusion of structural and functional data has been explored to better understand neural mechanisms. A study by Zhang et al. discusses various machine learning methodologies for fusing multimodal brain imaging data, highlighting applications in understanding brain arealization and disease biomarker exploration [12].

Research in fault-tolerant multimodal learning remains limited, with few studies exploring advanced model architectures for effective handling of missing data. Ma et al. introduced the SMIL framework, specifically designed to address the challenge of missing modalities across different phases, including training, testing, or both [13]. Similarly, Sohn et al. proposed a novel approach to multimodal representation learning, which mitigates the effects of absent modalities by minimizing information variation. Their method has been successfully integrated into various machine learning algorithms and deep networks with recurrent encoding mechanisms [14].

These studies collectively emphasize the importance of multimodal fusion in diverse fields, demonstrating its ability to improve system efficiency and resilience. Although researchers commonly fuse latent representations through summation or concatenation, these approaches exhibit limitations and require further refinement as they neglect the uncertain and complex interdependencies between modalities, leading to suboptimal and less robust fused representations. In this work, we propose a probabilistic fusion approach that integrates the probability density functions of individual modalities to achieve an optimal fusion of latent representations by taking interdependencies into account.

B. Multimodal variational Autoencoder

Multimodal variational autoencoders (MMVAEs) have shown promise in learning robust representations from incomplete data, but their use in fault-tolerant industrial systems remains limited. Wu and Goodman introduced the MMVAE using product-of-experts fusion to handle missing modalities, though their evaluation focused on image-text tasks, not sensor data [15]. Similarly, Marti-Juan et al. proposed a MMVAE with adversarial training for healthcare, demonstrating resilience to synthetic noise but not industrial faults [16]. In robotics, Park et al. applied MMVAEs for sensor fusion but acknowledged their fragility to real-world hardware failures [17]. Moreover, a survey by Zhan et al. noted the absence of benchmarks for fault-tolerant multimodal methods [18].

While Variational Autoencoders primarily focused on learning unimodal latent distributions, the need to model complex, multifaceted data has driven the development of multimodal VAE architectures. Recent advancements in multimodal variational autoencoders have led to the development of various architectures and methodologies to enhance multimodal fusion. Shi et al. introduced the Variational Mixture-of-Experts Autoencoder, a model that effectively learns shared and private latent spaces, enabling coherent joint and cross-generation across multiple modalities [19]. Guerrero-López et al. proposed the Multimodal Hierarchical VAE with Factor Analysis Latent Space, which employs multiple VAEs to learn private representations for each data view while sharing information through a low-dimensional latent space, enhancing flexibility and modularity [20]. Guo et al. presented DAE-Fuse, an adaptive discriminative autoencoder framework designed for multi-modality image fusion. This model generates sharp and natural fused images by incorporating discriminative blocks into the encoder-decoder architecture, demonstrating superiority in both quantitative and qualitative evaluations [21]. Vedantam et al. developed the Generative Query Network, which learns to represent scenes from multiple viewpoints, effectively capturing the underlying structure of the environment through a multimodal generative model [22].

These studies underscore the progress made in utilizing Variational Autoencoders for multimodal fusion. However, while unimodal VAEs have been widely employed for denoising tasks, the application of multimodal VAEs remains relatively underexplored, particularly for real-world industrial failure scenarios, where complex, noisy, and incomplete multimodal data are prevalent. In this work, we propose a multimodal denoising variational autoencoder that leverages the

distributional properties of different modalities, extracting and fusing them to establish a unified latent space representation. This joint distribution enables the effective reconstruction of corrupted data, enhancing fault tolerance and robustness.

III. Multimodal denoising varaitional Autoencoder

In this section, we begin by defining the problem associated with multimodal reconstruction in the presence of corrupted data. Following this, we introduce the baseline multimodal autoencoder framework as a foundation for our approach. We then extend this framework by incorporating a variational computation module and implementing distributional fusion. Lastly, we present the failure injection, designed to enhance the reliability of the reconstruction.

A. Problem defintion

Let $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a dataset of N multimodal data samples, where each sample $\mathbf{x}_i = (\mathbf{x}_i^{(0)}, \ldots, \mathbf{x}_i^{(m)}, \ldots \mathbf{x}_i^{(M)})$ consists of M modalities with $\mathbf{x}_i^{(m)} \in \mathbb{R}^{d_m}$. The data may be subject to modality-specific corruption functions $g_m : \mathbb{R}^{d_m} \to \mathbb{R}^{d_m}$, which transform clean data into corrupted versions. The goal is to find the functions $f_m : \mathbb{R}^{d_m} \to \mathbb{R}^{d_m}$ that reconstruct the original, uncorrupted data $\mathbf{x}_i$ from the corrupted inputs $\tilde{\mathbf{x}}_i$. This is achieved by minimizing a reconstruction loss

$$\mathcal{L}_{\text{recon}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{m=1}^{M} \left(\left|\left| \mathbf{x}_i^{(m)} - f_m(g_m(\mathbf{x}_i^{(m)})) \right|\right|_2^2 \right). \quad (1)$$

B. Multimodal Autoencoder Architecture

Given N different modalities, our goal is to learn a latent representation that effectively integrates information from all modalities and reconstructs the original data. Each modality i has an associated encoder E_i and decoder D_i. The latent representations produced by the encoders are aggregated via a compute function $\mathcal{C}$, which in this case is a summation function but can be replaced with other aggregation techniques. The overall architecture is presented in Figure 1.

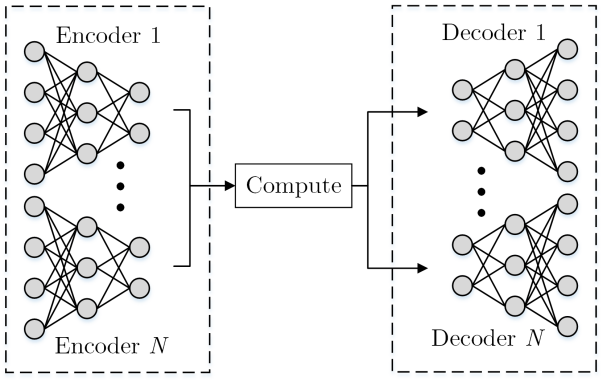


Fig. 1: Multimodal autoencoder architecture.

Each modality i is represented by its input data $\mathbf{x}_i \in \mathbb{R}^{d_i}$. The encoder function for modality i is defined as:

$$\mathbf{z}_i = E_i(\mathbf{x}_i; \theta_i^E) \in \mathbb{R}^{d_z}, \quad (2)$$

where θ_i^E represents the parameters of the encoder, and $\mathbf{z}_i$ is the latent representation of modality i. The latent representations from all modalities are aggregated using a compute function $\mathcal{C}$, formally expressed as:

$$\mathbf{z} = \mathcal{C}(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N). \quad (3)$$

We define the aggregation function as a simple summation:

$$\mathbf{z} = \sum_{i=1}^{N} \mathbf{z}_i. \quad (4)$$

However, this can be replaced with other fusion techniques, such as concatenation, mean pooling, or attention mechanisms.

The shared latent representation $\mathbf{z}$ is passed through individual decoders D_i to reconstruct the inputs for each modality:

$$\hat{\mathbf{x}}_i = D_i(\mathbf{z}; \theta_i^D), \quad (5)$$

where θ_i^D denotes the parameters of the decoder.

The optimization objective of the multimodal autoencoder is to minimize the reconstruction loss, which is commonly formulated as the sum of individual reconstruction losses across all modalities:

$$\mathcal{L}_{\text{recon}} = \sum_{i=1}^{N} \mathbb{E}_{\mathbf{x}_i \sim p(\mathbf{x}_i)} \left[||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2 \right]. \quad (6)$$

C. Variational fusion

In this section, we describe mathematically the fusion of two probability density functions (PDFs) generated by two encoders in a multimodal variational autoencoder.

The means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are computed by passing the latent samples $\mathbf{z}_1$ and $\mathbf{z}_2$ through neural networks:

$$\boldsymbol{\mu}_1 = f_{\mu_1}(\mathbf{z}_1), \quad \boldsymbol{\mu}_2 = f_{\mu_2}(\mathbf{z}_2) \quad (7)$$

The variances $\boldsymbol{\sigma}_1^2$ and $\boldsymbol{\sigma}_2^2$ are computed similarly, with an exponential and clipping operation for positivity and stability:

$$\boldsymbol{\sigma}_1^2 = \exp(f_{\sigma_1}(\mathbf{z}_1)), \quad \boldsymbol{\sigma}_2^2 = \exp(f_{\sigma_2}(\mathbf{z}_2)), \quad (8)$$

that the two encoders produce latent representations $\mathbf{z}_1$ and $\mathbf{z}_2$ are used to model Gaussian distributions:

$$p_1(\mathbf{z}_1) = \mathcal{N}(\mathbf{z}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \quad (9)$$

$$p_2(\mathbf{z}_2) = \mathcal{N}(\mathbf{z}_2; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \quad (10)$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors, and $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are the covariance matrices (assumed diagonal for simplicity, so $\boldsymbol{\Sigma}_1 = \text{diag}(\boldsymbol{\sigma}_1^2)$ and $\boldsymbol{\Sigma}_2 = \text{diag}(\boldsymbol{\sigma}_2^2)$).

The fusion of the two PDFs is achieved by:

$$p_{\text{fused}}(\mathbf{z}) \propto p_1(\mathbf{z}) \cdot p_2(\mathbf{z}) \quad (11)$$

For Gaussian distributions, the product of two Gaussians is also a Gaussian. The resulting fused distribution $p_{\text{fused}}(\mathbf{z})$ is given by:

$$p_{\text{fused}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_{\text{fused}}, \boldsymbol{\Sigma}_{\text{fused}}) \tag{12}$$

where the fused mean $\boldsymbol{\mu}_{\text{fused}}$ and covariance $\boldsymbol{\Sigma}_{\text{fused}}$ are computed as:

$$\boldsymbol{\Sigma}_{\text{fused}} = \left(\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}\right)^{-1} \tag{13}$$

$$\boldsymbol{\mu}_{\text{fused}} = \boldsymbol{\Sigma}_{\text{fused}} \left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right) \tag{14}$$

The fused distribution is then:

$$p_{\text{fused}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_{\text{fused}}, \boldsymbol{\sigma}^2_{\text{fused}}) \tag{15}$$

We incorporate a KL regularization term to ensure that both distributions operate within a comparable space, maintaining similar scales, proximity to the origin, and avoiding excessive dispersion. This is crucial for product-based fusion, as an overly dominant distribution (e.g., one with excessively large variance or a distant mean) could skew the fusion outcome. By applying KL regularization to both posteriors, we promote well-behaved, properly calibrated distributions, leading to more stable and reliable fusion. The KL divergences for the individual distributions are computed and averaged:

$$\text{KL}_1 = \text{KL}\left(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2) \| \mathcal{N}(\mathbf{0}, \mathbf{I})\right) \tag{16}$$

$$\text{KL}_2 = \text{KL}\left(\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2) \| \mathcal{N}(\mathbf{0}, \mathbf{I})\right) \tag{17}$$

$$\mathcal{L}_{\text{KL}} = \frac{\text{KL}_1 + \text{KL}_2}{2} \tag{18}$$

Thus, the total loss function of the model is given by:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda \mathcal{L}_{\text{KL}} \tag{19}$$

where λ is a hyperparameter balancing the reconstruction and regularization terms.

D. Failure injection

To convert a multimodal autoencoder into a denoising multimodal autoencoder, we introduce a noise-corruption process to the input data of each modality before encoding. The goal is to learn a robust latent representation that can reconstruct the original, uncorrupted data from its noisy version.

Let $\mathbf{x}_1$ and $\mathbf{x}_2$ represent the clean input data from two modalities. During training, we introduce a corruption process that selectively adds noise to one modality at a time or leaves both modalities uncorrupted. This is achieved using a corruption mask $\mathbf{m}$, which determines which modality is corrupted. The mask $\mathbf{m}$ is a binary vector defined as:

$$\mathbf{m} = (m_1, m_2), \quad m_i \in \{0, 1\} \tag{20}$$

where $m_1 = 1$ indicates that modality $\mathbf{x}_1$ is corrupted, $m_2 = 1$ indicates that modality $\mathbf{x}_2$ is corrupted, $m_1 = m_2 = 0$ indicates no corruption. The constraint $m_1 + m_2 \leq 1$ ensures that only one modality is corrupted at a time, or none.

The noisy inputs $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ are generated as:

$$\tilde{\mathbf{x}}_1 = m_1 \cdot g_1(\mathbf{x}_1) + (1 - m_1) \cdot \mathbf{x}_1, \tag{21}$$

$$\tilde{\mathbf{x}}_2 = m_2 \cdot g_2(\mathbf{x}_2) + (1 - m_2) \cdot \mathbf{x}_2, \tag{22}$$

where $g(\cdot)$ are failure functions that apply modality-specific transformations to the input data. These functions are based on the work of [23] and represent realistic failures, such as sensor distortions, occlusions, or other domain-specific corruptions.

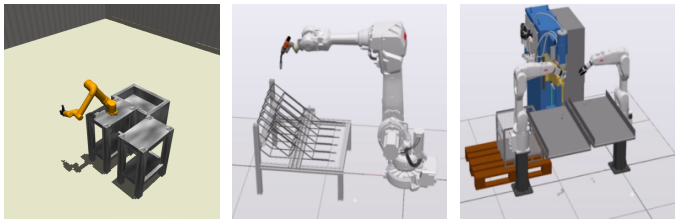
The goal of the denoising multimodal architecture is to reconstruct the clean data $\mathbf{x}_1$ and $\mathbf{x}_2$ from the noisy inputs $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$. The model is trained to minimize the reconstruction error between the reconstructed outputs $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ and the original clean data $\mathbf{x}_1$ and $\mathbf{x}_2$.

IV. Evaluation

In this section, we will evaluate the multimodal variational denoising autoencoder. To evaluate the neural denoising fusion architecture, we first present our dataset and experimental setup for the optimization process. After, we analyze the training and testing performance of the baseline as well as our proposed approach with and without failures.

A. Dataset

We employ three datasets, focusing on industrial robotic systems available at [24]–[26]. The three datasets have different levels of complexity and are presented in Figure 2. All datasets consist of N multimodal data pairs with M modalities. Each modality $\mathcal{M}_i$, where $i \in [M]$, is represented by a vector $\mathbf{x}_k^{(i)} \in \mathbb{R}^{d_i}$. The multimodal data pairs are denoted as $\{\mathbf{x}_k^{(1)}, \ldots, \mathbf{x}_k^{(M)}\}_{k=1}^N$. The joint distribution of the modalities is described by $\mathcal{P}_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(M)}}$, representing the probability of co-occurrence across modalities. Each modality also has a marginal distribution $\mathcal{P}_{\mathbf{x}^{(i)}}$, describing its independent probability distribution.



(a) MuJoCo (b) ABB single robot (c) ABB dual robot

Fig. 2: Three image modality samples of the three distinct multimodal datasets.

This study focuses on reconstructing bimodal information from static images S and sensory data T, represented as $m \in \{s, t\}$. To capture temporal dependencies, the sensory data T is segmented into overlapping windows of length $w_l = 20$ and step size $w_s = 1$. This reduces the dataset size to $N = \frac{N - w_l}{w_s}$ excluding initial samples up to the sequence length. The temporal modality T is structured as a matrix $T \in \mathbb{R}^{w_l \times d_t}$, where d_t is the dimensionality of the temporal data. The spatial modality S is represented as a higher-dimensional matrix $S \in \mathbb{R}^{3 \times s_r \times s_c}$ with $s_r = s_c = 256$ defining the spatial resolution.

B. Experimental Setup

The multimodal fusion architecture consists of two autoencoders, each dedicated to a specific modality (spatial or temporal), and a fusion module responsible for combining their latent representations. The image encoder is configured with a sequence of 2D convolutional layers following the channel structure 3-256-128-64-32-32-32. Each layer uses a kernel size of 5, a stride of 2, padding of 1, and incorporates a bias term. ReLU activation functions are applied after each layer to introduce non-linearity, and the output is flattened for compatibility with other components. The temporal encoder, designed for sequential data, employs fully connected layers with the structure n_0-256-256-256-256-256-288, where n_0 represents the input size of the sensor modality. ReLU activations are also applied after each layer in this encoder.

To simulate realistic failure scenarios, modality-specific failures are injected based on [23] during training. This is achieved by applying transformations $g_1(\cdot)$ and $g_2(\cdot)$ to the inputs of the spatial and temporal modalities, respectively. A binary corruption mask ensures that only one modality is corrupted at a time, or none, reflecting real-world conditions where failures often affect individual modalities independently.

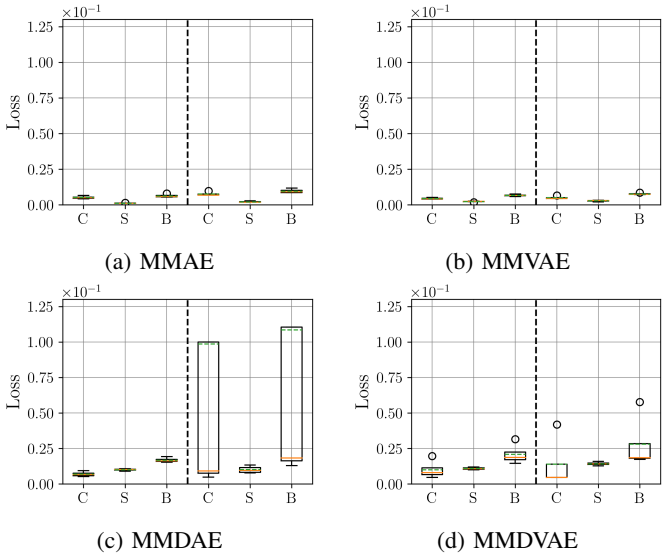
The training procedure spans a total of 200 epochs, employing a batch size of 16. The Adam optimizer is used for efficient parameter updates, with an initial learning rate of $\eta = 10^{-3}$. To enhance training stability and convergence, the learning rate is reduced by a factor of 0.1 every 100 epochs. The mean squared error (MSE) loss is utilized to evaluate the reconstruction quality.

C. Results

We begin our experiments using the Mujoco dataset by first analyzing the performance over 10 trials in Figure 3. Therefore, we evaluate the four multimodal models: the Multimodal Autoencoder (MMAE), the Multimodal Variational Autoencoder (MMVAE), the Multimodal Denoising Autoencoder (MMDAE), and the Multimodal Denoising Variational Autoencoder (MMDVAE). Both denoising models include corrupted input data. For each model, we track the reconstruction loss to assess the efficiency and robustness of the learned representations. Specifically, we present boxplots that illustrate the distribution of performance metrics, taking into account the mean performance over the last 250 training batches.

The consistently narrower boxplots for Variational Autoencoders compared to standard Autoencoders indicate lower variance in their performance across trials. This empirically shows that VAEs produce more stable and consistent results, regardless of whether the input data contains missing or corrupted information. In contrast, the classic Multimodal Autoencoder exhibits higher variability, implying that its performance is more sensitive to input variations.

The significant gap in mean performance highlights this inconsistency: while some trials may yield strong results for the standard Autoencoder, others show significant degradation, likely due to overfitting or poor generalization, as presented in Table I. This variability causes a higher spread in results,



(a) MMAE (b) MMVAE

(c) MMDAE (d) MMDVAE

Fig. 3: Boxplots of the four models using the mean test performances over the final 250 training batches of 10 trials. The performance of camera (C), sensor (S), and both combined (B) within the MuJoCo dataset are shown independently.

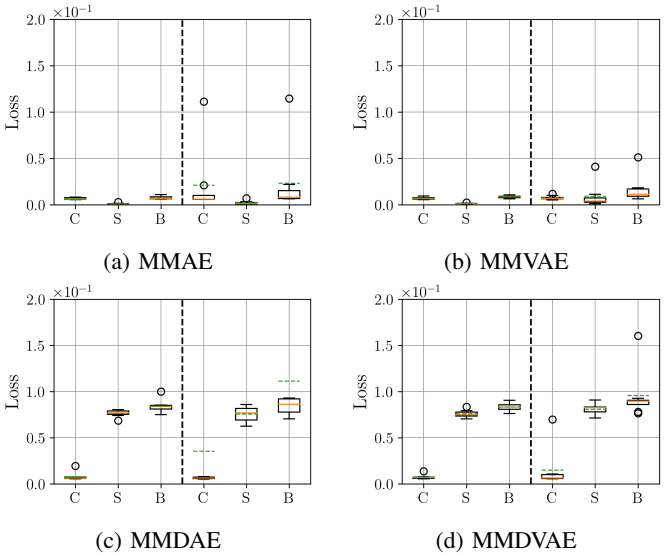
pulling the mean away from the median. However, since the median performance remains similar between models, this suggests that in typical cases, both approaches achieve comparable results. The key difference lies in the reliability of the Variational Autoencoders, which maintain more robust performance across different runs and input conditions.

TABLE I: Train (upper) and test (lower) performances (in 10^{-2}) using the mean over the final 250 batches of 10 trials. The performance of camera (C), sensor (S), and both combined (B) within the MuJoCo dataset are shown independently.

		MMAE	MMVAE	MMDAE	MMDVAE
Training	C	0.51 ± 0.12	0.44 ± 0.20	0.68 ± 1.36	1.00 ± 3.94
	S	0.11 ± 0.04	0.22 ± 0.06	0.99 ± 0.97	1.08 ± 0.95
	B	0.62 ± 0.13	0.66 ± 0.21	1.67 ± 1.66	2.08 ± 4.04
Testing	C	0.75 ± 0.13	0.49 ± 0.10	9.85 ± 15.78	1.38 ± 1.63
	S	0.20 ± 0.04	0.27 ± 0.05	1.00 ± 0.23	1.42 ± 0.12
	B	0.95 ± 0.14	**0.76 ± 0.05**	10.85 ± 15.98	**2.80 ± 1.73**

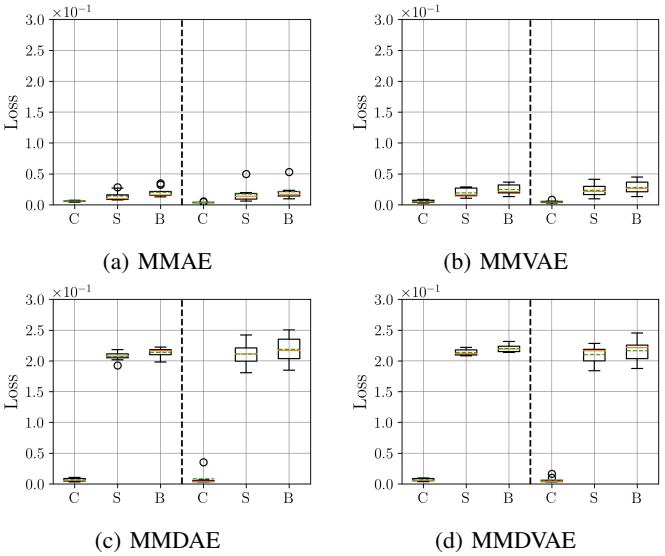
The next set of experiments is conducted using the ABB single-robot welding station dataset. We begin our analysis with a boxplot evaluation to provide a comparison of the models' performance distributions. The results are presented in Figure 4, highlighting key differences in model behavior.

Since this dataset exhibits increased complexity compared to the MuJoCo dataset, the performance results are less clear-cut and do not follow the obvious trends observed previously. However, the consistently smaller boxplots indicate a more robust convergence behavior across trials, suggesting that the models adapt more reliably. Notably, the MMVAE outperforms the Multimodal Autoencoder on average across 10 trials, even in denoising scenarios. This suggests that the probabilistic



(a) MMAE (b) MMVAE (c) MMDAE (d) MMDVAE

Fig. 4: Boxplots of the four models using the mean test performances over the final 250 training batches of 10 trials. The performance of camera (C), sensor (S), and both combined (B) within the single robot dataset are shown independently.



(a) MMAE (b) MMVAE (c) MMDAE (d) MMDVAE

Fig. 5: Boxplots of the four models using the mean test performances over the final 250 training batches of 10 trials. The performance of camera (C), sensor (S), and both combined (B) within the dual robot welding dataset are shown independently.

nature of the MMVAE helps it learn more generalizable representations, making it more resilient to variations and missing information. To provide a more detailed assessment of these results, we present the corresponding statistical data in Table II, offering deeper insights into the numerical performance differences between models.

TABLE II: Train (upper) and test (lower) performances (in 10^{-2}) using the mean over the final 250 batches of 10 trials. The performance of camera (C), sensor (S), and both combined (B) within the single robot dataset are shown independently.

		MMAE	MMVAE	MMDAE	MMDVAE
Training	C	0.65 ± 0.11	0.70 ± 0.13	0.80 ± 0.44	0.75 ± 0.24
	S	0.10 ± 0.12	0.14 ± 0.16	7.63 ± 3.71	7.58 ± 3.40
	B	0.75 ± 0.19	0.84 ± 0.20	8.43 ± 3.75	8.33 ± 3.40
Testing	C	2.10 ± 3.50	0.75 ± 0.21	3.55 ± 7.80	1.50 ± 2.11
	S	0.22 ± 0.20	0.90 ± 1.27	7.57 ± 0.84	8.08 ± 0.62
	B	2.32 ± 3.54	**1.65 ± 1.38**	11.12 ± 7.61	**9.58 ± 2.55**

Similarly, we extend our analysis to the dual-robot welding station dataset. Given the increased complexity introduced by the interaction between two robotic systems, this dataset presents additional challenges in learning robust multimodal representations. As in previous experiments, we utilize visualization techniques (Figure 5) and statistical evaluations (Table III) to compare convergence behavior, stability, and overall performance across multiple trials. By examining the models' performance under these conditions, we aim to evaluate their ability to generalize to more intricate and dynamic scenarios.

To evaluate the effectiveness of the proposed variational denoising approach, we compare the reconstruction capabilities of MMDAE and MMDVAE on input images. In Figure 6, we present four reconstructed samples demonstrating MMDVAE's superior ability to preserve fine details and structural integrity compared to MMDAE. The visual comparisons reveal that MMDVAE better handles edge sharpness, while MMDAE tends to produce specific artifacts or blurred areas. This performance gap stems from MMDVAE's generative nature, which enables more robust feature extraction from corrupted inputs compared to a point-to-point mapping.

TABLE III: Train (upper) and test (lower) performances (in 10^{-2}) using the mean over the final 250 batches of 10 trials. The performance of camera (C), sensor (S), and both combined (B) within the dual robot dataset are shown independently.

		MMAE	MMVAE	MMDAE	MMDVAE
Training	C	0.58 ± 1.69	0.53 ± 1.30	0.63 ± 2.71	0.65 ± 2.34
	S	1.41 ± 2.75	1.92 ± 2.84	20.73 ± 10.26	21.38 ± 10.07
	B	1.99 ± 3.26	2.45 ± 3.16	21.36 ± 10.57	22.03 ± 10.33
Testing	C	0.38 ± 0.14	0.46 ± 0.36	0.8 ± 2.54	0.62 ± 0.78
	S	1.72 ± 1.51	2.31 ± 2.27	21.09 ± 2.65	21.02 ± 2.69
	B	**2.1 ± 1.49**	2.77 ± 2.29	21.89 ± 3.65	**21.64 ± 2.80**

The results indicate that as dataset complexity increases, the performance of both models (classic and variational autoencoders) becomes increasingly similar, suggesting limited robustness to high-complexity data structures. However, the MMDVAE exhibits a slight but consistent advantage, likely due to its probabilistic framework better capturing data variability, though the overall benefit remains marginal.

V. Conclusion

In this paper, we introduced a novel approach for multimodal representation fusion using a multimodal denoising variational autoencoder framework, where the fusion is

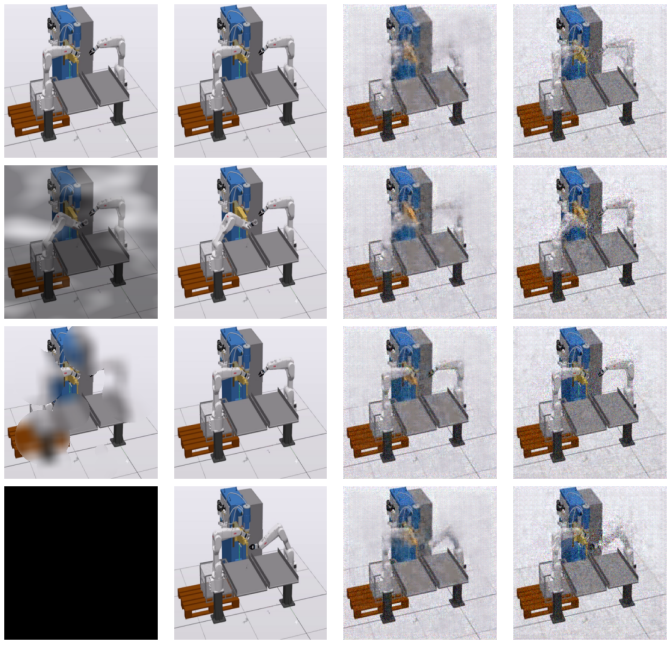


Fig. 6: Comparison of reconstruction performance on four test samples. Columns show (left to right): Input image, target output, MMDAE reconstruction, and MMDVAE reconstruction.

achieved through the multiplication of probability density functions corresponding to each modality. By modeling each modality as a Gaussian distribution, we derived a robust and expressive fused representation that optimally combines information from all modalities. To address real-world challenges, we incorporated a fault-tolerant training mechanism, which simulates modality-specific failures during training, ensuring the model's resilience to corrupted or missing data. Experimental results on industrial datasets demonstrated the effectiveness of our approach, achieving superior reconstruction accuracy and robustness compared to existing methods.

While this work demonstrates advancements in multimodal fusion and fault tolerance, several directions remain for future exploration. The current framework focuses on bimodal data. Extending it to handle more than two modalities could further enhance its applicability in complex systems. Additionally, integrating marginal information into the fusion process could further improve this multimodal denoising approach.

References

[1] R. Wolniak, "Downtime in the automotive industry production process–cause analysis," *Quality Innovation Prosperity*, vol. 23, no. 2, pp. 101–118, 2019.

[2] D. O. S. Torres, D. Altinses, and A. Schwung, "Data imputation techniques using the bag of functions: Addressing variable input lengths and missing data in time series decomposition," in *2025 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, 2025, pp. 1–7.

[3] N. Lakal, A. H. Shehri, K. W. Brashler, S. P. Wankhede, J. Morse, and X. Du, "Sensing technologies for condition monitoring of oil pump in harsh environment," *Sensors and Actuators A: Physical*, vol. 346, p. 113864, 2022.

[4] D. Altinses and A. Schwung, "Deep multimodal fusion with corrupted spatio-temporal data using fuzzy regularization," in *IECON 2023-49th Annual Conference of the IEEE IES*. IEEE, 2023, pp. 1–7.

[5] D. Altinses, D. O. S. Torres, S. Lier, and A. Schwung, "Neural data fusion enhanced pd control for precision drone landing in synthetic environments," in *2025 IEEE International Conference on Mechatronics (ICM)*. IEEE, 2025, pp. 1–7.

[6] O. Kullu and E. Cinar, "A deep-learning-based multi-modal sensor fusion approach for detection of equipment faults," *Machines*, vol. 10, no. 11, p. 1105, 2022.

[7] K. M. Alsaif, A. A. Albeshri, M. A. Khemakhem, and F. E. Eassa, "Multimodal large language model-based fault detection and diagnosis in context of industry 4.0," *Electronics*, vol. 13, no. 24, p. 4912, 2024.

[8] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[9] S. Li and H. Tang, "Multimodal alignment and fusion: A survey," *arXiv preprint arXiv:2411.17040*, 2024.

[10] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[11] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *NPJ digital medicine*, vol. 3, no. 1, p. 136, 2020.

[12] N. Luo, W. Shi, Z. Yang, M. Song, and T. Jiang, "Multimodal fusion of brain imaging data: Methods and applications," *Machine Intelligence Research*, vol. 21, no. 1, pp. 136–152, 2024.

[13] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2302–2310.

[14] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information," *Advances in neural information processing systems*, vol. 27, 2014.

[15] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," *Advances in neural information processing systems*, vol. 31, 2018.

[16] G. Martí-Juan, M. Lorenzi, G. Piella, A. D. N. Initiative *et al.*, "Mc-rvae: Multi-channel recurrent variational autoencoder for multimodal alzheimer's disease progression modelling," *NeuroImage*, vol. 268, p. 119892, 2023.

[17] D. Park, Y. Hoshi, and C. C. Kemp, "A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1544–1551, 2018.

[18] Y. Zhan, R. Yang, J. You, M. Huang, W. Liu, and X. Liu, "A systematic literature review on incomplete multimodal learning: techniques and challenges," *Systems Science & Control Engineering*, vol. 13, no. 1, p. 2467083, 2025.

[19] Y. Shi, B. Paige, P. Torr *et al.*, "Variational mixture-of-experts autoencoders for multi-modal deep generative models," *Advances in neural information processing systems*, vol. 32, 2019.

[20] A. Guerrero-López, C. Sevilla-Salcedo, V. Gómez-Verdejo, and P. M. Olmos, "Multimodal hierarchical variational autoencoders with factor analysis latent space," *arXiv preprint arXiv:2207.09185*, 2022.

[21] Y. Guo, R. Xu, R. Li, Z. Wu, and W. Su, "Dae-fuse: An adaptive discriminative autoencoder for multi-modality image fusion," *arXiv preprint arXiv:2409.10080*, 2024.

[22] R. Vedantam, K. Desai, S. Lee, M. Rohrbach, D. Batra, and D. Parikh, "Probabilistic neural symbolic models for interpretable visual question answering," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6428–6437.

[23] D. Altinses and A. Schwung, "Multimodal synthetic dataset balancing: a framework for realistic and balanced training data generation in industrial settings," in *IECON 2023-49th*. IEEE, 2023, pp. 1–7.

[24] D. Altinses, "Synthetic multimodal dataset using mujoco: Ur5 robot motion," Nov. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.14041622

[25] ——, "Synthetic multimodal dataset using abb studio: Single robot welding station," Nov. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.14041488

[26] ——, "Synthetic multimodal dataset using abb studio: Dual robot welding station," Nov. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.14041416