Enhancing Sensor Robustness in Automotive Systems: A Multimodal Generative Approach

Mustapha Bounoua[12], Christophe Beaugeant[1], Giulio Franzese[2], Pietro Michiardi[2]

1: Ampere, Sophia-Antipolis, France
2: Eurecom, Sophia-Antipolis, France

Abstract:

Modern automotive systems increasingly rely on a wide array of sensors to deliver safe and efficient operation. However, sensor data quality can be compromised due to environmental conditions, sensor malfunctions, or occlusions, which can jeopardize vehicle performance. This paper proposes a multimodal generative approach to enhance sensor robustness by leveraging multimodal latent diffusion models. These models facilitate modality cross-generation and enhancement, enabling the reconstruction of impaired sensor data through information from other modalities. We introduce a novel application in the automotive context, focusing on improving night vision capabilities using data from Lidar and radar sensors to generate enhanced camera images. This approach ensures robust sensor functionality, offering a general solution for sensor data integrity in automotive systems.

Keywords: Multimodal Generative Models, Sensor Robustness, Automotive Systems, Latent Diffusion, Night Vision

1. Introduction

Advanced driver-assistance systems (ADAS) rely on sophisticated sensor integration. Cameras, radar, and Lidar are essential components, providing comprehensive data for navigation, obstacle detection, and decision-making processes. However, the reliability of these systems is often compromised by issues such as sensor degradation, environmental factors, and data loss, which necessitate robust solutions for ensuring data integrity.

In this paper, we present a multimodal generative approach using latent diffusion models to address these challenges. By leveraging data from multiple sensor modalities, we aim to reconstruct missing or impaired sensor data, thereby enhancing the overall robustness of automotive systems. Our focus is on modality cross-generation and enhancement, where information from one modality can be used to generate or improve another. We explore a specific application in enhancing night vision capabilities, using radar and Lidar data to augment camera imagery in low-light conditions.

2. Background and Related Work

Multimodal sensors generative modeling:

The integration of multimodal sensors in the automotive industry is pivotal to the advancement of autonomous driving systems. By combining data from multiple sensor types, such as cameras, Lidar, radar, and vehicles can achieve a comprehensive understanding of their surroundings, which is critical for safe and reliable navigation in diverse driving conditions. This multimodal approach leverages the strengths of each sensor type: cameras provide high-resolution visual data, Lidar offers precise depth information, radar excels in detecting objects in adverse weather conditions, and GPS ensures accurate localization.

The use of generative models with multimodal sensor data has demonstrated great potential in improving autonomous vehicle performance. For example, [1] employed multimodal large models for hazard detection, [2] utilized them for trajectory prediction and [3] to enhance perception. Works like [4] [5] [6] focused on sensor fusion to enhance vehicle perception. [7] and [8] introduced multimodal world models to improve prediction quality in autonomous driving. [9] proposed an end-to-end autonomous driving paradigm based on generative models.

Diffusion-based Multimodal Generative Models:

Combining diffusion models [10] [11] with multimodal learning has become an active area of research. For example, [12] demonstrated image generation from textual descriptions by leveraging a diffusion process in latent spaces, merging the power of diffusion models with cross-modal capabilities.

Further work like [13] extended this approach by performing the diffusion process in a lower-dimensional latent space, enabling faster, high-quality image and text to image generation. These models are not only efficient at generating high-resolution

data but can be used to estimate information theoretic measures [14] [15]. Other works have focused on scaling to multiple modalities, such as [16] and [17] enabling both the joint and conditional generation of all modalities.

c) Training the Diffusion Model

Diffusion models [11] have merged as powerful generative models capable of fitting complex distributions. In this work, a score-based diffusion model is trained on the joint latent representations z defined earlier. In the forward diffusion, a gaussian noise is progressively added to the data, where the model learns to denoise the data iteratively. The forward process can be defined via a stochastic differential equation (SDE) [18] [19]:

$$dZ_t = \alpha(t)Z_t + g(t)\,dW_t$$

Where $\alpha(t)Z_t$ and $g(t)$ are the drift and diffusion terms, respectively, and w_t is a Wiener process.

This yields a forward diffusion process $\{z_t\}_{t=0}^{T}$ indexed by a continuous time variable $t \in [0,T]$ with initial condition $Z_0 \sim p_z$ and prior $Z_T \sim p_T$ which is typically gaussian distribution.

The reverse diffusion process is possible via a reverse SDE [18]:

$$dZ_t = [\alpha(t)Z_t - g^2(t)\nabla \log p_t\,(Z_t)]dt + g(t)d\widetilde{W}_t$$

The score function $\nabla \log p_t\,(z)$ can be approximated using a neural network $s_\theta(z,t)$. Once the score function is accessible the simulation of the reverser process allows the sampling of new samples z_0 by flowing the arrow of time from T to 0. Learning the score function can also be done via the denoising loss [11] .

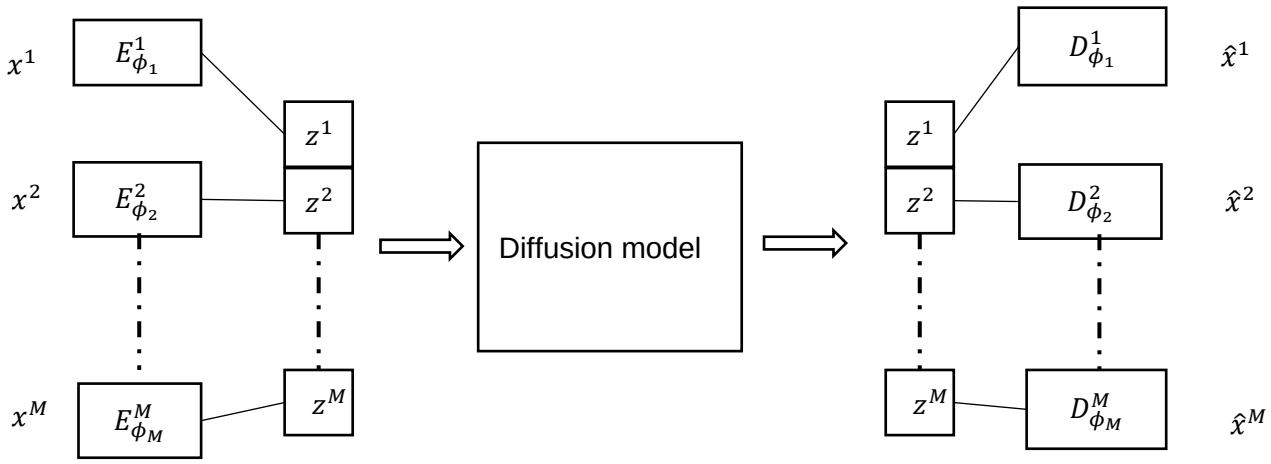


Figure 1 General architecture showing the different components of MLD.

3. Methodology

a) Multimodal Latent Diffusion

Multimodal latent diffusion (MLD) [16] is designed to learn joint latent representations from multiple sensor modalities, enabling the generation and enhancement of sensor data. The process involves a two-stage approach like [13]: The first stage consists of encoding each modality into a latent space then the training of a diffusion model on the concatenated encodings. A masking mechanism is applied to learn the different modes enabling conditional and unconditional generation.

b) Encoding Sensor Modalities

Consider $X = \{X^1, \dots, X^M\}$ a set of M data modalities collected from different sensors. Each modality is encoded using a modality-specific encoder: $z^i = \mathrm{E}^{\mathrm{i}}_{\Phi_1}(\mathrm{x}^{\mathrm{i}})$ and a decoding phase to the input space $\widehat{\mathrm{x}^{\mathrm{i}}} = D^i_{\phi_1}(x^i)$.

The encoded representations are concatenated to form a joint latent representation $Z = [Z^1, .. Z^M]$ which yields the probability distribution p_z. The training of the autoencoder can be conducted independently during a first stage. The first stage consisting of modality encoding serves two purposes: The reduction of the data dimension and the unification of the presentation of the different modalities as can be heterogenous.

d) Modality Cross-Generation

Modality cross generation is crucial to ensure the reconstruction of ensures data whenever a modality is impaired. This can be casted as the conditional generation case: given a generic partition of all modalities into non overlapping sets $A_1 \cup A_2$, where $A_2 = (\{1, \ldots, M\} \setminus A_1)$, conditional generation requires samples from the conditional distribution $p(z_{A_1}|z_{A_2})$, which are based on masked forward and backward diffusion processes.

The mask $m(A_1)$ contains M vectors u_i , one per modality, and with the corresponding cardinality. If modality $j \in A_1$, then u_j, otherwise $u_j = 0$. The effect of masking is to "freeze" throughout the diffusion processes the part of the random variable z_t corresponding to the conditioning latent modalities z^{A_2}. More formally, we define the masked forward diffusion SDE:

$$dz_t = m(A_1) \odot [\alpha(t) Z_t + g(t)\, dW_t]$$

To sample new samples in a conditional manner, we derive the masked reverse SDE:

$$dZ_t = m(A_1) \odot \Big[\big[\alpha(t) Z_t - g^2(t) \nabla \log p_t\left(Z_t | z_{A_2}\right)\big] dt + g(t) dW_t\Big]$$

e) Randomized Training

To learn all the score function we adopt a randomized approach where randomly select a set of conditioned modalities A_2 and learn the corresponding score function. We can restrict the set of learned score functions to meet the needs of specific applications, as the generation of certain modalities may be the primary goal.

f) Modality Enhancement

Modality enhancement extends the concept of modality cross generation by starting the reverse diffusion process from an intermediate latent state $z_{t'}$ rather than the terminal noise state z_T with $t' \in [0, T]$. This approach is particularly useful when aiming to refine or augment the quality of an existing modality while retaining the influence of the initial data sample. In this setting, the masked forward diffusion SDE remains the same. Consequently, the same model trained for cross generation purposes can be reused in zero shot manner.

The choice of t' determines the extent of enhancement: starting from a smaller t' (closer to 0) retains more of the original data characteristics, while a higher value t' (closer to T) allows for more significant generative modifications. Thus, modality enhancement provides a flexible mechanism for balancing between preservation of original modality details and the introduction of new features via the generative process.

4. Multimodal Integration for Night Vision

Perception systems in automotive applications rely heavily on camera data to provide visibility. However, in low-light conditions, cameras often struggle to capture high-quality images impacting object detection and navigation.

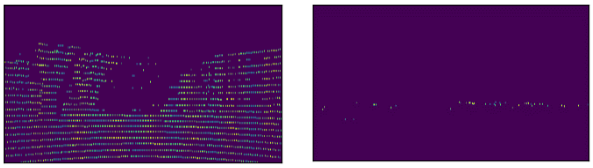
Our approach relies on MLD to integrates the information from Lidar and radar sensors to reconstruct or enhance night vision capabilities. By generating high-quality camera images using radar and Lidar information, we improve visibility and object detection accuracy in low-light environments. We consider the different sensors (Camera, Lidar, Radar) as different modalities and apply MLD to reconstruct or enhance the camera modality during nighttime.

A crucial challenge in night to day application is the absence of paired night -day dataset. To solve this problem, during training we synthetically generate an additional modality which consists of low illumination version of the daytime sample.

g) Implementation details

Our approach employs a diffusion transformer (DiT) architecture [20] , leveraging the power of transformer networks to manage multiple modalities and capture long-range dependencies. The model architecture is designed to process at the same time multiple modalities, utilizing token-based representations. We use the autoencoder from [13] to encode the different modalities.

Night-time camera (Synthetic) **Day-time Camera**



Lidar **Radar**

Figure 2: Illustrative examples from the nuScenes dataset used for training

The model is initialized with weights pre-trained on the ImageNet dataset [21]. We extend the model by adding additional tokens for each modality and fine-tune it using multimodal sensor data. The fine-tuning process is conducted on 512x512 image resolution for 200,000 iterations, with a batch size of 24 and a learning rate of 1×10−4. We employ the randomized approach described in Section 2 to learn the reconstruction of the camera modalities given the availability of either Lidar, radar, or both. In all the experiments, we use classifier free guidance [22] with **cfg = 2.0** and DDIM [23] sampler with **50** steps.

h) Data set and Experimental Setup

The nuScenes [24] dataset is a comprehensive autonomous driving dataset that includes sensor data from cameras, Lidar, radar, and more, collected from urban environments in various weather and lighting conditions.

Condition	**Day**	**Night**
Train	114251	16620
Test	3840	524

Table 1: nuScenes Dataset sample size after preprocessing.

In our approach, we train models using the front camera frames sampled with **12Hz**. Radar, and Lidar data are projected onto the same intrinsic parameters as the front camera, ensuring a unified perspective across modalities. We exclusively use the daytime data from the nuScenes dataset for training. In **Table 1**, we provide the dataset size information after preprocessing. To simulate nighttime conditions, we apply a low-illumination algorithm, like [3]. During testing, we reverse this process by reconstructing daytime modalities from the simulated nighttime data. We evaluate the quality of this reconstruction using metrics like Peak Signal-to-Noise Ratio (PSNR) [25], Structural Similarity Index Measure (SSIM) [26], and Learned Perceptual Image Patch Similarity (LPIPS) [27].

i) Experimental results

Quantitative results:

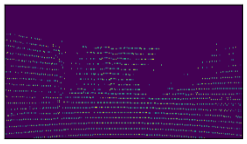
Table 2 presents the performance results of MLD light improvement across different modes. In our experiments, we reconstruct the daytime camera modality using nighttime camera data, testing various scenarios that incorporate additional information from radar and Lidar, Lidar only, radar only, and no additional sensors. We observe that Lidar contributes the most to improving camera reconstruction, though radar also yields comparable results. The performance is lowest when no additional sensors are used, indicating the importance of supplementary sensor data in enhancing reconstruction quality.

MLD	PSNR (↑)	SSIM (↑)	LPIPS (↓)
All sensors	**26.965**	**0.705**	**0.201**
Lidar	<u>26.751</u>	<u>0.698</u>	<u>0.207</u>
Radar	26.589	0.694	0.211
Only camera	26.565	0.680	0.221

Table 2: Quantitative results of MLD conditional generation using nighttime modality with different conditions: Radar and Lidar, Lidar only, Radar only, and no additional sensor. Bold indicates the best performance, and inline the second-best.

Night Image

Generated Camera



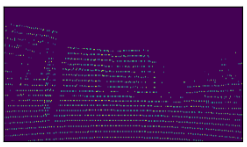
Lidar

Ground Truth

Radar

Night Image

Generated image

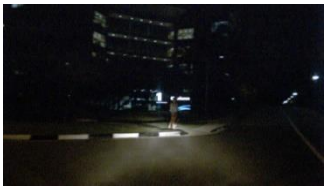


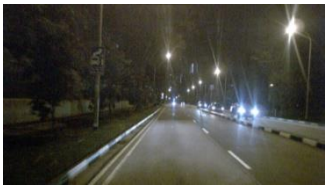
Lidar

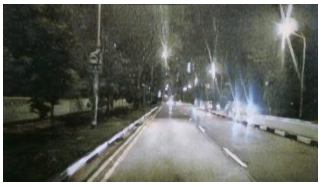
Ground Truth

Radar

Figure 3: Quantitative results of MLD conditional generation using nighttime synthetic modality to reconstruct the day-time camera modality.







Night Camera **Improved night camera**

Figure 4: Quantitative results of MLD conditional generation using real nighttime modality and the different sensors to improve night perception.

Qualitative results:

Our approach demonstrated improved visibility and object detection accuracy compared to baseline methods. **Figure 3** presents MLD conditional generation capabilities using the synthetic nighttime

image as condition. MLD is nearly able to perfectly reconstruct the image using by leveraging the additional information form the other modalities. **Figure 4** illustrates MLD performance on real nighttime images, displaying clearer and more detailed nighttime image in challenging illumination conditions.

Discussion and Future Work

The integration of multimodal data significantly enhances night vision performance. The generative model effectively fills in missing details and reduces noise, making the system more reliable in low-light conditions.

While our approach shows promise, there are limitations to consider. The diffusion-based approach ensures generation capabilities via an iterative process which raises computation and latency issues. A potential solution would consist of refining this method for better real-time performance via model distillation approach like Consistency models [28].

5. Conclusion

Our proposed multimodal generative approach provides a robust solution to the challenges faced by automotive sensor systems. By leveraging the strengths of different sensor modalities, we can generate and enhance impaired sensors, improving the overall reliability and performance of automotive systems. Multimodal generative models offer several advantages over traditional sensor fusion techniques. Our approach demonstrates the potential of these models in addressing real-world challenges in automotive applications.

References

[1] M. a. A. H. I. E. M. a. G. S. a. R. A. Abu Tami, "Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events," *Vehicles,* 2024.

[2] B. a. L. K. a. S. E. a. P. M. Ivanovic, "Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach," *IEEE Robotics and Automation Letters,* 2020.

[3] J. Li, B. Li, Z. Tu, X. a. G. Q. Liu, F. Juefei-Xu, R. Xu and H. Yu, "Light the Night: A Multi-Condition Diffusion Framework for Unpaired Low-Light Enhancement in Autonomous Driving," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2024.

[4] M. a. A. A. a. F. M. a. M. P. Da Silva--Filarder, "Multimodal variational autoencoders for sensor fusion and cross generation," *20th IEEE International Conference on Machine Learning and Applications (ICMLA),* 2021.

[5] D. a. L. Y. a. J. T. a. T. P. a. C. K. a. I. S. Roy, "Multi-modality sensing and data fusion for multi-vehicle detection," *IEEE Transactions on Multimedia,* 2022.

[6] Y. a. C. Y. Huang, "Autonomous Driving with Deep Learning: {A} Survey of State-of-Art," *arXiv,* 2020.

[7] D. a. Y. Y. a. J. T. a. Z. J. M. Bogdoll, "MUVO: A Multimodal World Model with Spatial Representations for Autonomous Driving," *arXiv,* 2024.

[8] A. a. R. L. a. Y. H. a. M. Z. a. F. G. a. K. A. a. S. J. a. C. G. Hu, "Gaia-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080,* 2023.

[9] W. a. S. R. a. G. X. a. Z. C. a. C. L. Zheng, "GenAD: Generative End-to-End Autonomous Driving," *arXiv,* 2024.

[10] J. a. W. E. a. M. N. a. G. S. Sohl-Dickstein, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," *International Conference on Machine Learning (ICML),* 2015.

[11] J. H. a. A. J. a. P. Abbeel, "Denoising Diffusion Probabilistic Models," *Advances in Neural Information Processing Systems 33 (NeurIPS 2020),* 2020.

[12] A. a. P. M. a. G. G. Ramesh, "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint arXiv:2204.06125,* 2022.

[13] R. a. B. A. a. L. D. a. E. P. a. O. B. Rombach, "High-Resolution Image Synthesis with Latent Diffusion Models," *arXiv,* 2021.

[14] G. a. B. M. a. M. P. Franzese, "MINDE: Mutual Information Neural Diffusion Estimation,," *The Twelfth International Conference on Learning Representations,* 2024.

[15] M. a. F. G. a. M. P. BOUNOUA, "SOI: Score-based O-{INFORMATION} Estimation," *ICML,* 2024.

[16] M. a. F. G. a. M. P. Bounoua, "Multi-Modal Latent Diffusion," *Entropy,* 2024.

[17] F. a. N. S. a. X. K. a. L. C. a. P. S. a. W. Y. a. Y. G. a. C. Y. a. S. H. a. Z. J. Bao, "One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale," *ICML,* 2023.

[18] Y. S. a. J. S.-D. a. D. P. K. a. A. K. a. S. E. a. B. Poole, "Score-Based Generative Modeling through Stochastic Differential Equations," *International Conference on Learning Representations,* 2021.

[19] G. a. R. S. a. Y. L. a. F. A. a. R. D. a. M. F. a. P. M. Franzese, "How Much Is Enough? A Study on Diffusion Times in Score-Based Generative Models," *Entropy,* 2023.

[20] W. Peebles and S. Xie, "Scalable Diffusion Models with Transformers," *arXiv preprint arXiv:2212.09748,* 2022.

[21] J. a. D. W. a. S. R. a. L. L.-J. a. L. K. a. F.-F. L. Deng, "ImageNet: A large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition,* 2009.

[22] J. a. S. T. Ho, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598,* 2022.

[23] J. S. a. C. M. a. S. Ermon, "Denoising Diffusion Implicit Models," *arXiv,* 2022.

[24] H. C. a. V. B. a. A. H. L. a. S. V. a. V. E. L. a. Q. X. a. A. K. a. Y. P. a. G. B. a. O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," *CVPR,* 2020.

[25] R. C. a. W. R. E. Gonzalez, Digital Image Processing, Pearson Education, 2008.

[26] Z. a. B. A. C. a. S. H. R. a. S. E. P. Wang, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing,* 2004.

[27] R. a. I. P. a. E. A. A. a. S. E. a. W. O. Zhang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric,* p. 2018, CVPR.

[28] Y. a. D. P. a. C. M. a. S. I. Song, "Consistency Models," *PMLR,* 2023.

[29] Z. a. L. C. a. X. Y. a. W. J. Huang, "Multi-Modal Sensor Fusion-Based Deep Neural Network for End-to-End Autonomous Driving With Scene Understanding," p. 2020, IEEE Sensors Journal.