2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Mutual Information-calibrated Conformal Feature Fusion for Uncertainty-Aware Multimodal 3D Object Detection at the Edge

Alex C. Stutts, Danilo Erricolo, Sathya Ravi, Theja Tulabandhula, and Amit Ranjan Trivedi

***Abstract*— In the expanding landscape of AI-enabled robotics, robust quantification of predictive uncertainties is of great importance. Three-dimensional (3D) object detection, a critical robotics operation, has seen significant advancements; however, the majority of current works focus only on accuracy and ignore uncertainty quantification. Addressing this gap, our novel study integrates the principles of conformal inference (CI) with information theoretic measures to perform lightweight, Monte Carlo-free uncertainty estimation within a multimodal framework. Through a multivariate Gaussian product of the latent variables in a Variational Autoencoder (VAE), features from RGB camera and LiDAR sensor data are fused to improve the prediction accuracy. Normalized mutual information (NMI) is leveraged as a modulator for calibrating uncertainty bounds derived from CI based on a weighted loss function. Our simulation results show an inverse correlation between inherent predictive uncertainty and NMI throughout the model's training. The framework demonstrates comparable or better performance in KITTI 3D object detection benchmarks to similar methods that are not uncertainty-aware, making it suitable for real-time edge robotics.**

I. Introduction

The rapid development of artificial intelligence (AI) capabilities, as demonstrated with image recognition and large language models (LLMs), has enabled its adoption across various domains. However, concerns about its reliability persist for safety-critical applications, including robotics. Given that the accuracy of data-driven models cannot be assured, it becomes essential not only to question *what if the model is wrong?*, but also to determine *how wrong* it might be by assessing its predictive uncertainties. Quantifying uncertainty in deep learning has, therefore, gained traction. Notably, data-driven models can suffer from two main types of uncertainties: *epistemic* and *aleatoric* [1]. Epistemic uncertainty arises from inherent data variance and can often be mitigated with additional training data. Conversely, aleatoric uncertainty stems from random data distortions, such as blurriness, occlusions, and overexposure in images, and cannot be resolved merely by augmenting the training data.

However, much of the current work has neglected considering platforms with time, cost, area, computing, and power constraints. Consequently, those existing uncertainty estimation methods, often reliant on distribution-based approximations, struggle under edge deployment due to their need for iterative sampling. Therefore, uncovering true, statistically confident uncertainties in point (mean) predictions that are intuitive and visualizable under considerable resource constraints remains challenging for critical edge robotics.

Acknowledgement: This work was supported in part by COGNISENSE, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. Authors are with the University of Illinois Chicago (UIC), Chicago, IL, Email: astutt2@uic.edu

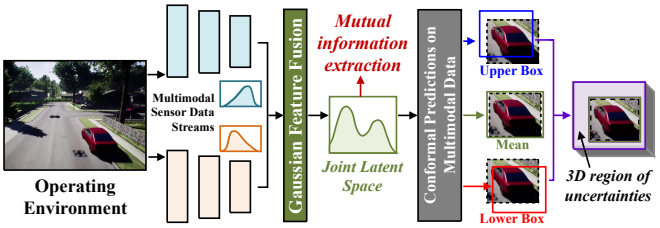

Fig. 1: **Uncertainty-Aware Multimodal Inference at the Edge:** In this work, we present a generalizable, multimodal conformal inference framework for lightweight uncertainty awareness and apply it to 3D object detection. The proposed methodology is deeply rooted in information and statistical theory, allowing the framework to take full advantage of the benefits of conformal prediction in quantifying uncertainty while under considerable resource constraints.

To tackle these challenges, we explore conformal inference (CI) [2]–[4]. Rooted in information theory and probabilistic prediction, CI has emerged as a prominent uncertainty quantification method that is simple, generalizable, and scalable [5]. Unlike conventional statistical inference, which depends on intimate knowledge of the data distribution for uncertainty estimation and is vulnerable to modeling inaccuracies, CI produces reliable, uncertainty-aware prediction intervals without distributional assumptions given a finite set of training data. CI assesses the conformity of each incoming data point to the existing dataset and formulates uncertainty intervals based on a preset coverage rate. Importantly, CI is compatible with any core model with an inherent uncertainty notion, yielding both model-agnostic and statistically sound estimations.

Despite these advantages, a key limitation of CI is its tendency to provide overly cautious uncertainty estimates that may prevent a prediction model from making meaningful decisions. For example, overly conservative uncertainty estimates in autonomous navigation can lead to suboptimal path planning, such as taking longer routes than necessary. While multimodal sensors have become prevalent in various robotics tasks to enhance the robot's perception and decision-making capabilities, they present a unique opportunity for CI to optimally calibrate the predicted uncertainty estimates by exploiting mutual information (MI) of multimodal sensor data streams. MI is an information-theoretic metric that measures the dependence between the marginal distributions of two random variables through their joint distribution. In this case, it can measure how much one sensor modality explains the output and prediction from another while operating in the same environment. Thus, leveraging MI to calibrate and

tighten CI's predictive uncertainty bounds while maintaining the guaranteed coverage rate is attractive.

Towards this goal, we consider 3D object detection a driving application and present a systematic framework for including MI in optimizing CI-based uncertainty bounds. 3D object detection is essential for many autonomous systems to provide a semantic understanding of their environment through identifying, localizing, and categorizing various objects. However, various propositions in our work are also generalizable to other autonomy tasks. The framework is overviewed in 1.

Our work makes the following key contributions:

- We introduce a 3D object detection framework that integrates uncertainty-aware projections obtained through conformal prediction. Evaluated on the demanding 3D KITTI vision benchmark suite [6], this framework surpasses state-of-the-art models in inference runtime while achieving a competitive accuracy. Given these attributes, our approach is particularly suited for edge robotics platforms with limited time and computational resources.
- We introduce a multitask loss function that can train a model to simultaneously provide point predictions and adaptive uncertainty confidence bounds that each take the form of 3D bounding boxes. The uncertainty boxes are demonstrated to enhance average precision and are combined to be more visually intuitive. Furthermore, we weight the loss function with an uncertainty-based distance metric, averaged over every dimension of each output, to influence the model to prioritize training samples that introduce more uncertainty.
- Integrating conformal inference with information-theoretic measures, specifically MI, we discuss a method to fuse data from multimodal sensors using a multivariate Gaussian product of latent variables in a variational autoencoder (VAE). The proposed VAE-based multimodal data fusion captures salient features of each modality and enables us to compute normalized mutual information (NMI). This, in turn, allows us to optimally calibrate the uncertainty bounds in a sample-adaptive manner.

In Sec. II, we discuss the current art of 3D object detection. In Sec. III, we present the proposed framework of uncertainty-aware multimodal 3D object detection. Sec. IV presents the simulation results and Sec. V concludes.

II. Current Art on 3D Object Detection

In this study, we focus on 3D object detection as a case study to demonstrate the efficacy of MI-based conformal feature fusion in achieving uncertainty awareness in multimodal sensing, particularly at the edge. 3D object detection is fundamental for existing and emerging robotic platforms, such as robotaxis, to understand environments comprehensively by detecting, localizing, and classifying objects. While 2D object detection offers basic object localization and recognition, 3D detection further enriches applications by adding depth and distance insights. This necessitates a sophisticated perception system, integrating diverse sensors like RGB cameras, LiDAR, and mmWave radar, which mutually enhance their performance.

For deep learning-based 3D object detection, we specifically focus on RGB camera images and LiDAR point cloud data. Prior works have developed state-of-the-art 3D object detection framework through early [7], [8], intermediate [9]–[12], and late [13] information fusion of LiDAR and camera streams, as LiDAR features are rightfully superior to camera features in assessing depth for 3D tasks [14]. Early fusion improves data preprocessing and detection results, but often requires an additional network for initial image data processing, which increases inference runtime. Intermediate fusion offers deeper integration of multimodal features, which enhances bounding box prediction accuracy, but properly doing so remains an open problem due to the considerable distinctions in feature information and view points. Late fusion is more computationally efficient, but its performance is limited due to the lack of capturing the deep covariance between the modalities.

Notably, the above frameworks vary in their processing of LiDAR as well, with three primary methods identified as point-based [15]–[17], grid-based [18]–[20], and range-based [21], [22] methods. Point-based methods involve direct predictions based on downsampled points and extracted features, which has influenced many subsequent state-of-the-art works but makes it difficult to balance appropriate sampling with efficiency. Grid-based methods rasterize point cloud data into grid representations such as voxels (volumetric pixels), pillars (vertically extended voxels), or bird's-eye view (BEV) 2D feature maps, which can provide richer and more organized 3D information, potentially leading to more accurate predictions, but require more time and memory to process. Range-based methods consist of processing 2D range views (spherical projections of point clouds), which inherently contain 3D distance as opposed to simple RGB and can therefore be easily integrated with existing efficient 2D backbones but nonetheless suffer from common 2D issues (e.g., occlusion and scale variation) that exacerbate aleatoric uncertainty. Among these prior works, PointPillars [19], introduced in 2018, remains the fastest inference model on the 3D KITTI vision benchmark suite [6] with a 16 ms runtime. PointPillars is a LiDAR-only model that also demonstrated comparable accuracy to other state-of-the-art models published around the same time.

Most previous 3D object detection frameworks focus primarily on accuracy; there are relatively few works that have explored uncertainty quantification [23]–[28]. While these works underscore the significance of uncertainty and its potential to improve performance, their methodologies largely hinge on Bayes' theorem, maximum likelihood, or coarse statistics such as standard deviation. Such methods, deeply tied to data, model, and specific assumptions (e.g., gaussianity), can face numerical instability and might not be optimal for resource-limited systems, such as edge robotics. Addressing this critical need, in this paper, we discuss an uncertainty-aware 3D object detection framework comparable to PointPillars in speed and accuracy while

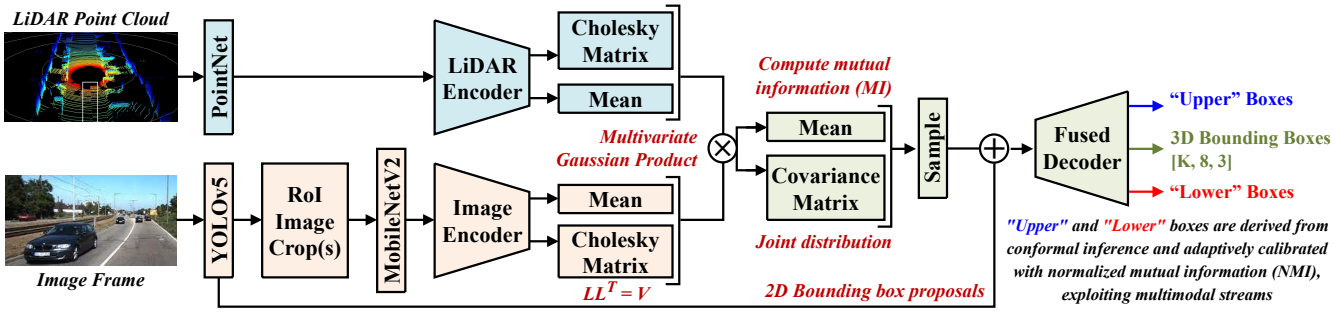

Fig. 2: **Model Architecture (Section 3):** The network designed for this work utilizes a variational autoencoder (VAE) featuring dual encoders for LiDAR point cloud and RGB camera image features. To extract the multimodal features, we rely on PointNet [15], YOLOv5s [29], and MobileNetV2 [30], keeping the network modular at a small expense of speed so that the feature extractors can be interchanged based on scene conditions. The information from the data streams are fused through a multivariate Gaussian product of their artificial 4D latent variables to approximate a proper covariant joint distribution. Mutual information between the multimodal features is computed using the joint mean and covariance, and a sample is extracted and concatenated with 2D bounding box proposals. Finally, a single decoder propagates the fused data to output K mean 3D bounding boxes of size $[8, 3]$ along with conformal inference-based upper- and lower- bound uncertainty estimates. K represents any number of detected objects.

providing statistically rigorous and generalizable uncertainty estimations via conformal inference.

III. Uncertainty-Aware Multimodal 3D Object Detection by Conformal Inference

This section provides an overview of the proposed uncertainty-aware 3D object detection framework with RGB camera and LiDAR sensors. The model architecture is shown in Fig. 2 and consists of a variational autoencoder (VAE) with parallel encoders for each sensor's extracted features and a single decoder that propagates information fused latent samples concatenated with 2D bounding box proposals to produce 3D bounding boxes and uncertainty bounds. To extract LiDAR features, the model relies on PointNet [15]. To obtain 2D bounding box proposals and subsequently extract camera features from cropped regions-of-interest (RoI), the model uses YOLOv5s [29] and MobileNetV2 [30]. This approach takes inspiration from PointFusion [9], as we opted for point-based LiDAR point cloud processing and intermediate LiDAR-camera fusion. These design choices were made primarily in consideration of efficiency and providing a solid information theoretic testbed for conformal inference.

A. Feature Fusion by Multivariate Gaussian Product

To effectively merge features extracted from RGB camera images and LiDAR point clouds, we adopt an approach inspired by [31]. They showed that the univariate Gaussian product of RGB and infrared image features optimally combines information from both modalities, ensuring the network remains resilient even when one data stream is suboptimal. Leveraging the Variational Autoencoder's (VAE) ability to approximate Gaussian distributions through reparameterization and Kullback-Leibler (KL) divergence of artificial latent variables (e.g., mean and variance) [32], we extend this statistical approach with a multivariate Gaussian product. Shifting from univariate to multivariate variables requires significant changes and optimizations.

Thus, our enhancement lies in operations on multivariate mean and covariance, ensuring richer representations of multimodal data. Instead of using the VAE's dual encoders for camera and LiDAR data to output variance, we utilize them to produce 4D Cholesky decompositions [33] of the presumed covariance matrices for each encoded feature set. A Cholesky decomposition (L) represents the square root of a covariance matrix and ensures symmetry and positive definiteness–two necessary criteria for subsequent matrix operations. Moreover, it encapsulates off-diagonal relationships of the latent variables, which often provide a truer representation of the covariance matrix but are commonly zeroed out in VAEs under the assumption of conditional independence.

From the Cholesky decompositions, we derive symmetric 4D covariance matrices for each set of encoded features using the matrix product $LL^T = V$. To fuse the feature information, we utilize both latent means and covariances. We refer to [34], which explains how to compute the mean μ and covariance V of a joint Gaussian distribution, given by equations in (1), from the product of n marginal distributions. Importantly, the equations are generalizable, suggesting that our framework can, in principle, handle an arbitrary number of sensor modalities.

$$\mu_{\text{joint}} = V_{\text{joint}} \sum_{i=1}^{n} V_i^{-1} \mu_i \tag{1a}$$

$$V_{\text{joint}}^{-1} = \sum_{i=1}^{n} V_i^{-1} \tag{1b}$$

Additionally, since these matrix computations involve inversion, we must address potential numerical instabilities that can ruin the approximations and cause divergence. To mitigate these concerns, we regularize the model with identity covariance and perform an eigen decomposition, denoted as $Q\Lambda Q^T = V$, on the joint covariance matrix whenever a Cholesky decomposition cannot be performed. With the latter step, we can ensure positive definiteness and avoidance of near-singularity by reconstructing the matrix after we have set the non-positive eigenvalues to a small positive constant (e.g., $1e$-6). Finally, with the proper joint mean and covariance, we can compute the mutual information (see below) between camera and LiDAR features and subsequently forward a sample from their fused distribution to the decoder along with 2D bounding box proposals.

B. Uncertainty Calibration by Mutual Information (MI)

Given the close relationship between conformal inference and information theory, we anticipate incorporating MI should improve our uncertainty-aware framework. MI quantifies the dependence between two random variables by examining the relationship between their marginal distributions and their joint distribution [35]. Effectively, MI assesses the uncertainty of one random variable in explaining the information of another. Previous studies have demonstrated that maximizing MI between the input feature space and latent space enhances the model's utilization of the latent representations [36], [37].

In this study, we utilize MI as a criterion for calibrating the conformal uncertainty intervals. To compute MI, we determine the determinants ($|\cdot|$) of the covariance matrices constructed for both the camera and LiDAR data and the covariance matrix of their combined joint distribution.

$$\text{MI} = \frac{1}{2}\log_2\left(\frac{|V_{RGB}||V_{LiDAR}|}{|V_{joint}|}\right) \tag{2}$$

Afterward, we approximate the Shannon entropy [35] of the two feature sets' covariances with (3) and use them to normalize the MI (i.e., compute NMI [38]) to be within the range of [0,1] with (4).

$$\text{H} = \frac{1}{2}\log_2\left((2\pi e)^4|V|\right) \tag{3}$$

$$\text{NMI} = \frac{2\text{MI}}{\text{H}_{RGB} + \text{H}_{LiDAR}} \tag{4}$$

It is important to note that, in theory, MI is upper bounded by the maximum of the the Shannon entropies of the random variables involved. However, because the VAE's latent random variables typically have unbounded support (because activation functions such as ReLU and others have unbounded ranges), it is possible to run into stability issues where a network could continue optimizing its parameters leading to divergent MI estimates. To fix this, we add the $softsign()$ activation function shifted by +1 to bind the latent variables to the range $[0, 2]$ and stabilize the network. This activation function resembles the hyperbolic tangent but is less steep and therefore saturates slower. In the next subsection, we discuss the placement of the NMI metric into the loss function.

C. Uncertainty Weighted Loss by Conformal Inference (CI)

CI offers a model-agnostic method for uncertainty quantification that seamlessly integrates with any foundational model possessing intrinsic uncertainty measures, such as quantile regression. The intervals guarantee marginal coverage of the truth based on a user-defined coverage rate [2]–[4]. Marginal coverage represents the average probability, taken over all considered samples, that true values will fall within the prediction intervals. It is analytically guaranteed by using a portion of the training data as a calibration set to compute conformity scores of new observations to prior information, which are used to calibrate the uncertainty intervals.

The conformalized joint prediction (CJP) method presented in our prior work [39] demonstrated a unique form of multivariate cross-conformal inference where a model is jointly trained to output point (mean) predictions and conditional quantiles that serve as upper and lower prediction bounds, capturing true aleatoric and epistemic uncertainty. To construct the prediction bounds, the method requires calibration of the sample data during training so as to guide the model to center predictions and maintain marginal coverage. Cross-conformal inference involves performing a number of calibration steps over all of the training data, striking a balance between the statistical efficiency of full-conformal prediction and the speed of split-conformal prediction. The method performs the calibrations dynamically over the randomized training batches as part of a multi-task loss function that simultaneously prioritizes reconstruction, KL divergence, and uncertainty interval centeredness, tightness, and coverage. As a result, it is shown that the intervals are highly tunable, flexible, and adaptive.

We make the following impactful modifications to the loss function presented in our prior work. First, we weight the reconstruction loss with a small uncertainty penalty to guide the network to prioritize resolving higher uncertainty in certain training batches. Secondly, we regularize the KL divergence with 4D covariance instead of variance. Finally, we dynamically tune the balance between interval sharpness (i.e., uncertainty distance) and marginal coverage with normalized mutual information (NMI). For completeness, the loss function is provided as:

$$\begin{aligned}\mathcal{L}_{Total} = &\ \text{SmoothL1}_{loss}(y, \hat{y}) \times (1 + 0.01U) \\ &+ \text{KL}_{div}(\mu_{joint}, V_{joint}) \\ &+ \text{INTSCORE}_{loss}(y, Q_l, Q_h, \{\alpha_l, \alpha_h\}) \\ &+ \text{COMCAL}_{loss}(y, p^{cov}_{avg}, Q_l, Q_h, NMI)\end{aligned} \tag{5}$$

where

$$\text{KL}_{div} = \frac{1}{2}\left(Tr(V) + \mu_{joint}\mu^T_{joint} - 4 - \log(|V|)\right) \tag{6}$$

$$\begin{aligned}\text{INTSCORE}_{loss} = &\ (Q_h - Q_l) + \frac{2}{\alpha}(Q_l - y)\mathbb{I}\{y < Q_l\} \\ &+ \frac{2}{\alpha}(y - Q_h)\mathbb{I}\{y > Q_h\}\end{aligned} \tag{7}$$

$$\text{COMCAL}_{loss} = (1 - NMI) \times \text{CAL}_{obj} + NMI \times \text{SHARP}_{obj} \tag{8a}$$

and

$$\begin{aligned}&\text{CAL}_{obj} = \\ &\mathbb{I}\{p^{cov}_{avg} < p\} \times \frac{1}{N}\sum_{i=1}^{N}[(y_i - Q_{l,h}(x_i))\mathbb{I}\{y_i > Q_{l,h}\}] + \\ &\mathbb{I}\{p^{cov}_{avg} > p\} \times \frac{1}{N}\sum_{i=1}^{N}[(Q_{l,h}(x_i) - y_i)\mathbb{I}\{y_i < Q_{l,h}\}]\end{aligned} \tag{8b}$$

$$\text{SHARP}_{obj} = \mathbb{I}\{p \leq 0.5\} \times \frac{1}{N}\sum_{i=1}^{N} Q_l(x_i) - Q_h(x_i) \\ + \ \mathbb{I}\{p > 0.5\} \times \frac{1}{N}\sum_{i=1}^{N} Q_h(x_i) - Q_l(x_i) \tag{8c}$$

In the equations, x represents input samples, y represents 3D bounding box labels, $\hat{y}$ represents predictions, U represents a singular uncertainty distance metric calculated by averaging the prediction interval length of each output dimension per training batch, Q_h and Q_l are each dimension's upper and lower quantile estimates used to calculate U, α_h and α_l are the 95^{th} and 5^{th} percentile coverage bounds that assert Q_h and Q_l, p is the chosen marginal coverage rate ($\alpha_h - \alpha_l = 90\%$), $\mathbb{I}$ is the indicator function, $Tr()$ is the trace function, and p_{avg}^{cov} is the estimated probability that the label values lie within $[Q_l, Q_h]$, averaged over the randomized training batches. $Q_{l,h}$ is meant to indicate that CAL_{obj} is computed separately for both Q_l and Q_h and then added together.

Focusing on the two lesser-known loss components—INTSCORE_{loss} is used to influence the model to maintain centered quantile intervals while COMCAL_{loss} is used to control the balance between minimizing the uncertainty intervals and increasing marginal coverage, as reflected in the sub-objectives CAL_{obj} and SHARP_{obj}. Notably, we insert NMI from Section III-B, averaged in each training batch, into COMCAL_{loss} to dynamically control the calibration balance during training as opposed to setting a static value. Intuitively, the model is influenced to be less uncertain when the MI between the RGB camera and LiDAR features is high. Therefore, uncertainty and MI should be inversely correlated.

IV. Results and Discussions

This section details our observations from applying the framework presented in Section III to 3D object detection involving RGB cameras and LiDAR point cloud inputs. Towards this, the primary goal of the proposed framework is to enable lightweight, conformalized uncertainty awareness while including principles of entropy, MI, and feature fusion based on a multivariate Gaussian product. This combination of theories is used to improve the model's uncertainty estimates through CI while operating under edge device constraints. Notably, uncertainty estimation can become difficult and unstable for a task such as 3D object detection, where there is a varied number of multivariate objects to be assessed per input sample. Therefore, taking an approach deeply rooted in information and statistical theory is imperative.

As projected in Section III-B, it is shown in Fig. 3 that the average uncertainty and normalized mutual information (NMI) obtained via conformalized feature fusion are inversely correlated over the duration of training the model described in Fig. 2. It is important to note that while mutual information is static given a discrete input feature space, here we are deriving it from artificial latent representations that are optimized during training. Hence, the value of NMI can change during training as the embedded information is better understood. While the estimated NMI increased between

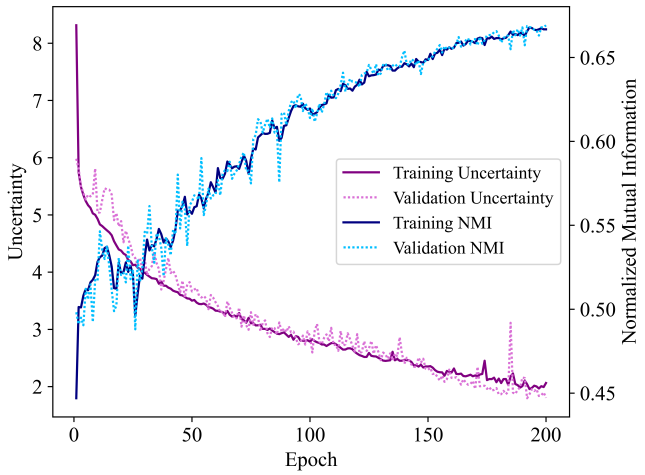

Fig. 3: **Average Uncertainty and Normalized Mutual Information (NMI) vs. Epoch:** Explicit uncertainty and NMI obtained from performing conformal inference and intermediate feature fusion are averaged across training batches in each epoch. Uncertainty and NMI are inversely correlated, influencing the model to be more confident in predictions when mutual information is high and *vice versa*.

the camera and LiDAR data, the overall uncertainty in the predictions decreased. This uncertainty metric U is used to weight the SmoothL1 reconstruction loss, while the NMI is used to calibrate this uncertainty in (8a). To the best of our knowledge, this is the first work demonstrating a stable combination of explicitly translatable uncertainty weighting and mutual information in a multitask loss function where both influence each other.

Table I quantitatively compares our proposed framework to similar works predicting 3D bounding boxes for cars in the seminal KITTI 3D detection dataset. The easy, moderate, and hard percentage scores are of average precision in 3D bounding box regression (AP_{3D}), which is based on precision-recall calculations with an intersection-over-union (IoU) threshold of 0.7 in various scene conditions. Most metrics are taken directly from the KITTI source website, where the various referenced models have been submitted for result reproduction. Our model is approximately 38% faster than PointPillars without suffering an equal accuracy loss, making it suitable for edge robotics. The runtime metrics we provide are adjusted for hardware differences amounting to an approximate 3× performance gap, given PointPillars used an NVIDIA GTX 1080 Ti desktop, and we used an NVIDIA RTX 4090 laptop. For reference, the 1080 Ti is only 3-5× faster than the latest NVIDIA Jetson AGX Orin edge computing device in relevant tasks. Unlike the other works, our model maintains a relatively consistent accuracy across each benchmark case, a unique result of using a VAE and conformal inference.

Furthermore, by factoring the marginal coverage of the upper- and lower-bound uncertainty boxes calibrated with NMI into the IoU calculations, the average precision increased by at least 39%. Accordingly, we propose a new evaluation metric, *mean average uncertainty* (MAU), to track the combined average uncertainty in predicting each corner of K 3D bounding boxes (i.e., $\frac{1}{K}\sum_{i=1}^{K} u_i$). A key obser-

Table I: **Comparison of Proposed 3D Object Detection Framework to Similar Work on KITTI Cars** (AP_{3D})

Reference	Modality	LiDAR Rep.	Fusion Type	Uncertainty	Runtime (ms)	Easy (%)	Mod. (%)	Hard (%)
PointFusion [9]	Cam+LiDAR	Points	Intermediate	No	–	74.71	61.24	50.55
ContFuse [40]	Cam+LiDAR	Grid (BEV)	Intermediate	No	60	83.68	68.78	61.67
MVX-Net [10]	Cam+LiDAR	Grid (Voxels)	Intermediate	No	–	83.2	72.7	65.2
EPNet [41]	Cam+LiDAR	Points	Intermediate	No	100	89.81	79.28	76.40
MMF [42]	Cam+LiDAR	Grid (BEV)	Intermediate	No	100	89.05	82.50	77.59
MV3D [11]	Cam+LiDAR	Multiple	Intermediate	No	360	74.97	63.63	54.00
3D-CVF [43]	Cam+LiDAR	Grid (BEV)	Intermediate	No	60	89.20	80.05	73.11
AVOD [12]	Cam+LiDAR	Grid (BEV)	Intermediate	No	100	83.07	71.76	65.73
CLOCs [13]	Cam+LiDAR	Multiple	Late	No	100	89.16	82.28	77.23
PointPillars [19]	LiDAR	Grid (Pillars)	Intermediate	No	16	82.58	74.31	68.99
Ours	Cam+LiDAR	Points	Intermediate	Yes	**9.87***	62.84	58.66	60.89
Ours w/ NMI-calibrated Uncertainty					14.82*	87.64 (MAU=3.52)	**89.83** (MAU=3.59)	**92.26** (MAU=3.62)

*These models were characterized on an NVIDIA RTX 4090 laptop; the metrics are adjusted for an NVIDIA GTX 1080 Ti desktop.

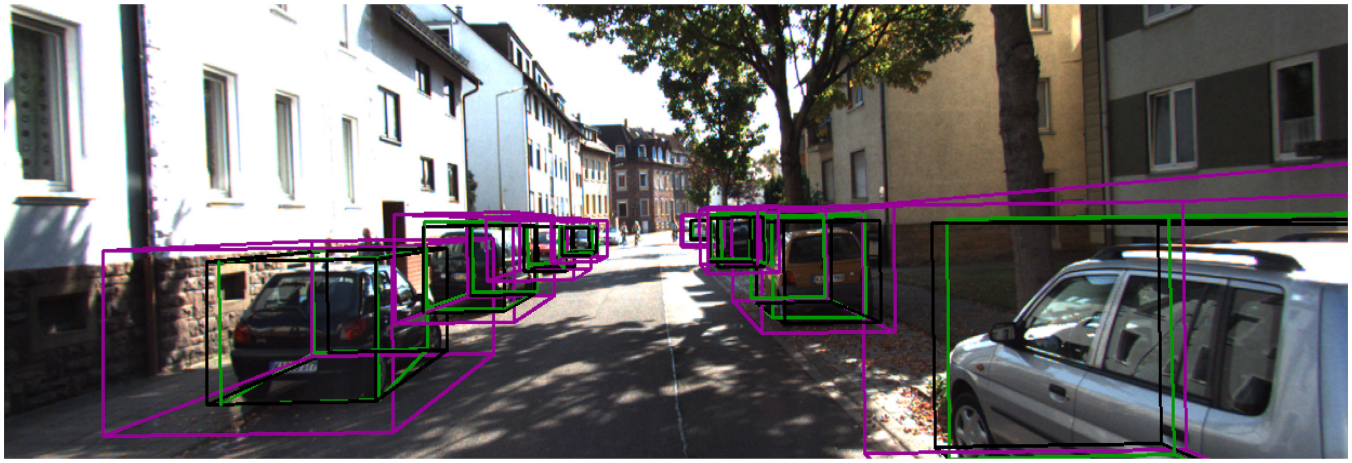

Fig. 4: **Uncertainty in 3D Bounding Box Regression:** Ground truth (black), predicted (green), and uncertainty (purple) 3D bounding boxes are visualized in a sample KITTI image of 8 cars with various occlusion and truncation status. As described in Section III-C, the uncertainty boxes represent a combination of upper- and lower-bound conditional quantiles obtained via conformal inference.

vation here is that, with uncertainty included, the average precision slightly increased with more difficult predictions while MAU also increased. This indicates that the model appropriately prioritized predictions where uncertainty was greater. However, it is worth noting that there are fewer annotations in the moderate and hard cases, so the model has fewer chances of being imprecise compared to the easy case. Overall, we show that robust uncertainty-awareness can improve the reliability of a model's predictions in making critical decisions and considerably improve accuracy. By maintaining a generalizable methodology, our work can be integrated to improve metrics in other models for various tasks.

Fig. 4 provides a qualitative assessment of the uncertainty in predicting the 3D bounding boxes. We display the ground truth box in black, the predicted box in green, and a combined uncertainty box in purple that encompasses the upper- and lower-bound boxes. This level of accuracy and precision in estimating and visualizing uncertainty in 3D object detection has not been demonstrated previously. A primary benefit of such assessment is that even if the model appears to be predicting well, a large uncertainty estimate can direct it to assert caution appropriately, such as when the sensors are not performing well or are impaired externally.

V. Conclusion

We presented a novel framework for quantifying, calibrating, and leveraging uncertainty in data-driven, multimodal deep learning at the edge. The proposed methodology, applied to 3D object detection, includes conformal inference, elements of information theory, and Gaussian feature fusion. Our research demonstrates that integrating uncertainty awareness not only increases reliability, but also improves prediction accuracy and precision. The approach is both generalizable and scalable, allowing it to be adapted to any task or dataset where uncertainty awareness should be considered, especially when under considerable resource constraints such as in edge robotics. The integration of information theory and conformal inference offers benefits that extend beyond individual results in the deep learning domain.

REFERENCES

[1] A. Kendall and Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 5580–5590.

[2] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Berlin, Heidelberg: Springer-Verlag, 2005.

[3] G. Shafer and V. Vovk, "A Tutorial on Conformal Prediction," *J. Mach. Learn. Res.*, vol. 9, p. 371–421, jun 2008.

[4] A. N. Angelopoulos and S. Bates, "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification," 2021. [Online]. Available: https://arxiv.org/abs/2107.07511

[5] V. Manokhin, "Awesome Conformal Prediction," Apr. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6467205

[6] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[7] C. Ruizhongtai Qi, W. Liu, C. Wu, H. Su, and L. Guibas, "Frustum PointNets for 3D Object Detection from RGB-D Data," 06 2018, pp. 918–927.

[8] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential Fusion for 3D Object Detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4603–4611.

[9] D. Xu, D. Anguelov, and A. Jain, "PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 244–253.

[10] V. A. Sindagi, Y. Zhou, and O. Tuzel, "MVX-Net: Multimodal VoxelNet for 3D Object Detection," *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7276–7282, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:102483619

[11] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-View 3D Object Detection Network for Autonomous Driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[12] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3D Proposal Generation and Object Detection from View Aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1–8.

[13] Su Pang and Daniel Morris and Hayder Radha, "Clocs: Camera-lidar object candidates fusion for 3d object detection," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10 386–10 393, 2020.

[14] J. Mao, S. Shi, X. Wang, and H. Li, "3D Object Detection for Autonomous Driving: A Comprehensive Survey," *International Journal of Computer Vision*, vol. 131, no. 8, pp. 1909–1963, Apr. 2023.

[15] R. Charles, H. Su, K. Mo, and L. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," 07 2017, pp. 77–85.

[16] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic Graph CNN for Learning on Point Clouds," *ACM Trans. Graph.*, vol. 38, no. 5, 2019.

[17] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–779, 2018.

[18] Y. Zhou and O. Tuzel, "VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection," 06 2018, pp. 4490–4499.

[19] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast Encoders for Object Detection From Point Clouds," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 689–12 697.

[20] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely Embedded Convolutional Detection," *Sensors*, vol. 18, no. 10, 2018.

[21] G. P. Meyer, A. Laddha, E. Kee, C. Vallespi-Gonzalez, and C. K. Wellington, "LaserNet: An Efficient Probabilistic 3D Object Detector for Autonomous Driving," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 669–12 678.

[22] M. Rapoport-Lavie and D. Raviv, "It's All Around You: Range-Guided Cylindrical Network for 3D Object Detection," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 2992–3001.

[23] G. P. Meyer and N. Thakurdesai, "Learning an Uncertainty-Aware Object Detector for Autonomous Driving," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 521–10 527.

[24] H. Chen, Y. Huang, W. Tian, Z. Gao, and L. Xiong, "MonoRUn: Monocular 3D Object Detection by Reconstruction and Uncertainty Propagation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 374–10 383.

[25] I. Oleksiienko and A. Iosifidis, "Uncertainty-Aware AB3DMOT by Variational 3D Object Detection," *arXiv preprint arXiv:2302.05923*, 2023.

[26] Y. Zhong, M. Zhu, and H. Peng, "Uncertainty-Aware Voxel based 3D Object Detection and Tracking with von-Mises Loss," *arXiv preprint arXiv:2011.02553*, 2020.

[27] Y. Liu, N. Mishra, M. Sieb, Y. Shentu, P. Abbeel, and X. Chen, "Autoregressive Uncertainty Modeling for 3D Bounding Box Prediction," *arXiv preprint arXiv:2210.07424*, 2022.

[28] D. Feng, L. Rosenbaum, F. Timm, and K. Dietmayer, "Leveraging Heteroscedastic Aleatoric Uncertainties for Robust Real-Time LiDAR 3D Object Detection," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 1280–1287.

[29] G. Jocher, "YOLOv5 by Ultralytics," 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," 2017. [Online]. Available: https://arxiv.org/abs/1704.04861

[31] L. Ren, Z. Pan, J. Cao, and J. Liao, "Infrared and visible image fusion based on variational auto-encoder and infrared feature compensation," *Infrared Physics and Technology*, vol. 117, 2021.

[32] C. Doersch, "Tutorial on variational autoencoders," 2016. [Online]. Available: https://arxiv.org/abs/1606.05908

[33] G. H. Golub and C. F. van Loan, *Matrix Computations*, 4th ed. JHU Press, 2013.

[34] P. A. Bromiley, "Products and Convolutions of Gaussian Probability Density Functions," 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:18045887

[35] T. M. Cover and J. A. Thomas, *Elements of Information Theory* . USA: Wiley-Interscience, 2006.

[36] A. L. Rezaabad and S. Vishwanath, "Learning Representations by Maximizing Mutual Information in Variational Autoencoders," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 2729–2734.

[37] S. Zhao, J. Song, and S. Ermon, "InfoVAE: Balancing Learning and Inference in Variational Autoencoders," in *AAAI Conference on Artificial Intelligence*, 2019.

[38] N. X. Vinh, J. Epps, and J. Bailey, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance," *J. Mach. Learn. Res.*, vol. 11, p. 2837–2854, 2010.

[39] A. C. Stutts, D. Erricolo, T. Tulabandhula, and A. R. Trivedi, "Lightweight, Uncertainty-Aware Conformalized Visual Odometry," *arXiv preprint arXiv:2303.02207*.

[40] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep Continuous Fusion for Multi-sensor 3D Object Detection," in *Computer Vision–ECCV 2018: 15th European Conference, Proceedings*, 2018, pp. 663–678.

[41] T. Huang, Z. Liu, X. Chen, and X. Bai, "Epnet: Enhancing point features with image semantics for 3d object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part XV 16*, 2020, pp. 35–52.

[42] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-Task Multi-Sensor Fusion for 3D Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[43] J. H. Yoo, Y. Kim, J. Kim, and J. W. Choi, "3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part XXVII 16*, 2020, pp. 720–736.