Uncertainty Modelling and Out-of-Domain Identification for Object Detection in Automated Driving

Zur Erlangung des akademischen Grades eines

DOKTORS DER INGENIEURWISSENSCHAFTEN (Dr.-Ing.)

von der KIT-Fakultät für Maschinenbau
des Karlsruher Instituts für Technologie (KIT)

angenommene

DISSERTATION

von

M.Sc. Ahmed Hammam

Tag der mündlichen Prüfung: 29.01.2026

Hauptreferent: Prof. Dr.-Ing. Christoph Stiller
Korreferent: Prof. Dr.-Ing. Arne Rönnau

Abstract

Highly automated driving (HAD) systems rely on perception modules to interpret real-time sensor data and provide information about the vehicle's environment. Although neural networks often achieve high accuracy in locating and classifying these objects, the confidence scores which they produce do not reliably reflect the uncertainty of their predictions. This limitation may lead the model to produce overconfident incorrect predictions, particularly in complex or unfamiliar situations, which can pose risks to system safety. Such predictions may propagate to downstream components, including trajectory planning and control modules. Another concern is that objects which are not present in the training data may either not be detected by the model or may be detected but misclassified as one of the in-domain classes with high certainty. This limits the model's ability to distinguish out-of-domain (OOD) objects from in-domain ones and increases the risk of erroneous system behavior.

This thesis addresses methods to improve the reliability of single-camera-based environment perception in automated driving systems by leveraging uncertainty estimation and out-of-domain instance detection and identification. The focus is on a monocular, front-facing camera configuration. The work proposes a multi-task architecture comprising a shared backbone and specialized decoders for semantic segmentation, instance segmentation, and depth estimation. Central to the architecture, this work proposes the use of a Dirichlet layer which outputs class probabilities, an uncertainty measure, and out-of-domain detection scores, enabling more reliable uncertainty estimates. Additionally, an intermediate layer variational inference module is proposed, introducing controlled stochasticity into the latent representations. While the Dirichlet layer models uncertainty at the output level, the ILVI module enables the estimation of epistemic uncertainty within the feature space. Together, these components provide complementary coverage of uncertainty across the network's representational hierarchy.

The proposed architecture is evaluated on public benchmarks and analyzed across a comprehensive set of metrics to assess uncertainty estimation, out-of-domain (OOD) detection capability, and the preservation of segmentation performance. Results show that the proposed Dirichlet layer increases certainty reliability, reducing overconfident misclassifications and improving the alignment between predicted confidence and actual model accuracy. When combined with the ILVI module, the system effectively identifies out-of-domain objects. Furthermore, it is worth noting that the architecture achieves improved semantic and instance segmentation performance compared to baseline models.

The integration of these components into a unified, single-camera architecture demonstrates that uncertainty modeling and OOD detection can be incorporated without degrading segmentation performance on in-domain data. The proposed contributions are modular and model-agnostic, making them applicable to a broad range of perception architectures for enhancing both predictive quality and system reliability.

Kurzfassung

Systeme zum hochautomatisierten Fahren basieren auf Perzeptionsmodulen, die Sensordaten in Echtzeit interpretieren und Informationen über die Fahrzeugumgebung bereitstellen. Obwohl tiefe neuronale Netze häufig eine hohe Genauigkeit bei der Lokalisierung und Klassifizierung von Objekten erreichen, spiegeln die ausgegebenen Konfidenzwerte die tatsächliche Unsicherheit der Vorhersagen nicht zuverlässig wider. Diese Unzulänglichkeit kann dazu führen, dass Modelle insbesondere in komplexen oder unbekannten Szenen falsche Vorhersagen mit hoher statt niedriger Konfidenz liefern. Dies stellt ein Sicherheitsrisiko für das Gesamtsystem dar und kann zu Fehlern in nachgelagerten Komponenten wie der Trajektorienplanung oder dem Regelungsmodul führen. Ein weiteres Problem besteht darin, dass Objekte, die nicht in den Trainingsdaten enthalten sind, entweder nicht erkannt oder fälschlicherweise mit hoher Konfidenz als bekannte Klassen eingestuft werden. Dies schränkt die Fähigkeit des Modells ein, zwischen bekannten und unbekannten Objekten (Out-of-Domain, OOD) zu unterscheiden, und erhöht das Risiko eines fehlerhaften Systemverhaltens.

Die vorliegende Arbeit behandelt Methoden zur Verbesserung der Zuverlässigkeit einer auf Kameradaten basierenden Umgebungsperzeption für Systeme zum automatisierten Fahren durch die Schätzung von Unsicherheiten sowie die Erkennung von OOD-Instanzen. Der Fokus liegt dabei auf einer monokularen, nach vorne gerichteten Kamerakonfiguration. Die Arbeit schlägt eine Multi-Task-Architektur mit einem gemeinsamen Backbone und spezialisierten Dekodern für semantische Segmentierung, Instanzsegmentierung und Tiefenschätzung vor. Im Zentrum dieser Architektur steht eine Dirichlet-Schicht, die Klassenwahrscheinlichkeiten, Unsicherheitsschätzungen und ein Detektionsmaß für OOD-Objekte ausgibt und so eine verlässlichere Unsicherheitsbewertung ermöglicht.

Zudem wird ein Intermediate-Layer-Variational-Inference-Modul (ILVI) vorgeschlagen, das kontrollierte stochastische Eigenschaften in die latenten

Repräsentationen des Netzes einführt. Während die Dirichlet-Schicht Unsicherheiten auf der Ausgabeebene modelliert, ermöglicht das ILVI-Modul die Schätzung epistemischer Unsicherheiten im Merkmalsraum. Zusammen decken diese Komponenten komplementäre Unsicherheitsaspekte über die gesamte hierarchische Repräsentation des Netzes ab.

Die vorgeschlagene Architektur wird mittels öffentlich verfügbarer Benchmarks evaluiert und anhand einer Vielzahl von Metriken analysiert, um die Qualität der Unsicherheitsschätzung, die Fähigkeit zur Erkennung von OOD-Objekten sowie die Beibehaltung der Segmentierungsleistung zu beurteilen. Die Ergebnisse zeigen, dass die vorgeschlagene Dirichlet-Schicht die Zuverlässigkeit der Unsicherheitsschätzung verbessert, übermäßig selbstsichere Fehlklassifikationen reduziert und eine bessere Übereinstimmung zwischen vorhergesagter Konfidenz und tatsächlicher Modellgenauigkeit erzielt. In Kombination mit dem ILVI-Modul ist das System in der Lage, OOD-Objekte effektiv zu identifizieren. Darüber hinaus ist hervorzuheben, dass die Architektur eine verbesserte Leistung bei der semantischen Segmentierung und der Instanzsegmentierung im Vergleich zu Basismodellen erreicht.

Die Integration dieser Komponenten in eine einheitliche Architektur zur Verarbeitung von Daten einer Monokamera zeigt, dass Schichten zur Unsicherheitsmodellierung und OOD-Erkennung implementiert werden können, ohne die Segmentierungsleistung für bekannte Objekte zu beeinträchtigen. Die vorgeschlagenen Beiträge sind modular und modellagnostisch, wodurch sie für eine Vielzahl von Perzeptionsarchitekturen geeignet sind und sowohl die Vorhersagequalität als auch die Zuverlässigkeit des Gesamtsystems verbessern.

Acknowledgements

I would like to express my deepest gratitude to everyone whose support and encouragement have made this dissertation possible. First and foremost, I thank my wife for always motivating me to persevere and excel. My heartfelt appreciation also goes to my parents and siblings, who have consistently stood behind me and inspired me throughout this journey.

I extend my sincere thanks to the Institute of Measurement and Control (MRT) at the Karlsruhe Institute of Technology (KIT) and all its members. In particular, I am deeply grateful to Professor Stiller for his guidance, generosity, and support. He not only provided a stimulating scientific environment but also granted me the freedom to pursue this work successfully. His invaluable mentorship played a crucial role in shaping my research, and I am thankful for the many ways his insights fostered my development.

My work has also benefited tremendously from my time at Opel Automobile GmbH, under the direction of Nikolas Wagner. I am deeply grateful to Frank Bonarens and Seyed Eghbal Ghobadi for giving me the opportunity to join their team, where I gained invaluable knowledge in artificial intelligence and autonomous vehicles, and participated in collaborative consortium projects and cross-company initiatives. Their trust and encouragement, as well as their openness to publishing research internationally, were instrumental in broadening my professional horizons. I also wish to thank my colleagues at Opel, Juncong Fei, Patrick Feifel, Lukas Stäker, and Philip Heidenreich, for their continued assistance. I am especially indebted to Stefan Berger for his thorough feedback and thoughtful reviews of my work. Finally, I would like to sincerely thank Felix Hauser for his unwavering support and generous help in all aspects throughout the entire journey of this thesis.

It has been a privilege to collaborate with such outstanding individuals and organizations. I owe a great deal of my academic and professional growth to their generosity, insights, and willingness to support my endeavors.

Table of Contents

Abbreviations and Notations

Abbreviations

A2D2	Audi Autonomous Driving Dataset
AC	Accurate Certain
AD	Automated Driving
ADS	Automated Driving systems
AE	Autoencoder
AP	Average Precision
ASPP	Atrous Spatial Pyramid Pooling
AU	Accurate Uncertain
AUROC	Area Under the Receiver Operating Characteristic Curve
AvU	Accuracy vs. Uncertainty
BDD	Berkeley DeepDrive
BNN	Bayesian Neural Network
CE	Cross-Entropy
CNN	Convolutional Neural Network
DETR	Detection Transformer
DNN	Deep Neural Network

DR	Detection Rate
ECE	Expected Calibration Error
ELBO	Evidence Lower Bound
FCN	Fully Convolutional Network
FN	False Negatives
FNR	False Negative Rate
FP	False Positives
FPR	False Positive Rate
GAN	Generative Adversarial Network
IC	Inaccurate Certain
ID	In-Domain
ILVI	Intermediate Layer Variational Inference
IoU	Intersection over Union
IU	Inaccurate Uncertain
iOOD	Instance-level Out-of-Domain
KITTI	Karlsruhe Institute of Technology and Toyota Technological Institute
KL	Kullback-Leibler
MC	Monte Carlo
MCE	Max Calibration Error
MLE	Maximum Likelihood Estimation
ML	Machine Learning

mIoU	Mean Intersection over Union
NMS	Non-Maximum Suppression
OOD	Out-of-Domain
ROC	Receiver Operating Characteristic
RoI	Region of Interest
RSE	Relative Squared Error
SSL	Self-Supervised Learning
SOTA	State-of-the-Art
TN	True Negatives
TP	True Positives
TPR	True Positive Rate
VAE	Variational Autoencoder
VI	Variational Inference

Notations

c	Class
d	Distance
f	Function
h	Latent layer
y	Model output
z	Logit value

α	Dirichlet concentration parameters
δ	Threshold
λ	Weighting factor
μ	Gaussian distribution mean
ϕ	Variational parameters
σ	Gaussian distribution variance
θ	Model parameters
B	Prediction Bin
Dir	Dirichlet distribution
F	Maximum Likelihood Estimate
H	Height
K	Set of classes
N	Pixels count
W	Width
$\mathcal{L}$	Model Loss
$\mathcal{N}$	Normal distribution
$\mathcal{X}$	Input space
$\mathcal{Y}$	Target space

1 Introduction

Automated driving systems (ADS) are being developed with the aim of optimizing traffic flow, reducing emissions, and improving road safety. These systems consist of multiple interconnected modules, including perception, prediction, planning, and control [JGB+20]. The perception module provides an abstracted and structured representation of the environment based on raw sensor data. The accuracy and reliability of this module are essential, as its outputs serve as inputs for all downstream components. Errors in perception can propagate through the system and may result in suboptimal planning or control actions.

A central component of perception is image segmentation, which assigns a class or instance label to each pixel in the input image. This enables the system to localize and distinguish between relevant entities in the environment, such as road users, infrastructure, or static obstacles. Segmentation models based on deep learning have demonstrated strong performance in controlled or benchmarked datasets [CVC+22]. However, in real-world deployments, these models are required to generalize to conditions not seen during training, including variations in illumination, weather, sensor noise, and scene composition. This introduces a discrepancy between the training and operational data distributions, leading to potential model failures [FHSR+20].

In such cases, the use of uncertainty estimation has been proposed as a mechanism to quantify the reliability of model predictions. Uncertainty estimates allow the system to differentiate between high-confidence and low-confidence outputs and can be used to adjust behavior accordingly. For example, lower confidence in object classification may result in reduced speed or increased distance to surrounding objects. In this context, epistemic uncertainty, which arises from model limitations or insufficient training coverage, is of particular interest. It reflects uncertainty due to a lack of knowledge and can potentially be reduced with additional data or model refinement. In contrast, aleatoric uncertainty, which stems from inherent sensor or environmental noise, is not

addressed in this work, as it typically requires specialized sensor models or multimodal inputs that lie outside the scope of this study.

Another important consideration is the presence of out-of-domain (OOD) inputs, which are samples that do not belong to the training distribution and may include previously unseen object categories or scene configurations. Standard classification and segmentation models are not designed to identify such inputs and may assign incorrect labels with high confidence. This behavior poses a risk in safety-critical systems such as ADS, where incorrect classification of unknown entities can lead to hazardous decisions. The ability to detect OOD inputs enables the system to invoke predefined fallback strategies, such as conservative planning or escalation to human supervision.

Finally, perception systems in ADS must operate under real-time constraints, typically requiring inference at frame rates suitable for timely reaction [FHSR+20]. This necessitates the use of models and methods that impose minimal computational overhead, as additional complexity may exceed the capabilities of automotive-grade embedded hardware. This work is situated within these constraints and addresses the need for uncertainty-aware and OOD-resilient perception under single-camera, real-time settings.

1.1 Research Questions and Contributions

This work addresses three central research questions related to uncertainty estimation and OOD detection in perception systems for ADS. Each research question corresponds to a contribution aimed at enhancing the reliability of uncertainty estimates, reducing false positive detections, and accurately detecting OOD objects in real-world driving scenarios. The overall goal is to improve uncertainty quantification and OOD detection capabilities of segmentation methods while preserving their core segmentation performance and efficiency.

The first research question asks how the reliability of uncertainty estimation in deep neural networks (DNNs) for automated driving (AD) can be improved. In response, this work introduces an uncertainty estimation methodology based on Dirichlet distributions combined with intermediate layer variational inference (ILVI). This approach is designed for both semantic segmentation and instance

segmentation tasks. The methodology focuses on producing more trustworthy uncertainty estimates.

The second research question examines how reliable uncertainty estimation can reduce false positive detections in AD systems. False positives (FP) and false negatives (FN) can hinder vehicle safety and performance. To address this challenge, the proposed contribution refines network performance to decrease FN while using uncertainty thresholding to filter out predictions that are likely to be incorrect. This strategy helps preserve accurate detections and lowers incorrect classifications, ultimately improving detection accuracy.

The third research question considers the challenges of identifying OOD objects in real-time driving scenarios. Using Dirichlet distributions, this work leverages Dirichlet strength, corresponding to the total concentration of the predicted Dirichlet distribution, to detect and identify OOD objects. This capability allows the system to recognize when it encounters an object that differs from its training data. By doing so, it addresses one of the barriers of perception systems to operate reliably when encountering OOD objects in the environment.

This thesis is scoped to the perception module of automated driving systems, with a focus on semantic and instance-level segmentation augmented by epistemic uncertainty estimation and OOD detection. Epistemic uncertainty, which captures the model's lack of knowledge, plays a central role in evaluating the reliability of predictions and supporting confidence-aware decision-making. All methods are developed under a camera-only constraint, using input from a single front-facing monocular camera, without reliance on LiDAR, radar, or multi-camera configurations. The proposed architecture and techniques are designed to ensure real-time applicability, where real-time is defined as introducing no additional inference-time computational overhead. This enables seamless integration into existing perception pipelines without compromising runtime performance or requiring specialized hardware. Additionally, the methods are designed to be model-agnostic, avoiding architectural changes that would limit their generalizability or hinder practical adoption.

1.2 Outline of the Thesis

This work is organized into five chapters, each addressing key aspects of uncertainty estimation and OOD objects' identification.

Chapter 2 provides an overview of the related work on DNNs for semantic and instance segmentation, along with a review of state-of-the-art (SOTA) approaches for uncertainty estimation and OOD identification.

Chapter 3 introduces the core contributions of this work. It describes the incorporation of Dirichlet distributions and ILVI on a DNN architecture for improving uncertainty estimation and OOD detection and identification. The mathematical foundations underlying the methodology are also discussed, providing a theoretical basis for the proposed approach.

Chapter 4 presents a comprehensive experimental evaluation of the proposed architecture on standard benchmark datasets. The improvement of uncertainty estimation and OOD identification is analyzed using performance metrics. A comparative analysis with existing methods is conducted to highlight the improvements introduced in this work.

Chapter 5 concludes the dissertation by summarizing the main findings and discussing their implications for AD perception systems. The limitations of the proposed approach are acknowledged and potential directions for future work are suggested.

2 Related Work and Fundamentals

The development of ADS relies heavily on accurate perception, which is achieved through advanced deep learning techniques. Object detection and segmentation form the core of perception systems, enabling vehicles to interpret and react to their surroundings. However, real-world driving conditions introduce significant challenges that necessitate proper uncertainty estimation and OOD identification to ensure the reliability of predictions [HAA+22].

Despite their success in computer vision tasks such as semantic and instance segmentation, DNNs are prone to incorrect predictions. A more critical concern arises when these incorrect outputs are accompanied by high certainty estimates, providing misleading information to downstream components of the system. This overconfidence can lead to inappropriate or hazardous decisions in ADS. To mitigate these risks and enhance the reliability of ADS, it is essential to incorporate uncertainty estimation techniques that quantify the model's confidence and robust OOD detection mechanisms that identify novel or unfamiliar inputs and enable appropriate safety measures.

This chapter reviews deep learning models for perception and approaches for uncertainty estimation and OOD detection. It begins by examining DNN segmentation models utilized in ADS perception. Following this, various uncertainty estimation techniques are explored, highlighting their strengths and limitations. Finally, the chapter investigates SOTA approaches for OOD detection.

2.1 Deep Neural Networks

DNNs have accomplished many milestones and are continuing to prove their validity in many supervised machine learning tasks [MS24, TZMF25]. The goal for a supervised network is to predict the target value $y \in \mathcal{Y}$ for an input $x \in \mathcal{X}$. In this work, the input space $\mathcal{X}$ corresponds to the space of images. Given a supervised machine learning problem with the task of classification, the target $\mathcal{Y}$ consists of a finite set of K classes where the task for the network is to predict the class of each out of the set of classes C. For the purpose of this work, a DNN is defined as a function $f_\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by $\theta \in \mathbb{R}^n$, with $n \in \mathbb{N}$ denoting the total number of model parameters, which maps an input $x \in \mathcal{X}$ to an output $f_\theta(x) \in \mathcal{Y}$.

2.1.1 Semantic Segmentation Models

Semantic segmentation is a fundamental task in computer vision that involves assigning a class label to each individual pixel in an image. Unlike image-level classification, which predicts a single label for the entire image, semantic segmentation provides dense, pixel-wise predictions that allow for a detailed understanding of scene content. Each pixel is categorized into a predefined set of classes, such as road, pedestrian, vehicle, building, or vegetation, enabling the network to produce a structured interpretation of the environment.

Fully Convolutional Networks

The introduction of fully convolutional networks (FCN) by Long et al. [LSD15] revolutionized semantic segmentation. FCN replaced fully connected layers with convolutional ones, enabling end-to-end training for pixel-wise classification. By using deconvolutional layers to upsample feature maps, FCN produce dense outputs.

Encoder-Decoder Architectures

Encoder-decoder architectures, such as U-Net [RFB15], improve upon FCN by using skip connections between the encoder and decoder to retain spatial information, resulting in better segmentation accuracy, particularly for fine structures. SegNet [BKC17] further optimizes this approach by using max-pooling indices from the encoder for upsampling in the decoder, reducing computational complexity while maintaining high segmentation quality.

DeepLab, introduced by Chen et al. [CPK+17], improves segmentation by employing atrous (dilated) convolutions to capture multi-scale context without significantly increasing computational cost. Its atrous spatial pyramid pooling (ASPP) module is particularly effective at handling objects of varying sizes. Later versions, such as DeepLabv3+ [CZP+18], further enhance performance by incorporating encoder-decoder architectures.

Transformers

Transformer-based models further enhance segmentation accuracy by integrating attention mechanisms [PTR+21]. Attention U-Net [OSF+18] adds attention gates to focus on relevant regions, while models like SegFormer [XWY+21] leverage Transformers' ability to capture global context alongside convolutional neural networks' (CNNs) local feature extraction, setting new benchmarks in the field.

2.1.2 Instance Segmentation Models

Instance segmentation is a computer vision task that unifies the goals of object detection and semantic segmentation. While semantic segmentation assigns a class label to each pixel in an image, it does not differentiate between multiple objects of the same class. Instance segmentation addresses this limitation by not only classifying each pixel but also distinguishing between individual instances of the same object class. Common approaches for instance segmentation extend the capabilities of traditional DNN-driven semantic segmentation networks.

Mask R-CNN

Mask R-CNN, introduced by He et al. [HGDG17], has had a significant impact on the field of instance segmentation and continues to serve as a benchmark for subsequent approaches. It extends the Faster R-CNN [RHGS15] object detection framework by adding a parallel branch for predicting segmentation masks for each region of interest (RoI), alongside the existing branches for classification and bounding box regression. A key innovation of Mask R-CNN is the use of RoIAlign, a feature extraction technique that preserves spatial alignment, enabling precise pixel-level mask predictions essential for accurate instance segmentation.

Path Aggregation Network (PANet), proposed by Liu et al. [LQQ+18], enhances Mask R-CNN by improving information flow between different levels of the feature pyramid. A feature pyramid refers to a multi-scale representation of the input image, where features are extracted at various spatial resolutions to detect objects of different sizes. PANet introduces bottom-up path augmentation, which complements the top-down pathway of the original feature pyramid by reinforcing low-level spatial features with high-level semantic context. This bidirectional information flow strengthens the entire feature hierarchy, enabling more effective feature reuse across scales. Additionally, PANet incorporates adaptive feature pooling to better capture fine-grained details. These enhancements lead to improved accuracy in both object detection and instance segmentation tasks.

EfficientNet

EfficientNet, proposed by Tan and Le [TL19], is a family of CNN architectures designed to optimize both accuracy and efficiency. Its innovation lies in its compound scaling method, which uniformly scales network depth, width, and input resolution through a set of fixed scaling coefficients. This systematic approach to model scaling ensures that each additional computational resource is used more effectively. As a result, EfficientNet delivers high functional performance with lower computational cost across a wide range of computer vision tasks.

In the context of instance segmentation, EfficientNet can serve as a powerful backbone due to balancing high predictive accuracy with computational efficiency.. By integrating EfficientNet into frameworks such as Mask R-CNN or PANet, researchers have reported gains in segmentation quality with relatively modest increases in computational overhead [TPL20, ZS20]. Its lightweight design also makes it an appealing choice for real-world applications where inference speed and resource constraints are critical factors.

Transformer-Based Models

Transformer-based architectures also have been integrated in instance segmentation models. Detection Transformer (DETR) [CMS+20], introduced by Carion et al., reimagines the detection and segmentation tasks using a Transformer architecture. DETR directly predicts the bounding boxes and class labels as a set-based prediction problem, followed by mask prediction for each detected instance. This approach simplifies the pipeline by removing the need for traditional post-processing steps like non-maximum suppression (NMS).

2.2 Uncertainty Estimation

The DNN, as described previously, is commonly known as a point estimate network, where no probability distributional functions are considered in its development or operation. These types of networks are well established, with the ability to be developed and deployed with the current modern architectures. However, such networks are unable to express uncertainty about their predictions, which poses reliability challenges [GTA+23].

Uncertainty is categorized into aleatoric uncertainty, which originates from inherent noise in the data, and epistemic uncertainty, which arises from the model itself. Standard neural networks do not model these uncertainties, requiring alternative approaches to address this limitation [HW21].

This work examines epistemic uncertainty, which plays a crucial role in evaluating prediction reliability and quantifying certainty in decision-making processes.

2.2.1 Bayesian Neural Networks

Unlike standard neural networks that produce point estimates for their parameters, Bayesian Neural Networks (BNNs) model their parameters as probability distributions [Mac92, FBLT20]. This probabilistic formulation enables BNNs to capture uncertainty by representing a range of plausible parameter values and updating these distributions as new data is observed. As a result, BNNs naturally account for both data uncertainty (aleatoric uncertainty) and model uncertainty (epistemic uncertainty). They often yield better-calibrated predictions compared to their non-Bayesian counterparts, mitigating issues associated with overconfident or underconfident outputs [MMN19, OFR+19].

Despite these advantages, BNNs suffer from significant computational challenges. Performing exact Bayesian inference requires evaluating high-dimensional integrals over the parameter space. While this remains computationally feasible for small-scale models, it does not scale to networks of practical size, rendering exact inference intractable. Consequently, a variety of approximation techniques have been proposed to make Bayesian inference more scalable and applicable to deep learning models [MBK21]. Two of the well-known Bayesian approximation techniques, Monte Carlo dropouts and deep ensembles, are discussed next.

2.2.2 Bayesian Approximations

Monte Carlo Dropout

Gal et al. demonstrated that training deep neural networks with dropout is mathematically equivalent to performing approximate Bayesian inference using variational inference (VI) with a specific variational distribution [Gal16]. In this interpretation, dropout defines a structured variational family, where the posterior over the model parameters is approximated by a Bernoulli distribution applied to the network's weights or activations.

Originally introduced as a regularization technique to mitigate overfitting by randomly deactivating neurons during training, dropout was later reinterpreted within a Bayesian framework as a form of approximate posterior inference over the model weights. Specifically, training a neural network with dropout

can be viewed as performing variational inference, where dropout induces a variational distribution that approximates the true posterior over the weights. Consequently, the learned weights correspond to the optimal variational parameters of a BNN.

At inference time, uncertainty estimates can be obtained by performing multiple stochastic forward passes with dropout active and aggregating the predictions. This technique, known as Monte Carlo (MC) Dropout, enables predictive uncertainty estimation. However, its practicality in real-time applications is limited by the need for multiple forward passes, which increases computational cost linearly with the number of samples [GDS20].

Deep Ensembles

Ensembles of DNNs, known as deep ensembles, have been shown to improve predictive performance in various classification tasks [GHM+22]. The underlying idea is that the aggregated predictions of multiple, slightly different DNNs yield more reliable results than a single network. A randomization-based approach using deep ensembles for uncertainty estimation in classification was introduced in [LPB16]. In this method, the ensemble members are treated as a uniformly weighted mixture model, with the mean of their predictions representing the final output and the variance capturing the associated uncertainty.

Each member in the ensemble shares the same architecture but is initialized with different weights. The models can also be trained on different subsets of the training data to enhance diversity. Increasing ensemble diversity is recommended by combining different initializations and training subsets, leading to a more robust ensemble of networks.

After training the DNNs for the ensemble, the models' parameters need to be stored, requiring an increasingly large storage capacity. Furthermore, while a single prediction from a trained network is typically fast, generating predictions from an entire DNN ensemble increases the overall computational cost.

2.2.3 Dirichlet-based Deep Neural Networks

Dirichlet-based DNNs utilize the Dirichlet distribution to model both class predictions and the uncertainty associated with them in multi-class classification [Tsi21, KCZ+21]. Instead of directly producing a probability vector, the network outputs a set of concentration parameters that define a distribution over class probabilities. A high concentration in one parameter results in a sharply peaked distribution, indicating strong certainty in the corresponding class. Conversely, when concentration values are more evenly distributed across classes, the resulting distribution is broader, reflecting decreased certainty in the prediction [XSG21].

A primary motivation for employing Dirichlet-based networks is their ability to provide a structured representation of uncertainty, which is lacking in conventional softmax-based classifiers [KHN23, XLZL23]. In contrast to point estimate DNNs, Dirichlet networks explicitly capture the uncertainty by modeling a distribution over possible predictions, facilitating a clearer differentiation between confident and uncertain predictions [MPV21, UHF21].

During training, the network learns to produce highly concentrated Dirichlet distributions for correct classifications while assigning more uniform distributions to uncertain or incorrect predictions [ZWXH24, DCYL23]. This structured approach to uncertainty modeling not only improves calibration but also provides a principled way to assess uncertainty of the model [WJ21].

Prior Networks

Prior Networks address uncertainty estimation by directly modeling a Dirichlet distribution over class probabilities, without relying on sampling-based methods [MG18a]. During training, the network is tasked with predicting a Dirichlet distribution that matches a predefined target: sharply peaked distributions for confident predictions (i.e., when the input is from a known class) and broader, more uniform distributions for uncertain or OOD inputs.

The training objective minimizes the Kullback-Leibler (KL) divergence between the predicted Dirichlet distribution and the target Dirichlet distribution [CFT19].

This training strategy encourages the model to produce highly concentrated Dirichlet distributions for in-domain, correctly classified samples, and low-concentration, near-uniform distributions for ambiguous or OOD samples. As a result, the network can produce both accurate class predictions and meaningful uncertainty estimates in a single forward pass. Unlike Bayesian Neural Networks (BNNs), which require expensive posterior sampling, Prior Networks achieve efficient and interpretable uncertainty quantification without the associated computational overhead [HTI+18].

Evidential Networks

Similar to Prior Networks, Evidential Networks integrate the Dempster-Shafer Theory of Evidence into deep learning to provide a richer framework for uncertainty estimation [SKK18]. Instead of predicting point estimates of class probabilities, Evidential Networks predict the parameters of a Dirichlet distribution, where the concentration parameters are interpreted as the amount of evidence supporting each class. By treating DNN predictions as subjective opinions, Evidential Networks allow the model to accumulate and refine evidence for different classes [Ton22].

The network is trained using a specialized evidential loss function that encourages both accurate classification and calibrated uncertainty. This loss is composed of two main terms. The first term minimizes the expected negative log-likelihood under the predicted Dirichlet distribution, thereby encouraging the model to classify known samples with high certainty. The second term introduces an uncertainty regularization loss, typically based on a Kullback-Leibler (KL) divergence, which penalizes predictions that are overconfident on incorrect or ambiguous samples [TXD21, EDHH24].

This training strategy ensures that the network not only improves its predictive performance but also learns to express uncertainty appropriately when facing ambiguous, or OOD inputs [THD24, RJP+22].

2.3 Out-of-Domain Identification Approaches: State of the Art

Ensuring robust OOD detection is essential for safety-critical systems such as autonomous vehicles and ADS. OOD detection involves recognizing inputs that significantly differ from the training data distribution, thereby enabling the system to respond appropriately to novel or unseen scenarios. Below, several approaches in the literature are outlined.

2.3.1 Uncertainty Estimation

By quantifying the certainty of model predictions, uncertainty-aware methods enable the detection of OOD data, allowing models to identify inputs that differ from the training distribution. This capability is essential in preventing overconfident and erroneous decisions [SS20, LEvdLL23].

Methods such as MC Dropout [GG16] and Deep Ensembles [LPB17] are commonly employed for this purpose. Prior Networks and Evidential Networks have also been effectively utilized for OOD detection by providing refined uncertainty estimations.

Other approaches have also introduced energy-based models, which replace softmax confidence scores with an energy function, enhancing OOD detection by providing a more discriminative measure of uncertainty [LWOL20]. Similarly, normalizing flows and density-based approaches have been explored to learn the distribution of ID data, enabling more effective OOD detection [CZG20].

2.3.2 Reconstruction-based Methods

Reconstruction-based methods are widely used for OOD detection, leveraging the principle that models trained to reconstruct ID data struggle to accurately reproduce unseen or OOD inputs. These approaches train deep learning models to encode and decode ID data, and during inference, the reconstruction error serves as a key signal for identifying OOD inputs. A high reconstruction error

suggests that the input deviates significantly from the training distribution, indicating a possible OOD sample [SY14, AECR24].

Early approaches to reconstruction-based OOD detection used Autoencoders (AE), which learn a compressed representation of ID data and reconstruct inputs from this latent space [SY14]. More advanced methods employ variational autoencoders (VAE), which introduce a probabilistic framework to encode inputs into a structured latent space, allowing for uncertainty-aware reconstructions and improved detection of OOD samples [KW+19, ZL23].

Moreover, β-VAEs and hierarchical VAEs have been introduced to refine reconstruction-based OOD detection by incorporating structured priors and adaptive likelihoods, helping distinguish between normal and anomalous data [RRK22, LWX+22]. Other hybrid approaches integrate adversarial training with reconstruction-based models, where generative adversarial networks (GAN) guide the latent space to better separate ID and OOD samples [CSG22].

2.3.3 Self-Supervised Learning

Self-supervised learning (SSL) has emerged as a powerful approach for learning feature representations without requiring manual labels. By leveraging auxiliary pretext tasks, self-supervised models can extract meaningful representations from unlabeled data, improving generalization and robustness. In the context of OOD detection, SSL enables models to identify inputs that deviate from the learned data manifold, making it a promising method for detecting anomalous or novel samples [HM19, JBZB20].

One early approach of SSL-based OOD detection involves rotation prediction tasks. In [GSK18], it was demonstrated that training models to recognize image rotations helps them develop internal representations that capture key structural properties of the data. Poor performance on this pretext task suggests that an input may be OOD, as it does not conform to the learned manifold. Extending this idea, contrastive learning has gained popularity as a method for distinguishing between ID and OOD data [CKNH20, HFW+20]. Contrastive learning optimizes models to group similar samples while pushing dissimilar ones apart, allowing OOD inputs to be naturally assigned lower similarity scores [MYII23].

A more unified framework, self-supervised outlier detection, was introduced by Sehwag et al. [SCM21], which integrates rotation-based losses and contrastive objectives to achieve state-of-the-art performance in OOD detection. Similarly, [Buc23] explored domain generalization via SSL, demonstrating how relation-based SSL approaches can enhance OOD detection by modeling high-level structural relationships within data.

2.3.4 Open Set Recognition

Open set recognition (OSR) techniques aim to distinguish inputs belonging to known classes from those belonging to unknown or novel classes, closely aligning with the objectives of OOD detection. Unlike standard classification, which assumes all test inputs fall within predefined categories, OSR acknowledges that real-world environments frequently introduce entirely new object categories, necessitating a model's ability to recognize and handle unknowns [GHC20, WVH24].

Early OSR methods relied on threshold-based mechanisms that modified the final model layer to reject inputs with low predicted certainty [BB16]. However, such approaches proved inadequate in high-dimensional spaces, where DNNs tend to make overconfident predictions. Bendale and Boult addressed this by introducing OpenMax, which statistically models deep features to improve rejection of unknown classes. More recent distance-based and metric-learning strategies measure a sample's distance in feature space relative to known class clusters, improving the ability to reject unknowns [LLLS18, VHVZ21].

More advanced approaches in OSR integrate deep generative models such as VAE and GAN, which learn data distributions and detect novel inputs as deviations from learned patterns. Meta-learning has been explored to enhance OSR's generalization ability across diverse tasks [SMH+21].

3 Uncertainty Estimation and OOD Identification

Uncertainty estimation provides a measure of how uncertain or confident a model is about its predictions, complementing the predicted output with a reliability score. This is essential as the models often encounter inputs outside their training datasets. The primary goal of uncertainty estimation is to provide a reliable measure of a model's output trustworthiness.

The objectives of this work are to detect and segment ID objects precisely, provide reliable uncertainty estimates, and identify OOD objects accurately. The key challenge is improving uncertainty estimation and OOD identification without affecting segmentation performance. This requires a specially designed architecture capable of all three tasks.

This chapter describes the architecture designed to achieve accurate object segmentation, uncertainty estimation, and OOD identification. It begins with a detailed discussion of the architecture, which includes a single shared backbone and multiple decoders for the tasks of semantic segmentation, instance segmentation, and depth estimation.

A key component, the Dirichlet layer, is highlighted for its role in modeling uncertainty and providing a quantifiable measure of certainty for each prediction. Additionally, the chapter introduces the intermediate layer variational inference (ILVI) module, which enhances the model's ability to generalize from training data to unseen scenes by incorporating stochastic layers.

Towards the end of the chapter, the metrics used to assess the performance of the architecture are detailed. These metrics ensure a comprehensive evaluation of the model's accuracy, reliability, and robustness in handling uncertainty and OOD identification.

3.1 Architecture

This section discusses the proposed model architecture designed for generating uncertainty estimates for semantic and instance segmentation tasks, as shown in Figure 3.1.

3.1.1 Architecture Overview

The architecture is structured around a shared backbone that performs feature extraction from the input image. The backbone's purpose is to capture the features necessary for downstream tasks, including semantic segmentation, instance segmentation and depth estimation.

Shared Backbone

The shared backbone processes input images to generate feature maps, which are subsequently utilized by the decoders. Employing a single shared backbone across the architecture, as opposed to separate backbones for each decoder, ensures that all decoders operate with identical feature maps. This unified approach maintains consistency in the features and uncertainties represented across all outputs. Allocating a backbone to each decoder is likely to induce divergence in their learned feature spaces, resulting in inconsistent outputs. This may allow artifacts overlooked by one backbone to escape detection by others, thus compromising overall consistency [VGVG+21, Cra20].

Intermediate Layer Variational Inference

The ILVI module is a central component of the architecture that introduces stochasticity into the feature extraction process. By modeling latent variables as probabilistic distributions, the ILVI module enables the generation of diverse feature representations during both training and inference. This stochastic sampling allows the model to explore a broader range of latent feature spaces, contributing to improved generalization from training data to unseen scenes. Although the ILVI module itself does not directly provide

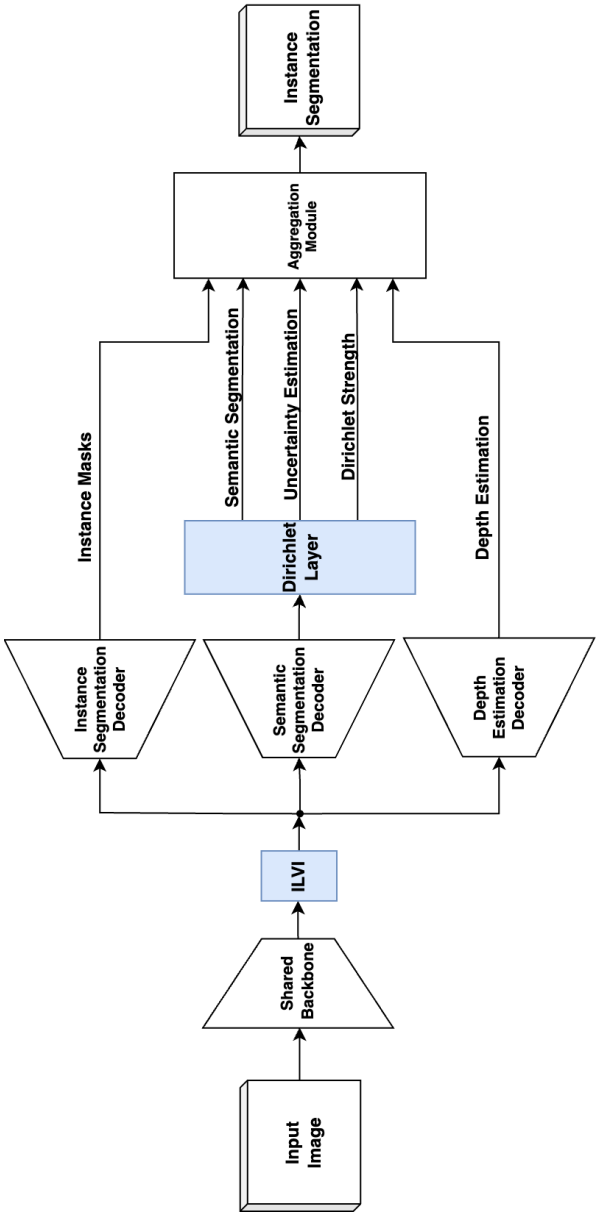


Figure 3.1: The proposed architecture performs instance segmentation, semantic segmentation, and depth estimation using a shared backbone and task-specific decoders, incorporating an ILVI module and Dirichlet layer. Adapted from Figure 2 in [HBGS23a] ©IEEE.

explicit uncertainty estimates, its stochastic outputs are utilized downstream to quantify the uncertainty in predictions.

Parallel Decoders

The architecture incorporates three parallel decoders for different tasks: semantic segmentation, instance segmentation, and depth estimation. The semantic segmentation decoder classifies each pixel in the image into one of the training classes, providing a semantic understanding of the environment. Its output is passed to the Dirichlet layer, which further classifies each pixel, estimates uncertainties, and identifies OOD pixels. The instance segmentation decoder identifies and delineates individual objects within the scene. The depth estimation decoder generates per-pixel depth estimates to determine the relative distance of objects from the vehicle.

Dirichlet Layer

The Dirichlet layer employs Dirichlet distributions to model prediction uncertainty, providing a quantifiable measure of certainty. It processes the outputs of the semantic segmentation decoder to classify each pixel, estimate uncertainty, and identify OOD pixels.

Aggregation Module

The final phase of the architecture, the aggregation module, combines outputs from all decoders. It aggregates segmented objects, their labels, depth data, and uncertainties into a unified scene representation. This module assigns classes to instance segmentation masks, calculates uncertainties, and identifies OOD objects using semantic segmentation data, and computes depth information for each mask.

3.1.2 Architecture Loss Function

Each module within the architecture contributes to the composite loss function, which is structured as follows:

$$\mathcal{L} = \mathcal{L}_{\text{ILVI}} + \mathcal{L}_{\text{Dir}} + \mathcal{L}_{\text{Depth}} + \mathcal{L}_{\text{Ins}}, \tag{3.1}$$

where the $\mathcal{L}_{\text{ILVI}}$ loss is for the ILVI layer, $\mathcal{L}_{\text{Dir}}$ refers to the Dirichlet layer loss, $\mathcal{L}_{\text{Depth}}$ is the depth decoder loss, and $\mathcal{L}_{\text{Ins}}$ is the instance segmentation loss. Each of the decoders and modules with their respective loss functions will be explained in detail in the following respective sections.

3.2 Intermediate Layer Variational Inference

This section introduces ILVI, which extends VI to the hidden layers of DNNs. By adding stochasticity to these layers, ILVI improves the model's ability to estimate uncertainties. This technique enhances the network's robustness and reliability by leveraging internal representations for better uncertainty estimation. The section covers the implementation, theoretical foundations, and impact of ILVI on model performance [HGBS21, HBGS22b].

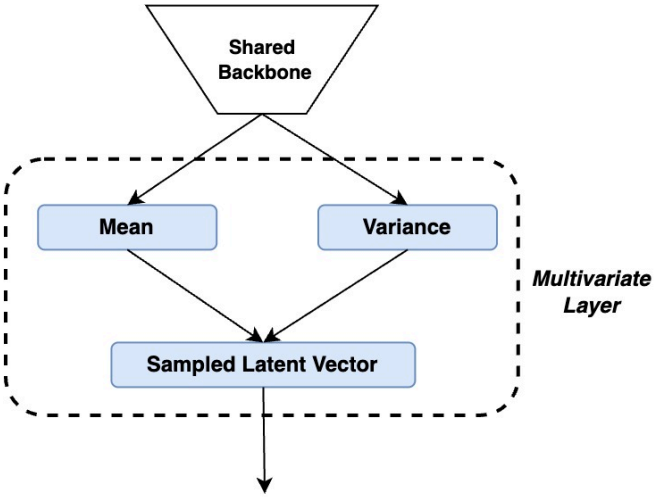


Figure 3.2: Illustration of the ILVI Layer integrated within the architecture. The shared backbone generates feature representations, which are further processed by the ILVI layer to produce a mean and variance. These are combined to sample latent variables from a multivariate distribution. The sampled latent vector serves as input for subsequent task-specific decoders. Adapted from Figure 3 in [HGBS21] ©ACM.

In ILVI, certain DNN layers are treated as multivariate entities with mean and variance, capturing stochasticity [KW13]. This controlled randomness forces diverse representations of latent features, enhancing robustness and uncertainty [MCL+21]. Parameterizing layers with these statistics and sampling directly from them improves speed and memory efficiency.

In this work, the ILVI layer is applied to the output of the shared backbone, allowing stochasticity to be introduced early in the architecture. This placement ensures that the sampled latent features are propagated to all three decoders, maintaining consistency across them. Integrating the ILVI layer here optimizes the sampling process, as the decoders have fewer layers than the shared backbone. During training, one latent sample per input is propagated through the decoders. During inference, latent samples are drawn and passed through to the decoders. The decoder outputs are then aggregated to estimate predictive uncertainty.

3.2.1 Theoretical Foundations of ILVI

To explain ILVI, an encoder–decoder network is considered as an example, where the ILVI layer corresponds to the last layer of the encoder. The primary task for the model is to classify each input image.

The dataset is defined as $X = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, consisting of N i.i.d. samples of input images x and corresponding labels y. The model predicts the labels by employing a probabilistic process that incorporates an unobserved continuous latent variable h.

In the *generative model*, a conditional prior distribution $p_\theta(h|x)$ is assumed over the latent variable, parameterized by the decoder parameters θ. Given a latent sample h, the decoder generates the prediction y from the conditional likelihood $p_\theta(y|h,x)$ [Doe16, BKM17, RMW14]. Both $p_\theta(h|x)$ and $p_\theta(y|h,x)$ are chosen from differentiable parametric families of distributions. However, the true parameters θ^* and the actual latent variables h remain unknown [KW13, RMW14].

The *posterior distribution* $p_\theta(h|x,y)$ quantifies the probability of the latent variable h given the input x and observed label y. Using Bayes' theorem:

$$p_\theta(h|x,y) = \frac{p_\theta(y|h,x)\, p_\theta(h|x)}{p_\theta(y|x)} = \frac{p_\theta(y|h,x)\, p_\theta(h|x)}{\int p_\theta(y|h,x)\, p_\theta(h|x)\, dh}, \tag{3.2}$$

where $p_\theta(y|x)$ denotes the marginal likelihood. Here, $p_\theta(y|x)$ is obtained by marginalizing the latent variable h and serves as the normalizing constant of

the posterior $p_\theta(h|x,y)$. Direct evaluation of this posterior is infeasible due to the intractable integral in the denominator [BN06, Mur12].

To overcome this intractability, *variational inference* introduces an *encoder distribution* $q_\phi(h|x,y)$, parameterized by ϕ, which serves as an approximation to the true posterior. The objective is to minimize the Kullback–Leibler (KL) divergence between the variational posterior and the intractable posterior:

$$\mathrm{KL}\big(q_\phi(h|x,y) \,\|\, p_\theta(h|x,y)\big) = \mathbb{E}_{q_\phi(h|x,y)}\left[\log \frac{q_\phi(h|x,y)}{p_\theta(h|x,y)}\right]. \tag{3.3}$$

Using Bayes' theorem to expand $p_\theta(h|x,y)$ in the KL divergence, the expression becomes:

$$\mathrm{KL}\big(q_\phi(h|x,y) \,\|\, p_\theta(h|x,y)\big) = \mathbb{E}_{q_\phi(h|x,y)}\left[\log \frac{q_\phi(h|x,y)\, p_\theta(y|x)}{p_\theta(y|h,x)\, p_\theta(h|x)}\right]. \tag{3.4}$$

Separating the terms involving the marginal likelihood and grouping the terms, the KL divergence can be expressed as:

$$\begin{aligned}\mathrm{KL}(q_\phi(h|x,y)\|p_\theta(h|x,y)) = {} & \mathbb{E}_{q_\phi(h|x,y)}\left[\log q_\phi(h|x,y)\right] \\ & + \log p_\theta(y|x) - \mathbb{E}_{q_\phi(h|x,y)}\left[\log p_\theta(y|h,x) p_\theta(h|x)\right].\end{aligned} \tag{3.5}$$

Reorganizing the terms, the KL divergence can be rewritten as:

$$\begin{aligned}\mathrm{KL}(q_\phi(h|x,y)\|p_\theta(h|x,y)) = {} & \log p_\theta(y|x) \\ & - \mathbb{E}_{q_\phi(h|x,y)}\left[\log p_\theta(y|h,x) p_\theta(h|x) - \log q_\phi(h|x,y)\right].\end{aligned} \tag{3.6}$$

This equation demonstrates the decomposition of the KL divergence, where the marginal likelihood $\log p_\theta(y|x)$ is separated, and the remaining terms are encapsulated within an expectation under the variational distribution $q_\phi(h|x,y)$.

The KL divergence in equation 3.6 is a non-negative quantity and equals zero only when the variational distribution $q_\phi(h|x, y)$ exactly matches the true posterior $p_\theta(h|x, y)$. Based on this property, the Evidence Lower Bound (ELBO) can be derived from the KL divergence between $q_\phi(h|x, y)$ and $p_\theta(h|x, y)$.

Evidence Lower Bound

The Evidence Lower Bound (ELBO) is a fundamental concept in variational inference, serving as an objective function to approximate the intractable posterior distribution $p_\theta(h|x, y)$. It represents a lower bound on the marginal log-likelihood of the observed data, $\log p_\theta(y|x)$, and is derived by reformulating the Kullback-Leibler (KL) divergence between the variational distribution $q_\phi(h|x, y)$ and the true posterior. By maximizing the ELBO, the model ensures that the variational approximation $q_\phi(h|x, y)$ is a close match to the true posterior $p_\theta(h|x, y)$, while simultaneously optimizing the parameters θ of the model.

The KL divergence is non-negative, as shown in the inequality below, which forms the foundation for defining the ELBO as a lower bound on the marginal log-likelihood of the observed data:

$$\begin{gathered} \mathrm{KL}(q_\phi(h|x, y) \| p_\theta(h|x, y)) \geq 0 \\ \log p_\theta(y|x) - \mathbb{E}_{q_\phi(h|x,y)} \left[\log p_\theta(y|h, x) p_\theta(h|x) - \log q_\phi(h|x, y) \right] \geq 0 \end{gathered} \tag{3.7}$$

From this inequality, the log marginal likelihood $\log p_\theta(y|x)$ can be expressed as:

$$\log p_\theta(y|x) \geq \mathbb{E}_{q_\phi(h|x,y)} \left[\log p_\theta(y|h, x) p_\theta(h|x) - \log q_\phi(h|x, y) \right]. \tag{3.8}$$

Rewriting the terms:

$$\begin{aligned} \log p_\theta(y|x) \geq\ & \mathbb{E}_{q_\phi(h|x,y)} \left[\log p_\theta(y|h, x) \right] \\ & + \mathbb{E}_{q_\phi(h|x,y)} \left[\log p_\theta(h|x) \right] - \mathbb{E}_{q_\phi(h|x,y)} \left[\log q_\phi(h|x, y) \right] \end{aligned} \tag{3.9}$$

where,

$$\mathbb{E}_{q_\phi(h|x,y)}\left[\log p_\theta(h|x)\right] - \mathbb{E}_{q_\phi(h|x,y)}\left[\log q_\phi(h|x,y)\right] = -\mathrm{KL}(q_\phi(h|x,y)\|p_\theta(h|x)). \tag{3.10}$$

Substituting this result:

$$\log p_\theta(y|x) \geq \mathbb{E}_{q_\phi(h|x,y)}\left[\log p_\theta(y|h,x)\right] - \mathrm{KL}(q_\phi(h|x,y)\|p_\theta(h|x)). \tag{3.11}$$

The resulting expression is defined as the ELBO:

$$\mathrm{ELBO}(\theta,\phi) = \mathbb{E}_{q_\phi(h|x,y)}\left[\log p_\theta(y|h,x)\right] - \mathrm{KL}(q_\phi(h|x,y)\|p_\theta(h|x)). \tag{3.12}$$

The objective is to maximize the ELBO with respect to both the variational parameters ϕ and the model parameters θ. By maximizing the ELBO, the KL divergence between $q_\phi(h|x,y)$ and $p_\theta(h|x,y)$ is minimized, ensuring that $q_\phi(h|x,y)$ serves as an approximation to $p_\theta(h|x,y)$ [BKM17].

The two components of the ELBO serve distinct purposes in model training. The expected log-likelihood term $\mathbb{E}_{q_\phi(h|x,y)}\left[\log p_\theta(y|h,x)\right]$ represents the model's task of predicting the observed labels y based on the input x and the latent representation h. This term quantifies the model's fit to the observed data. In contrast, the KL divergence term $\mathrm{KL}(q_\phi(h|x,y)\|p_\theta(h|x))$ acts as a regularizer by limiting the divergence of the variational posterior $q_\phi(h|x,y)$ from the prior $p_\theta(h|x)$. This regularization mitigates overfitting and enforces structure on the latent space.

KL Divergence for Gaussian Latent Variables

If the latent variable h follows a Gaussian distribution, the KL term in the variational lower bound in equation 3.12 can be expressed in closed form, as outlined in [KW13, KW+19]. The KL divergence for Gaussian latent variables allows for direct integration under specific assumptions, thereby simplifying the variational inference process. Specifically, assume that h is sampled from a Gaussian distribution $\mathcal{N}(\mu,\sigma^2)$, and the prior $p_\theta(h|x)$ is a standard normal Gaussian $\mathcal{N}(0,\mathbf{I})$. This approach facilitates a closed-form solution for the KL divergence term in the ELBO.

This formulation assumes that both the prior $p_\theta(h|x) = \mathcal{N}(0, I)$ and the approximate posterior $q_\phi(h|x, y)$ are multivariate Gaussian distributions. The dimensionality of the latent variable h is denoted by J, where μ and σ represent the variational mean and standard deviation, respectively, evaluated for each data point. The summation over J ensures that contributions from all dimensions of h are accounted for in the KL divergence computation.

The KL divergence between the approximate posterior $q_\phi(h|x, y)$ and the prior $p_\theta(h|x)$ in equation 3.12 is defined as:

$$\mathrm{KL}(q_\phi(h|x, y) \| p_\theta(h|x)) = \int q_\phi(h|x, y) \log \frac{q_\phi(h|x, y)}{p_\theta(h|x)} \, dh \tag{3.13}$$

Expanding the logarithm:

$$\mathrm{KL}(q_\phi(h|x, y) \| p_\theta(h|x)) = \int q_\phi(h|x, y) \left[\log q_\phi(h|x, y) - \log p_\theta(h|x)\right] \, dh \tag{3.14}$$

This results in two separate integrals:

$$\begin{aligned} \mathrm{KL}(q_\phi(h|x, y) \| p_\theta(h|x)) = &\int q_\phi(h|x, y) \log q_\phi(h|x, y) \, dh \\ &- \int q_\phi(h|x, y) \log p_\theta(h|x) \, dh \end{aligned} \tag{3.15}$$

The two terms in Equation 3.15 will be computed separately, starting with the second (entropy) term involving the log of the approximate posterior.

For the first term in equation 3.15, using the same posterior approximation, the log-probability is:

$$\log q_\phi(h|x, y) = -\frac{1}{2} \left[J \log(2\pi) + \log \det(\Sigma) + (h - \mu)^T \Sigma^{-1} (h - \mu)\right] \tag{3.16}$$

Here, $\Sigma = \text{diag}(\sigma^2)$, and for a diagonal covariance:

$$\log \det(\Sigma) = \sum_{j=1}^{J} \log(\sigma_j^2) \tag{3.17}$$

The expectation $\mathbb{E}_{q_\phi}\left[(h-\mu)^T\Sigma^{-1}(h-\mu)\right]$ simplifies to J, as $q_\phi(h|x,y)$ is centered around its mean. Substituting this, the second term evaluates as:

$$\int q_\phi(h|x,y) \log q_\phi(h|x,y)\, dh = -\frac{1}{2}J\log(2\pi) - \frac{1}{2}\sum_{j=1}^{J}\log(\sigma_j^2) - \frac{1}{2}J \tag{3.18}$$

In the second integral, assume that $q_\phi(h|x,y) = \mathcal{N}(h;\mu,\text{diag}(\sigma^2))$, and that the prior $p_\theta(h|x)$ is $\mathcal{N}(h;0,I)$. The prior $p_\theta(h|x)$ is characterized by a zero mean vector and a unit covariance matrix. Its log-probability is expressed as:

$$\log p_\theta(h|x) = -\frac{1}{2}\left[J\log(2\pi) + h^T h\right] \tag{3.19}$$

Taking the expectation over $q_\phi(h|x,y)$, the first term becomes:

$$\int q_\phi(h|x,y) \log p_\theta(h|x)\, dh = -\frac{1}{2}J\log(2\pi) - \frac{1}{2}\mathbb{E}_{q_\phi(h|x,y)}\left[h^T h\right] \tag{3.20}$$

For a Gaussian $q_\phi(h|x,y)$ with mean μ and diagonal covariance $\text{diag}(\sigma^2)$, the expectation $\mathbb{E}_{q_\phi}\left[h^T h\right]$ simplifies to:

$$\mathbb{E}_{q_\phi}\left[h^T h\right] = \sum_{j=1}^{J}(\mu_j^2 + \sigma_j^2) \tag{3.21}$$

Substituting this result:

$$\int q_\phi(h|x,y) \log p_\theta(h|x)\, dh = -\frac{1}{2} J \log(2\pi) - \frac{1}{2} \sum_{j=1}^{J} (\mu_j^2 + \sigma_j^2) \tag{3.22}$$

Combining equations 3.18 and 3.22 , the KL divergence in 3.15 becomes:

$$\begin{aligned} \mathrm{KL}\big(q_\phi(h|x,y) \parallel p_\theta(h|x)\big) = &\left[-\frac{1}{2} J \log(2\pi) - \frac{1}{2} \sum_{j=1}^{J} \log(\sigma_j^2) - \frac{1}{2} J \right] \\ &- \left[-\frac{1}{2} J \log(2\pi) - \frac{1}{2} \sum_{j=1}^{J} (\mu_j^2 + \sigma_j^2) \right] \end{aligned} \tag{3.23}$$

Simplifying this further:

$$\mathrm{KL}(q_\phi(h|x,y) \| p_\theta(h|x)) = \frac{1}{2} \sum_{j=1}^{J} \left(\mu_j^2 + \sigma_j^2 - \log(\sigma_j^2) - 1 \right) \tag{3.24}$$

Reparameterization Trick

Variational inference aims to optimize the ELBO by maximizing the expectation term:

$$\mathbb{E}_{q_\phi(h|x,y)} \left[\log p_\theta(y|h,x) \right].$$

While $q_\phi(h|x,y)$ is parameterized by the learnable variational parameters ϕ, computing the gradient of the expectation term $\mathbb{E}_{q_\phi(h|x,y)} \left[\log p_\theta(y|h,x) \right]$ with respect to ϕ poses a challenge.

This difficulty arises due to the stochastic nature of the sampling process required to compute the expectation. Since the latent variable h is sampled from $q_\phi(h|x,y)$, the sampling operation introduces a non-differentiable component.

Consequently, direct backpropagation through the sampling step is not feasible, complicating the computation of gradients with respect to ϕ.

The reparameterization trick, introduced by [KW13], resolves this issue by reformulating the sampling process to make it differentiable. Instead of directly sampling h from $q_\phi(h|x, y)$, the latent variable is expressed as a deterministic transformation of ϕ and a noise variable ϵ, sampled from a standard normal distribution:

$$h = \mu + \sigma \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{3.25}$$

where μ represents the mean vector, parameterized by ϕ and predicted by the encoder network, σ is the standard deviation vector, also parameterized by ϕ and predicted by the encoder network, and ϵ is a noise vector, independently sampled from the standard normal distribution $\mathcal{N}(0, I)$.

This reformulation decouples the stochasticity, introduced by ϵ, from the learnable parameters ϕ. As a result, gradients can propagate through μ and σ, enabling the efficient computation of gradients and the optimization of the ELBO via backpropagation.

ILVI Loss

Referring back to equation 3.1 and relating it to equation 3.12 and equation 3.24, the ILVI loss L_{ILVI} is defined as

$$L_{\text{ILVI}} = \text{KL}\big(q_\phi(h|x, y) \parallel p_\theta(h|x)\big)\,. \tag{3.26}$$

By minimizing L_{ILVI}, the model aligns the approximate posterior $q_\phi(h|x, y)$ with the prior $p_\theta(h|x)$. This alignment introduces regularization that mitigates overfitting and encourages structured exploration of the latent feature space [HBWP13, YZ18]. The resulting latent representations remain robust and generalizable.

The second component of the ELBO, $\mathbb{E}_{q_\phi(h|x,y)}[\log p_\theta(y|h, x)]$, quantifies the expected log-likelihood of the observed label y given the latent variable h and the input x. In practice, optimization is performed by minimizing the

negative ELBO; therefore, the task-specific decoder objectives correspond to the negative expected log-likelihood:

$$-\mathbb{E}_{q_\phi(h|x,y)}[\log p_\theta(y|h,x)] = \mathcal{L}_{\text{Dir}} + \mathcal{L}_{\text{Depth}} + \mathcal{L}_{\text{Ins}}. \tag{3.27}$$

3.3 Decoders

This section details the design and functionality of the decoders in the architecture. The three decoders - instance segmentation, semantic segmentation, and depth estimation - process features extracted by the shared backbone and ILVI. Each decoder's architecture and associated loss functions are described.

3.3.1 Instance Segmentation Decoder

Instance segmentation is performed by the instance decoder, which generates masks outlining each object in an image. It operates on features from the shared backbone. The decoder's structure refines these features to capture instance-specific information, distinguishing unique boundaries and characteristics of each object, even within the same semantic class.

The decoder's output consists of binary masks, each of size HxW, representing different object instances. These masks include only the pixels of the corresponding object, distinctly separating it from other objects and the background.

The training process compares the instance decoder's mask predictions with ground-truth masks using a loss function to quantify differences. This guides the decoder to produce accurate masks. Simultaneous training of the semantic segmentation and instance decoders ensures that the instance decoder captures fine details while incorporating semantic context. This integration enhances the model's ability to categorize and segment individual object instances, improving the accuracy and effectiveness of instance segmentation [JHP+22].

3.3.2 Semantic Segmentation Decoder

The semantic segmentation decoder and the Dirichlet layer work together to classify each pixel into predefined classes, creating a segmentation map. They process features from the shared backbone and ILVI to generate a per-pixel classification map.

Together, they estimate uncertainty and identify OOD pixels. The Dirichlet layer generates three per-pixel maps: semantic segmentation, uncertainty estimation, and OOD identification.

The loss function for the semantic segmentation decoder is based on training the Dirichlet distributions in the Dirichlet layer. This will be further explained in section 3.4.

3.3.3 Depth Estimation

Depth estimation is performed at a per-pixel level, focusing on objects such as vehicles, pedestrians, and other relevant items in the scene. The depth decoder provides depth estimates for each identified object, concentrating exclusively on object pixels while excluding elements like the road and sky. This object-centric approach enhances detected instances with depth information alongside class and OOD estimates, allowing the model to focus on the most critical elements in the driving environment.

Depth estimation is approached as a regression task, aiming to predict continuous depth values for each pixel.

The Smooth L1 loss, also known as Huber Loss [Hub64], is used for training the regression task. This loss blends the characteristics of L1 and L2 losses to mitigate sensitivity to outliers [RHGS15]. It is defined as follows:

$$L_{\text{smooth}}(d, \hat{d}) = \begin{cases} 0.5(d - \hat{d})^2 & \text{if } |d - \hat{d}| < \delta, \\ \delta|d - \hat{d}| - 0.5\delta^2 & \text{otherwise,} \end{cases} \tag{3.28}$$

where d denotes the ground truth value, $\hat{d}$ denotes the predicted value, δ is a hyperparameter that delineates the threshold between adopting a squared error

(for differences smaller than δ) and a linear error (for differences larger than δ).

Smooth L1 is advantageous over traditional L1 and L2 losses due to its combined robustness to outliers, gradient stability, differentiability, and balanced error penalization [Bar19, FLR+19].

3.3.4 Architecture Variants

To ensure the architecture's versatility, this work examines two structure variations to assess the proposed architecture. The first variant prioritizes speed and efficiency with a smaller structure, while the second targets higher segmentation accuracy with a larger design.

The first employs MobileNetV3 [HSC+19] as the shared backbone, with semantic and instance segmentation decoders adapted from Panoptic Deeplab [CCZ+20]. The second uses EfficientNet [TL19] for its backbone, incorporating decoder structure from Efficient Panoptic Segmentation [MV20].

The loss function for the instance segmentation is adapted to each variant. Subsequent sections will explore the two decoder structure variants used and the instance segmentation loss for each variant.

Variant 1: MobileNet as a Backbone

The first variant incorporates the ASPP within each decoder [CZP+18]. The ASPP module enables the network to understand objects at various scales without losing resolution. Each decoder refines the outputs of the ILVI module then decoded to their respective task.

The instance segmentation loss in this variant can be expressed as:

$$\mathcal{L}_{\text{Ins}} = \mathcal{L}_{\text{center}} + \mathcal{L}_{\text{Masks}}, \tag{3.29}$$

where $\mathcal{L}_{\text{center}}$ focuses on minimizing the distance between predicted and actual center points of instances using mean square error as a loss. The $\mathcal{L}_{\text{Masks}}$ aims to match the predicted masks with the ground truth using L1 loss, ensuring precise object boundary delineation [CZP+18].

Variant 2: EfficientNet as a Backbone

The EfficientNet variant of the architecture utilizes the EfficientNet as the shared backbone, complemented by a 2-way Feature Pyramid Network (FPN) [LDG+17].

The semantic segmentation and depth estimation decoders process the multi-scale feature maps from the FPN to produce dense per-pixel classification and depth estimate maps respectively.

The instance segmentation head is inspired by the Mask R-CNN architecture. It consists of the Region Proposal Network (RPN) and the RoI Align module. The RPN generates region proposals, identifying potential areas in the image where objects might be located using anchor boxes to predict object boundaries and scores. The RoI Align module refines these proposals by extracting fixed-size feature maps and applying convolutional layers to predict instance masks, class labels, and bounding boxes [HGDG17].

The instance segmentation loss function encompasses five components following [MV20], formulated as:

$$\mathcal{L}_{\text{Ins}} = \mathcal{L}_{\text{OS}} + \mathcal{L}_{\text{OP}} + \mathcal{L}_{\text{Cls}} + \mathcal{L}_{\text{Bbx}} + \mathcal{L}_{\text{Mask}}, \tag{3.30}$$

where $\mathcal{L}_{\text{OS}}$ denotes the objectness score loss, $\mathcal{L}_{\text{OP}}$ the object proposal loss, $\mathcal{L}_{\text{Cls}}$ the classification loss, $\mathcal{L}_{\text{Bbx}}$ the bounding box loss, and $\mathcal{L}_{\text{Mask}}$ the mask segmentation loss.

3.4 Dirichlet Layer

Accurately estimating uncertainty in deep learning remains challenging. Techniques like Monte Carlo dropout [GG16] and deep ensembles [LPB17] provide methods for assessing model uncertainty. These methods are computationally intensive, requiring significant memory and processing time. This becomes problematic in real-time applications, where both efficiency and accuracy are essential [SZSS21, BSD23, GDS20].

A limitation in many classification neural networks is the use of the softmax function, which, despite its widespread adoption for producing probabil-

ity distributions, often fails to capture model uncertainty accurately [PBZ21, WLC+21]

To overcome the limitations of softmax and high computational demand needed, the Dirichlet layer is introduced in this architecture. Dirichlet-based neural networks offer a more refined approach for managing prediction uncertainty, enhancing the reliability of model predictions [HBGS22a, UHF21, KCZ+21].

Additionally, the Dirichlet layer enhances the DNN's ability for OOD detection, vital for identifying inputs outside the training dataset and reducing risks from unfamiliar scenarios. It provides a measurable uncertainty level, allowing the model to determine if an input deviates from known data [CTS+21, WZH+22].

This section discusses the Dirichlet layer's role in uncertainty estimation and OOD identification. It starts with a study comparing Dirichlet distributions with the softmax function, using handwritten digits as an example. This study emphasizes the advantages of Dirichlet distributions over traditional methods. It details the implementation and functionality of the Dirichlet layer within the architecture, focusing on its impact on uncertainty estimation and OOD detection. Additional information on the Dirichlet layer, along with the loss function, is provided later in the section.

3.4.1 Handwritten Digits Experiment

This subsection presents a study comparing the performance of using the last decoder layer of the DNN as the traditional softmax versus using Dirichlet distributions for uncertainty estimation and OOD input detection. The experiment uses the MNIST dataset of handwritten digits for a classification task, as shown in Figure 3.4 [LBBH98].

The architecture consists of two convolutional layers followed by two fully connected layers. Two identical neural network architectures are employed, differing only in their final layers: one uses a softmax activation function to output class probabilities, while the other generates Dirichlet concentration parameters (α). The study analyzes the last layer outputs, softmax probabilities and Dirichlet concentrations, to compare how each model represents uncertainty.

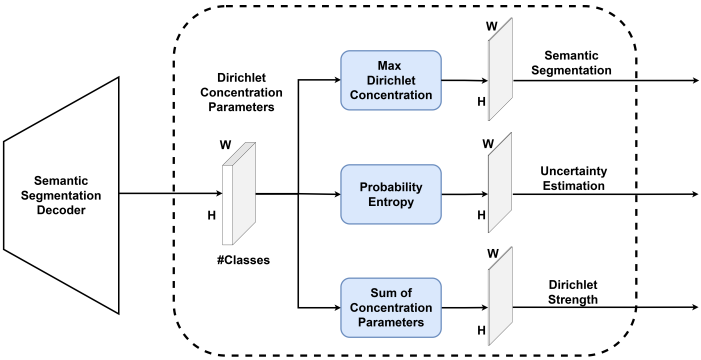


Figure 3.3: Illustration of the Dirichlet layer. The output of the semantic segmentation decoder is passed to the Dirichlet layer, where it is modeled as Dirichlet concentration parameters. These parameters define a distribution over class probabilities for each pixel in the input image. From the Dirichlet concentration parameters the outputs are generated: semantic segmentation predictions derived from the maximum concentration, uncertainty estimates computed using probability entropy, and Dirichlet strength calculated as the sum of the concentration parameters. Adapted from Figure 2 in [HBGS23a]©IEEE

.

Figure 3.4: MNIST Dataset

For this study, the DNNs are intentionally trained to classify only 9 out of the 10 digits from the MNIST dataset, deliberately excluding the '0' digit. This modification creates a framework for testing OOD detection capabilities. By withholding one digit from the training process, the network encounters unseen data during testing. This modification creates an experimental setup for testing OOD detection capabilities.

Both DNN configurations are optimized to achieve more than 99% classification accuracy on its trained classes. This optimization is crucial, as it ensures that improvements in uncertainty estimation or OOD detection do not compromise classification accuracy. The goal is to refine uncertainty estimation while maintaining high performance in the primary classification task.

Three results will be shown for each DNN; the last layer's distribution for correct predictions, incorrect predictions and OOD predictions. It is important to view the results of these distributions as these are the key indicators for distinguishing between correct, incorrect and OOD predictions.

Figure 3.5 shows the last layer outputs of both DNNs, comparing the average distributions for correct ID, incorrect ID, and OOD predictions for both softmax and Dirichlet distributions. The probabilities are sorted from highest to lowest. Ideally, for correct ID, the highest probability class should have a high probability, while all other classes should be close to zero. For incorrect ID and OOD predictions, the probabilities should be evenly distributed, indicating uncertainty.

Softmax performs well for correct ID predictions, generating a high probability for the correct class, but it fails for incorrect ID and OOD predictions by also producing a very high probability for one class. In contrast, Dirichlet distributions not only perform similarly for correct ID predictions but also yield more balanced probabilities for incorrect ID and OOD predictions, indicating better uncertainty representation. This demonstrates how Dirichlet distributions provide a more robust method for handling uncertainty and OOD detection compared to softmax, while maintaining high certainty on correct predictions.

Predictive entropy will be used in this study as a measure of certainty, where low predictive entropy indicates high certainty and high predictive entropy indicates low certainty. It calculates the average certainty for correct, incorrect, and OOD detections. Low predictive entropy means the model is confident, assigning a high probability to the true class. High predictive entropy for incorrect and OOD detections shows uncertainty, with probabilities spread across multiple classes. [GBC16]

Table 3.1: Average certainty for each distribution.

	Correct ID Predictions (↑)	Incorrect ID Predictions (↓)	OOD Predictions (↓)
Softmax	**98.1%**	73.2%	69.8%
Dirichlet Distributions	91.6%	**4.6%**	**4.2%**

Table 3.1 presents the average certainty for each distribution using predictive entropy. The softmax model achieves a high certainty level of 98.1% for

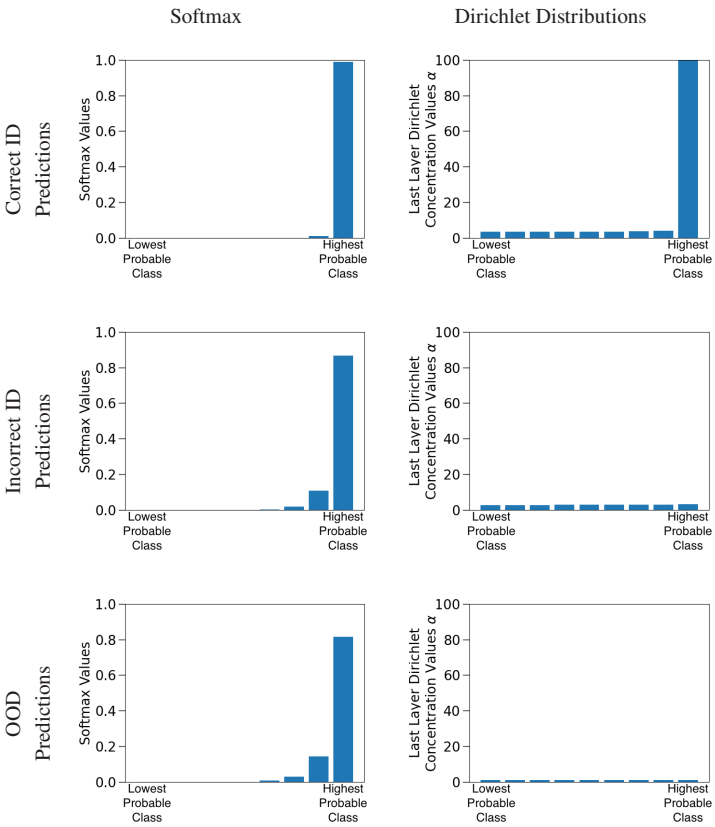


Figure 3.5: Comparison of Softmax and Dirichlet Distributions. This figure visualizes the final-layer outputs of deep neural networks for correct ID, incorrect ID, and OOD predictions. The softmax distribution exhibits overconfident outputs, even for incorrect and OOD cases, whereas the Dirichlet distribution yields a more calibrated representation that better reflects predictive uncertainty across these scenarios.

correct predictions but shows less distinction between correct and incorrect predictions, with a relatively high certainty of 73.2% and 69.8% for incorrect ID predictions and OOD predictions respectively. This supports findings that softmax often delivers overconfident predictions, especially for incorrect ID and OOD inputs [OFR+19].

In contrast, the Dirichlet model demonstrates a clear separation between correct ID predictions and the other two predictions, maintaining a lower certainty level of less than 5% for both incorrect ID and OOD predictions. Although improvements in uncertainty representation can sometimes reduce the certainty of correct predictions, the Dirichlet model still achieves over 90% certainty for correct ID predictions. This distinction underscores the Dirichlet model's superior ability to represent uncertainty, providing a clearer differentiation between correct, incorrect, and OOD predictions.

Table 3.2: Difference between the distributions using the Wasserstein metric.

	Correct ID vs. Incorrect ID (↑)	Correct ID vs. OOD (↑)	Incorrect ID vs. OOD (↑)
Softmax	3.74	2.41	1.65
Dirichlet Distributions	**10.86**	**11.06**	**5.22**

The Wasserstein distance metric is used to compare the output distributions. For comparability, the Dirichlet distributions will be normalized to a range of 0 to 1, matching the range of softmax probabilities. Table 3.2 shows the Wasserstein distance between each pair of distributions for each DNN.

The Wasserstein distance metric further emphasizes these differences between the distributions. The softmax model shows low separation between prediction types with distances of 3.74 (correct ID vs. incorrect ID), 2.41 (correct ID vs. OOD), and 1.65 (incorrect ID vs. OOD). The Dirichlet model, however, exhibits greater separation with distances of 10.86, 11.06, and 5.22, respectively.

Table 3.3: Dirichlet strength for the Dirichlet distributions' predictions.

	ID Classes	OOD Classes
Dirichlet Distributions	145.3	12.7

While Tables 3.1 and 3.2 highlight the improvements offered by Dirichlet-based modeling, they also indicate that the distinction between incorrect ID and OOD predictions remains relatively small. To address this limitation, Dirichlet strength is introduced as an additional metric to better separate OOD inputs from incorrect ID predictions. Table 3.3 presents the Dirichlet strength values used to assess the OOD detection performance of the Dirichlet-based

model. The results demonstrate a clear gap between ID and OOD classes: a high Dirichlet strength value of 145.3 reflects strong certainty in the ID predictions, while a low value of 12.7 corresponds to increased uncertainty for OOD inputs.

These findings highlight the Dirichlet distribution's superiority in representing uncertainty and distinguishing between correct, incorrect, and OOD predictions. And most importantly, Dirichlet distributions effectively identify OOD inputs unlike softmax.

The results demonstrate the benefits of using Dirichlet distributions, supporting their implementation for more complex tasks.

3.4.2 Dirichlet Distributions for Multi-Class Classification

The Dirichlet distribution is a multivariate continuous probability distribution used primarily in Bayesian statistics to model the probabilities of multiple outcomes, characterized by a set of parameters that influence its shape [BN06]. Considering a multi-class classification problem with k classes, the goal is to assign an input $\boldsymbol{x}$ to one of the k classes.

Given the probability simplex as $\mathcal{S} = \{(y_1, \ldots, y_k) : y_i \geq 0, \sum_i y_i = 1\}$, the Dirichlet distribution is a probability density function on vectors $y \in \mathcal{S}$ and categorized by concentration parameters $\alpha = \{\alpha_1, \ldots, \alpha_K\}$ as:

$$\mathrm{Dir}(y; \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} y_i^{\alpha_i - 1} \tag{3.31}$$

where the normalizing constant $\frac{1}{B(\alpha)}$ denotes the multivariate Beta function $B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}$, $\alpha_0 = \sum_{i=1}^{K} \alpha_i$ and Gamma function $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, \mathrm{d}t$, and y denotes the ground truth probability distribution [Min00].

To model the Dirichlet distribution within this architecture, the concentration parameters α are directly derived from the outputs of the semantic segmentation decoder.

3.4.3 Maximum Likelihood Estimation

To train the Dirichlet distributions, maximum likelihood estimation (MLE) is used to estimate the concentration parameters α of the Dirichlet distribution. This is achieved by maximizing the log-likelihood (equivalently, minimizing the negative log-likelihood) [Min00] as follows:

$$F(\alpha; y) = \log \prod \mathrm{Dir}(y; \alpha) = \log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \log y_i. \quad (3.32)$$

where y denotes the observed target probability vector (given by the ground-truth label distribution), and the optimization is performed with respect to α with y held fixed.

The goal is to train the DNN to generate reliable uncertainty estimations by treating correct and incorrect predictions separately. In other words, the model should produce low uncertainty (a strongly peaked distribution) for correctly classified instances, while assigning higher uncertainty (a more uniform distribution) to misclassifications.

To ensure low uncertainty for correct predictions, the DNN should exhibit a strong concentration toward the correct class, as shown in Figure 3.6a. This can be achieved by maximizing the likelihood using the ground truth labels and employing a *one-hot vector*. Conversely, for incorrect predictions, high uncertainty is achieved by maximizing the likelihood function with equal probabilities assigned to all classes, as shown in Figure 3.6b.

To address these points, an extended formulation of equation 3.32 for the semantic segmentation is presented as follows:

$$\mathcal{L}_{\mathrm{Dir}} = F(\alpha_{\mathrm{correct}}; y_{\mathrm{correct}}) + F(\alpha_{\mathrm{incorrect}}; y_{\mathrm{incorrect}}), \quad (3.33)$$

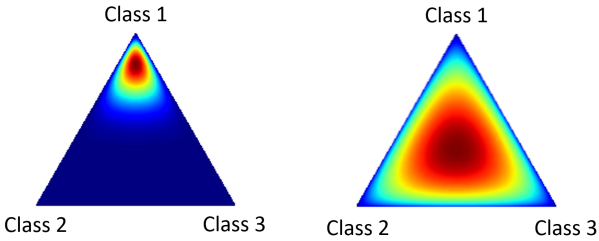
where $\alpha_{\mathrm{correct}}$ and $\alpha_{\mathrm{incorrect}}$ are the network's concentration parameters representing the correct and incorrect DNN predictions respectively, and y_{correct} and $y_{\mathrm{incorrect}}$ represent the ground truth labels for the correct classes and the equal probability vector to yield high uncertainty, respectively.

To mitigate the issue of overconfidence caused by the imbalanced class distribution in the training dataset, an additional weighting factor λ_{weight} is introduced, which is multiplied by the semantic segmentation loss $\mathcal{L}_{\text{Dir}}$ (equation 3.33). This weighting factor is computed as $\lambda_{\text{weight}_c} = 1 - \frac{N_c}{\sum N}$, where N represents the total pixel count and N_c represents the pixel count for class c.

3.4.4 Uncertainty Estimation

The Dirichlet distribution defines a distribution over the probability simplex, enabling the modeling of uncertainty in multi-class classification tasks. Unlike the softmax function, which converts raw logits into a single point estimate of class probabilities, the Dirichlet distribution captures the model's belief over possible class distributions, providing a more expressive representation of the model's certainty levels [HKH22]. The softmax formula is as follows:

$$\text{Softmax}(z_c) = p(y = c|z) = \frac{\exp(z_c)}{\sum_{i=1}^{K} \exp(z_i)}. \tag{3.34}$$



(a) Certain prediction. (b) Uncertain prediction.

Figure 3.6: Visualization of Dirichlet distributions for multi-class classification in deep neural networks, illustrating the distinction between certain and uncertain predictions. Subfigure (a) shows a concentrated Dirichlet distribution with a sharp peak at the correct class (Class 1), indicated by warmer colors (red and yellow), signifying high certainty. Subfigure (b) displays a uniform Dirichlet distribution, with cooler colors (blue and green) reflecting an even spread of probabilities across all classes and thus high uncertainty. Color intensity represents probability density: warmer regions indicate higher density, while cooler regions correspond to lower density. Adapted from Figure 1 in [HBGS22a] ©ACM.

where z_c is the logit for class c. Softmax converts raw prediction scores (logits) into probabilities by exponentiating and normalizing them. This amplifies minor differences between logits, leading to overconfident predictions. Consequently, softmax can produce skewed probabilities among classes, where small input changes yield significantly different output probabilities. This makes it prone to misrepresenting the model's true certainty [PBZ21, SKK18].

In contrast, the concentration parameters of the Dirichlet distribution directly affect the shape and variability of the probability distribution across classes. This ensures that the output probabilities are true realizations from a distribution, inherently capturing both the uncertainty and the correlations between classes. This is different to the softmax function, which computes probabilities based on the relative magnitudes of logits, failing to provide any insight into the underlying uncertainties or the relationships between different classes [Tsi21].

In this context, predictive entropy is utilized as the primary metric for uncertainty estimation. Predictive entropy is widely used in semantic segmentation models as it quantifies uncertainty by analyzing the variability within the probability distributions of predicted classes [GTA+21]. Predictive entropy is calculated as follows,

$$\hat{\mathbb{H}}[y|x] = -\sum_{c}^{K}(p(y = c|x))\log(p(y = c|x)) \tag{3.35}$$

where y is the output variable, $c \in \{1, \ldots, K\}$ indexes the K classes, and $p(y = c|x)$ denotes the predictive probability that the input pixel x belongs to class c. and $p(y = c \mid x)$ denotes the predictive probability that the input pixel x belongs to class c. For the Dirichlet formulation, the network predicts concentration parameters $\alpha(x)$ that define a distribution over class-probability vectors π on the simplex, $\pi \mid x \sim \mathrm{Dir}(\alpha(x))$. The predictive class probability is then given by the posterior mean of π,

$$p(y = c \mid x) \;=\; \mathbb{E}[\pi_c \mid x] \;=\; \frac{\alpha_c}{\alpha_0}, \qquad \alpha_0 = \sum_{k=1}^{K} \alpha_k.$$

A higher predictive entropy indicates greater uncertainty, signifying a probability distribution spread across multiple outcomes. Conversely, a lower pre-

dictive entropy suggests higher certainty, with a more concentrated distribution on a specific outcome [GDS20, GG16].

3.4.5 OOD Identification Using Dirichlet Strength

An important aspect of improving model reliability and safety is the ability to accurately identify OOD inputs that are outside the training distribution. DNNs often struggle to differentiate between ID and OOD instances. This leads to the DNN predicting incorrect predictions with high confidence [RLZ+22, HMD18].

This work uses Dirichlet strength, denoted as α_0, to differentiate between ID and OOD instances.

$$\alpha_0 = \sum \alpha \tag{3.36}$$

High Dirichlet strength values signify ID instances, while low values indicate OOD instances, reflecting uncertainty and potential deviation from the training data. Leveraging the Dirichlet distribution's properties, this approach aims to enhance the model's ability to assess and quantify uncertainty, providing a clearer distinction between known and unknown data [HBGS23a, CTS+21].

The dual use of predictive entropy and Dirichlet strength enhances the model's certainty interpretation by providing complementary measures to assess prediction reliability. Predictive entropy quantifies the inherent uncertainty in the model's output, while Dirichlet strength captures the concentration of belief relative to the learned data distribution. Together, these measures offer a more comprehensive assessment of the model's predictions, facilitating a clearer distinction between ID and OOD instances.

The following are two common cases:

1. *ID Data with High Uncertainty and High Dirichlet Strength*

   For in-domain data, the model may exhibit high predictive uncertainty while maintaining a high Dirichlet strength. This scenario indicates that the input is recognized as belonging to the training distribution, yet the model struggles to assign it to a specific class with confidence. A high

sum of concentration parameters reflects a strong belief that the input lies within the known categories. However, elevated predictive entropy reveals ambiguity in class assignment, often due to overlapping features or class boundaries. This highlights that strong certainty about the data's domain does not necessarily translate into accurate classification [XLZL23].

2. *OOD Data with Low Dirichlet Strength*

   OOD data differ from the training set and often contain features unfamiliar to the model. These features do not align with the learned distributions, making it challenging for the model to associate them with any trained category. This results in low Dirichlet strength, which serves as an important indicator for identifying inputs as OOD. Low Dirichlet strength signifies the presence of features that are dissimilar to the training data [CLT+21].

It is important to note that during the training phase of the model, no OOD data were included. The model was trained exclusively on ID data, without any exposure to or labeling of OOD examples. Thus, the DNN's ability to identify and differentiate OOD inputs is derived solely from its training on known, ID datasets. This approach underscores the robustness and adaptability of the proposed architecture, relying purely on the generalization capabilities of the DNN trained on ID data to handle entirely new and unseen inputs during operational deployment. This methodological choice enhances the clarity of testing the model's performance in real-world scenarios, where it must make predictions on data significantly different from those it has encountered during training.

3.5 Aggregation Module

The aggregation module is the final component of the architecture. It consolidates outputs from preceding decoders. This module provides a comprehensive analysis of each detected instance. It incorporates all outputs and calculates the following attributes for each instance: instance mask, instance class, mean distance, mean uncertainty estimate, and mean Dirichlet strength. The mod-

ule processes each instance identified by the instance decoder upon receiving inputs from the decoders.

1. Instance Mask: This is a binary representation for each detected object in the image, with shape (N, H, W), where N represents the number of objects. Each element of N is a mask delineating the specific pixels belonging to the object. This is crucial for accurate instance segmentation, enabling the network to differentiate between multiple instances in close proximity or with overlapping boundaries. It serves as the first step in the aggregation module and acts as the main identifier of objects in the scene.

2. Instance Class: After segmenting the image and obtaining the instance masks, the next step is to classify each mask. Each mask is assigned a label from a predefined set of classes defined within the training dataset. This classification is performed by applying the instance mask to the semantic segmentation output, isolating the portion of interest. The aggregation module then evaluates this masked segment to determine the most probable class label for the instance. It leverages the semantic information provided by the semantic segmentation output to ensure accurate and consistent classification.

3. Mean Uncertainty Estimate of the Instance Mask: To compute the uncertainty estimate for each instance mask, pixel-wise uncertainty values are first extracted across the entire image. Each instance mask is then applied to this uncertainty map to isolate the values corresponding to its pixels. The mean of these values is calculated to obtain a single uncertainty estimate for the instance. This aggregated measure reflects the model's certainty in its prediction for the given object. The procedure is applied systematically to all instance masks, enabling consistent evaluation of uncertainty across all segmented objects.

4. Mean Dirichlet Strength of the Instance Mask: To determine the Dirichlet strength value for each instance mask, the architecture uses the Dirichlet distribution concentration parameters from the Dirichlet layer. This serves as a way to know whether the detected object is an OOD object or not. Each instance mask isolates the pertinent Dirichlet distribution parameters for that specific instance. Summing these parameters and

taking their average per mask yields the Dirichlet strength value of the instance.

5. Mean Distance of the Instance Mask: To calculate the mean distance for each instance mask, the process involves averaging the depth values of the pixels within the mask. Initially, depth estimation provides a per-pixel depth map for the objects in the image. For each identified instance, the corresponding mask is applied to this depth map to extract depth values exclusively within the mask's boundaries. These extracted values are then averaged, resulting in the mean distance for the instance.

The Aggregation Module demonstrates the advantages of using a single shared backbone over a multi-encoder approach. A shared backbone ensures consistent feature representations across all decoders, facilitating seamless alignment of outputs from the semantic segmentation, instance segmentation, and depth estimation branches. This coherence is essential for the aggregation process, enabling a unified and reliable analysis of decoder outputs. The module assigns to each detected instance its corresponding mask, class label, mean depth, uncertainty estimate, and Dirichlet strength. This integration enhances the model's understanding of spatial relationships, improves classification accuracy, and supports more reliable certainty estimation.

3.6 Metrics

In this section, the metrics used to evaluate the performance of the architecture are discussed. To ensure a thorough evaluation, these metrics are divided into two main categories: performance metrics, and uncertainty and OOD metrics.

Performance metrics encompass metrics designed to evaluate specific functions of the architecture, namely semantic segmentation, instance segmentation and depth estimation. Each result is assessed using dedicated metrics, enabling comparative analysis of different outputs. The training of the architecture is conducted in a supervised manner using ground truth data, allowing these metrics to compare the DNN's output with the ground truth for performance evaluation.

Uncertainty and OOD metrics involve evaluating uncertainty estimation and OOD detection in the architecture. One contribution of this work is the use of a wide set of metrics that assess different aspects of the uncertainty and OOD performance of the DNN. These metrics are crucial for understanding how well the architecture can distinguish between ID and OOD data. Together, these metrics offer insights into the architecture's proficiency in estimating uncertainty.

3.6.1 Performance Metrics

Mean Intersection over Union (mIoU)

Intersection over Union (IoU) is a standard metric for quantifying the accuracy of a semantic segmentation model [RTG+19]. It calculates the ratio of the intersection (overlap) between the predicted segmentation and the ground truth to their union (combined coverage).

The IoU for a specific class is defined as:

$$\text{IoU(class)} = \frac{\text{Intersection Area}}{\text{Union Area}} \tag{3.37}$$

The intersection refers to the region where the model prediction and the ground truth overlap, while the union covers the total area occupied by either the prediction, the ground truth, or both. If the model prediction perfectly matches the ground truth, then the intersection and union will be identical, resulting in an IoU of 1.

The mIoU is the average of IoU values across all classes, providing a comprehensive score that reflects the segmentation accuracy for each class in the dataset [EEVG+15]:

$$\text{mIoU} = \frac{\sum \text{IoU(class)}}{\text{Total Number of Classes}} \tag{3.38}$$

Precision and Recall

Precision and recall are fundamental metrics used to evaluate the performance of instance segmentation models, especially in situations where classes are imbalanced [WKM+19].

Precision represents the accuracy of positive predictions made by the classifier. It is the ratio of correctly predicted positive observations to the total predicted positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.39}$$

where TP and FP represent true positives and false positives repectively.

Recall quantifies the completeness of the positive predictions made by the classifier. It is the ratio of correctly predicted positive observations to the actual positives in the dataset [Tha20]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.40}$$

where FN represents false negatives.

Depth Estimation Metrics

Depth estimation performance is evaluated using the relative squared error (RSE) metric, which measures the accuracy of predicted depth values compared to the ground truth. This metric quantifies the squared difference between the predicted depth $\hat{d}$ and the ground truth depth d, normalized by the ground truth depth. RSE is defined as:

$$\text{RSE} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\hat{d}_i - d_i}{d_i} \right)^2, \tag{3.41}$$

where N is the total number of pixels, $\hat{d}_i$ is the predicted depth for the i-th pixel, and d_i is the corresponding ground truth depth.

The relative squared error emphasizes large relative errors, making it particularly sensitive to significant deviations in depth predictions.

3.6.2 Uncertainty and OOD Metrics

Distributional Separation Efficiency

The ability to separate between correct and incorrect classifications is an important factor in the tests. An ideal network should be able to show high certainty on its correct predictions and low certainty on incorrect predictions. Accordingly, the aim is to quantify the efficiency of the DNN to differentiate between correct and incorrect predictions by plotting their corresponding certainty distribution for both cases. In this case, the Wasserstein distance metric is used to compare the two distributions. This metric, often used in optimal transport problems, calculates the minimum cost to transform one distribution into another. In this context, a high Wasserstein distance indicates that the distributions of certainty for correct and incorrect predictions are well-separated, reflecting the network's ability to differentiate its predictions based on certainty. Conversely, a low Wasserstein distance implies substantial overlap between the distributions, suggesting that the model assigns similar certainty levels to both correct and incorrect predictions [KPM+18].

Accuracy vs. Certainty Metrics

One of the most important factors for DNNs equipped with uncertainty estimation is not to be only accurate about its outputs, but to be also certain about it. This was one of the main contributions of [MG18b] by defining some metrics directly related to this aspect: accuracy vs. certainty.

For these metrics, the initial step involves extracting the segmentation and uncertainty maps from the network's outputs. Once generated, these maps are compared against the ground truth to calculate four key components: predictions that are both accurate and certain (n_{ac}), accurate but uncertain predictions

(n_{au}), incorrect but certain predictions (n_{ic}), and predictions that are both incorrect and uncertain (n_{iu}). Notably, these calculations can be applied at both the pixel level, where each individual pixel in the ground truth is compared with its counterpart in the predictive maps, and at the instance level, evaluating entire objects or regions. The terms *certainty* and *uncertainty* rely on adjustable thresholds, where the metrics are calculated on increasing thresholds to assess DNN's performance at different certainty threshold. These metrics are represented in the below table sorting outcomes based on their accuracy and certainty levels [MG18b]:

	Accurate	**I**naccurate
Certain	AC	IC
Uncertain	AU	IU

From these foundational elements, three conditional probabilities emerge. The first probability $p(\text{accurate}|\text{certain})$ quantifies the likelihood of an accurate prediction when made with certainty. The second probability $p(\text{uncertain}|\text{inaccurate})$ assesses the probability of an uncertain prediction given it is incorrect. And the third, accuracy vs. uncertainty (AvU), computes the likelihood that a network's output falls into one of two categories: confidently accurate or uncertainly inaccurate. The three conditional probabilities are calculated as follows [MG18b]:

$$p(\text{accurate}|\text{certain}) = \frac{n_{ac}}{n_{ac} + n_{ic}} \tag{3.42}$$

$$p(\text{uncertain}|\text{inaccurate}) = \frac{n_{iu}}{n_{ic} + n_{iu}} \tag{3.43}$$

$$\text{AvU} = \frac{n_{ac} + n_{iu}}{n_{ac} + n_{au} + n_{ic} + n_{iu}} \tag{3.44}$$

When iterating through various uncertainty thresholds, the value of each metric stands as an indicator of the network's efficacy at that particular threshold. A higher value in the metric points to a more robust performance.

The outcomes of these probabilities at thresholds from 0% certainty to 100% certainty are plotted and the area under the curve for each metric is calculated. The higher the area under the curve for each metric the better the uncertainty estimation of the DNN.

Receiver Operating Characteristic Curve Metrics

The Receiver Operating Characteristic (ROC) curve is a widely utilized metric in the evaluation of uncertainty estimation [APH+21, GTA+21], particularly in binary classifiers. It plots the relationship between the true positive rate (TPR) and the false positive rate (FPR), where TPR and FPR are calculated as follows:

$$\text{TPR} = \frac{TP}{TP + FN}, \qquad \text{FPR} = \frac{FP}{FP + TN} \tag{3.45}$$

with *TP* for true positives and *FN* for false negatives, *FP* representing false positives and *TN* true negatives. In this context, correct predictions are labeled as positive and incorrect ones as negative.

Two key metrics derived from the Receiver Operating Characteristic (ROC) curve are the Area Under the ROC Curve (AUROC) and the False Positive Rate at 95% True Positive Rate (FPR@0.95TPR). AUROC provides an aggregate measure of classification performance across all decision thresholds, where a value of 0.5 indicates random guessing and 1.0 denotes perfect discrimination. In contrast, FPR@0.95TPR quantifies the proportion of false positives when the true positive rate reaches 95%. This metric is particularly valuable in high-risk applications, where maintaining a high level of certainty in correct classifications is essential while minimizing incorrect detections.

Calibration of Deep Neural Networks

The calibration of a DNN is critical for ensuring the network's safety and reliability. Calibration refers to the network's ability to align its observed accuracy with its certainty, ensuring that the certainty level associated with a predicted class label accurately reflects its correctness [WGW23].

In this work, the calibration of the DNN is assessed using the expected calibration error (ECE) and maximum calibration error (MCE). The calibration process involves organizing segmentation predictions into predetermined bins B_i (where $i = 1, \ldots, B$) based on their certainty levels. For each bin B_i, accuracy (acc) and certainty (cert) are calculated as follows [GPSW17]:

$$\mathrm{acc}_i = \frac{\mathrm{TP}_i}{|B_i|}, \qquad \mathrm{cert}_i = \frac{1}{|B_i|} \sum_{j=1}^{|B_i|} \hat{c}_i \tag{3.46}$$

Here, $|B_i|$ denotes the number of examples in bin B_i, $\hat{c}_i$ represents the respective certainty or entropy score, and TP_i is the count of correctly classified instances in B_i. The ECE and MCE are then computed as:

$$\mathrm{ECE} = \frac{1}{B} \sum_{i=1}^{B} |\mathrm{acc}_i - \mathrm{cert}_i|, \qquad \mathrm{MCE} = \max |\mathrm{acc}_i - \mathrm{cert}_i| \tag{3.47}$$

These metrics quantitatively evaluate the degree of calibration in a DNN, indicating the alignment between the model's certainty levels and its actual performance, which is vital for the safety and efficacy of ADS functions [KLM19].

OOD Detection and Identification Performance

The evaluation of instance-level OOD detection relies on three key metrics: $\mathrm{AP@50\%_{OOD}}$, $\mathrm{Recall_{OOD}}$, and iOOD. These metrics assess the model's ability to correctly identify OOD objects while minimizing false detections of ID objects.

The $\mathrm{AP@50\%_{OOD}}$ metric measures the average precision of OOD instance detection when using a 50% IoU threshold. This means that for an OOD instance to be considered correctly detected, the predicted instance mask must have at least 50% overlap with the ground truth OOD mask. The metric captures the trade-off between precision and recall, ensuring that models can reliably detect OOD objects while avoiding excessive false positives.

The $\text{Recall}_{\text{OOD}}$ metric evaluates the model's ability to correctly identify OOD instances within a given dataset. It calculates the proportion of true OOD objects that have been successfully detected relative to the total number of ground-truth OOD instances. A high $\text{Recall}_{\text{OOD}}$ value indicates that the model is effective at recognizing unknown objects.

The iOOD metric provides an instance-level assessment of OOD detection by analyzing the model's certainty in identifying OOD objects. The iOOD focuses on the average certainty score assigned to instances that have been detected as OOD. It quantifies how certain the model is that a given detected object is OOD, offering insight into the reliability of its predictions. A higher iOOD score indicates that the model consistently assigns strong OOD certainty to correctly identified unknown instances, which the model's ability in distinguishing between known and unknown objects.

4 Experiments, Results and Discussion

This chapter presents a comprehensive evaluation of the proposed architecture for uncertainty estimation and OOD detection and identification in object segmentation tasks for ADS. The experiments are designed to assess the model's performance across multiple dimensions, including semantic and instance segmentation accuracy, the quality of uncertainty quantification, and the robustness of OOD detection at both pixel and instance levels.

The evaluation strategy follows a systematic and rigorous design, beginning with training on the Cityscapes dataset [COR+16], followed by testing on a diverse set of benchmark datasets, KITTI [GLSU13], A2D2 [GKM+20], and BDD100K [YCW+20], to assess the model's generalization to various real-world driving scenarios. To evaluate the OOD detection capabilities, the Lost and Found dataset [PRG+16] is employed. Both pixel-level and instance-level analyses are conducted: the former focuses on evaluating the uncertainty of individual OOD pixels, while the latter assesses the model's ability to reliably detect and segment entire unknown objects.

The chapter begins with a detailed description of the experimental setup, including the datasets, and evaluation setup. It then presents a thorough analysis of the results, encompassing segmentation performance, uncertainty calibration, and OOD detection effectiveness. Particular attention is given to how uncertainty-aware modeling contributes to reducing false positive detections and enhancing the segmentation of previously unseen objects. Finally, the results are benchmarked against state-of-the-art methods to highlight the advancements introduced by the proposed architecture.

4.1 Experiments Setup

This section outlines the datasets used for training and evaluation, followed by the details of the model training procedure and comparison strategy.

4.1.1 Datasets

Five datasets were employed to train and evaluate the methods, each supporting different aspects of perception in autonomous driving.

Cityscapes

A widely used benchmark containing high-resolution images from various German cities, annotated with pixel-level labels for 30 classes. These classes include roads, buildings, vehicles, and pedestrians. Cityscapes provides training, validation, and test splits, facilitating model development and consistent evaluation [COR+16].

KITTI

The Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) dataset focuses on autonomous driving and 3D scene understanding [GLSU13], gathered by a vehicle equipped with multiple sensors. It spans urban, suburban, and rural roads under varying lighting and weather conditions. KITTI contains labeled cars, pedestrians, and cyclists, introducing challenges such as occlusions and complex traffic dynamics.

BDD

The Berkeley DeepDrive (BDD) dataset with large-scale collection of dashcam recordings in urban environments, featuring diverse driving scenes such as highways, residential areas, and parking lots. This dataset captures different weather conditions, traffic situations, and lighting variations, making it suitable for testing model robustness in real-world settings [YCW+20].

A2D2

The Audi Autonomous Driving Dataset (A2D2) is a dataset assembled from Audi's fleet of test vehicles, covering urban, suburban, and highway scenarios under multiple weather and lighting conditions [GKM+20]. Its detailed annotations and variety of road scenes enable evaluation of both semantic and instance segmentation tasks in realistic driving environments.

Fishyscapes Lost and Found

An extension of the Lost and Found dataset [PRG+16], adapted by Fishyscapes to assess OOD detection [BSN+21]. It contains street scenes augmented with unfamiliar or anomalous objects not encountered during training, such as debris or barriers, thus enabling OOD detection experiments. This dataset focuses on measuring how well a model distinguishes ID versus anomalous inputs in real driving scenarios.

4.1.2 Model Training and Evaluation

All models were trained solely on the Cityscapes dataset, selected for its high-resolution imagery and extensive pixel-level annotations suited to semantic and instance segmentation tasks. KITTI, BDD, and A2D2 were retained only for evaluation, allowing an assessment of the trained model's generalization capabilities in diverse environments and conditions.

For OOD detection, the Fishyscapes Lost and Found dataset was used, presenting real-world road anomalies absent from Cityscapes. Both pixel-level and instance-level evaluations were conducted on OOD objects. Pixel-level evaluation focused on identifying unknown pixels within the scene, while instance-level evaluation targeted the complete detection and classification of anomalous objects.

Three models were tested for comparison. The first used cross-entropy (CE) loss as a baseline. The remaining two employed the prior and the evidential uncertainty estimation techniques, respectively, providing a state-of-the-art reference for uncertainty-aware approaches. Both SOTA methods and the baseline use predictive entropy as the model's uncertainty estimation.

All backbones were pretrained on ImageNet. Training used the Adam optimizer with a base learning rate of 5e-4 and a polynomial schedule (with a learning step set to 1), and a weight decay of 0.0005. Each run consisted of 200 epochs with a batch size of 8 on 1024×1024 crops. The only augmentation used was random cropping. All experiments were executed on a single NVIDIA A100 (40 GB) GPU using PyTorch.

4.2 Pixel-level Results

This section examines pixel-level performance for the evaluated models, specifically focusing on ID data. The analysis considers several configurations of the proposed architecture alongside a baseline and two state-of-the-art approaches. The configurations include the CE baseline model with ILVI, Dirichlet Maximum Likelihood Estimation (Dirichlet MLE), and a combined implementation of Dirichlet MLE with ILVI. This comparative approach aims to provide an understanding of the influence of ILVI and Dirchlet MLE on uncertainty estimation and semantic segmentation performance.

The evaluation is structured into four key subsections. The first subsection investigates distributional separation efficiency by analyzing the certainty distributions of correct and incorrect predictions. This is quantified using the Wasserstein distance metric, which measures the extent of separation between these distributions. The second subsection assesses segmentation performance using mIoU as the evaluation metric, allowing for a quantitative comparison of pixel-level classification accuracy across methods.

The third subsection addresses calibration, analyzing metrics such as ECE and MCE to evaluate how well predicted certainty aligns with actual accuracy. Certainty histograms are included to provide a visual representation of the models' calibration behavior. The fourth subsection explores the relationship between accuracy and certainty, utilizing metrics that measure the models' ability to correlate certainty levels with prediction correctness and uncertainty with incorrect predictions.

4.2.1 Distributional Separation Efficiency

The evaluation of the distributional separation efficiency focuses on analyzing the certainty distributions of correct and incorrect predictions. The separation between these distributions is quantified using the Wasserstein distance metric, which measures the degree of separation. Higher Wasserstein distance values indicate better separation, reflecting improved uncertainty estimation and a stronger ability to distinguish between correct and incorrect predictions. Lower values suggest significant overlap, signaling less reliable differentiation. The distribution plots are shown in Figure 4.1 and the Wasserstein distance values are recorded in Table 4.1.

The CE baseline achieves the lowest Wasserstein distance of 7.83, indicating minimal separation between correct and incorrect predictions. This is visually evident in the corresponding plot, where the distributions of correct and incorrect predictions overlap significantly across the certainty range. The substantial overlap in the high-certainty region highlights the baseline model's inability to reliably associate high certainty with correct predictions, often assigning high certainty to incorrect predictions.

The Prior model improves the Wasserstein distance to 23.15, representing a roughly threefold increase over the baseline. This improvement is reflected in the plot, where incorrect predictions are more concentrated in the low-certainty region, and correct predictions occupy the higher-certainty range more consistently. However, the overlap in the 50% to the 100% certainty region remains significant, limiting the model's ability to fully distinguish

Table 4.1: Distributional separation efficiency metrics comparing different models based on their Wasserstein distance. A higher Wasserstein distance indicates better separation between correct and incorrect predictions, reflecting improved uncertainty estimation.

	Wasserstein Distance (↑)
CE (Baseline)	7.83
Prior	23.15
Evidential	45.68
CE + ILVI	31.09
Dirichlet MLE	51.32
Dirichlet MLE + ILVI	59.26

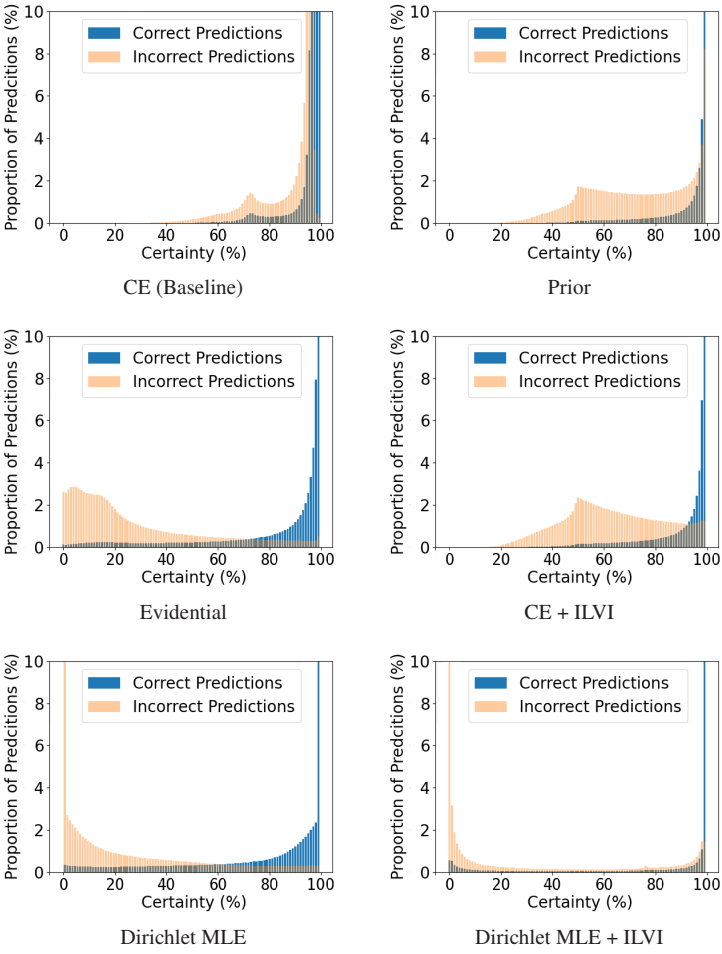


Figure 4.1: Distributional separation plots are presented, depicting the distribution of correct predictions (blue bars) and incorrect predictions (orange bars) across the methods. The x-axis represents certainty values of the methods, while the y-axis shows the proportion of predictions. Each plot illustrates the overlap or separation between correct and incorrect predictions, providing a visual comparison of the distributional behavior for each method. It is worth noting that the y-axis is only limited to 10% to better visualize the separation between the distributions.

prediction types. Furthermore, the incorrect predictions are not assigned low certainty, which adds to the overlap.

The Evidential model achieves a Wasserstein distance of 45.68, nearly doubling the Wasserstein distance value of the Prior model. The plot shows a clearer distinction between correct and incorrect predictions. The reduced overlap in the mid-certainty range signifies a significant improvement in uncertainty alignment, allowing the Evidential model to better distinguish between prediction correctness and uncertainty levels.

The CE + ILVI model achieves a Wasserstein distance of 31.09, demonstrating a notable improvement over the CE baseline and the Prior model. The corresponding plot shows clearer separation, with incorrect predictions more concentrated in the low-certainty region. However, compared to the Evidential model, the overlap in the mid-certainty range is more pronounced, which explains the relatively lower Wasserstein distance. The ILVI addition enhances the separation efficiency compared to CE alone.

The Dirichlet MLE model achieves a Wasserstein distance of 51.32, marking a significant improvement over all the previous models, including the Evidential model. The plot shows a pronounced separation between correct and incorrect predictions, with incorrect predictions tightly clustered in the low-certainty range and correct predictions concentrated at high-certainty values. The reduced overlap across all certainty regions highlights the method's robust ability to align uncertainty with prediction outcomes.

The Dirichlet MLE + ILVI model achieves the highest Wasserstein distance of 59.26, reflecting the most effective separation among all evaluated models. The plot reveals strong separation, with correct predictions tightly grouped at high-certainty values and incorrect predictions concentrated at low-certainty values. Compared to the CE baseline, the Dirichlet MLE + ILVI model achieves a more than sevenfold improvement in Wasserstein distance. This result underscores the complementary role of ILVI in enhancing the separation achieved by Dirichlet MLE and highlights its effectiveness in utilizing uncertainty to distinguish between prediction outcomes.

4.2.2 Segmentation Performance

The segmentation performance results, summarized in Table 4.2, assess the mIoU of the different methods. By comparing the results, the relative improvements across methods can be analyzed, providing insight into how effectively each approach performs with respect to segmentation.

The CE baseline achieves a mIoU of 68.72%, providing the initial benchmark for pixel-level segmentation. The Prior model increases the baseline mIoU to 70.85%, reflecting a relative improvement of approximately 3.1 percentage points over the baseline. The Evidential model further improves the mIoU to 71.42%, a smaller relative gain of 0.8 percentage points over the Prior model.

The CE + ILVI model achieves a mIoU of 69.86%, reflecting a relative improvement of approximately 1.7 percentage points over the baseline. While the addition of ILVI enhances segmentation performance compared to the CE baseline, the improvement is less pronounced compared to the Prior and Evidential models. This result suggests that the inclusion of ILVI gives the baseline CE a boost.

The Dirichlet MLE model achieves a mIoU of 71.26%, marginally lower than the Evidential model but maintaining competitive performance. The improvement of approximately 2.5 percentage points over the baseline highlights the benefits of employing Dirichlet MLE for uncertainty modeling, particularly in its ability to align uncertainty estimation with segmentation outcomes.

The Dirichlet MLE + ILVI model achieves the highest mIoU of 72.84%. This result reflects a substantial improvement of approximately 4.1 percentage points over the baseline and 1.4 percentage points over the highest SOTA Evidential

Table 4.2: Segmentation performance using Mean Intersection over Union.

	mIoU in % (↑)
CE (Baseline)	68.72
Prior	70.85
Evidential	71.42
CE + ILVI	69.86
Dirichlet MLE	71.26
Dirichlet MLE + ILVI	72.84

model. The enhanced performance indicates that the combination of Dirichlet MLE and ILVI addresses the challenges of segmentation tasks by leveraging the strengths of both methods. The relative improvement compared to Dirichlet MLE alone suggests that ILVI provides a complementary contribution that enhances segmentation performance further.

Figures 4.2 and 4.3 present the segmentation performance of the models on a Cityscapes sample with their corresponding uncertainty estimation. Table 4.3 shows the classes' color map for the segmentation sample results. The unlabeled class, shown in black, denotes pixels that do not correspond to any annotated semantic category. These pixels are omitted from training and evaluation, ensuring that they do not influence model learning or performance metrics.

road	sidewalk	vegetation	terrain	sky
building	wall	fence	traffic light	traffic sign
car	person	rider	bicycle	motorcycle
truck	trailer	train	bus	caravan
unlabeled				

Table 4.3: Color scheme representing class labels for the semantic segmentation task, including the unlabeled class shown in black.

Relating these figures to the distributional separation plots in Figure 4.1, the Dirichlet MLE + ILVI model produces a binary-like uncertainty representation that reflects the characteristics of the distributional separation plot. The plot reveals two peaks, one at high certainty and one at low certainty, corresponding to regions of high certainty and low certainty in the model's predictions respectively. This is directly mirrored in the uncertainty maps generated by the Dirichlet MLE + ILVI model, where white pixels represent high certainty and black pixels represent low certainty.

This binary-like behavior is less pronounced in the baseline and SOTA models, which tend to exhibit smoother uncertainty maps that lack the distinct separation observed in the Dirichlet MLE + ILVI approach. The correlation between the peaks in the Wasserstein distance plot and the binary-like uncertainty maps in the Dirichlet MLE + ILVI model highlights a strong relationship between

the model's uncertainty representation and its ability to differentiate certainty levels.

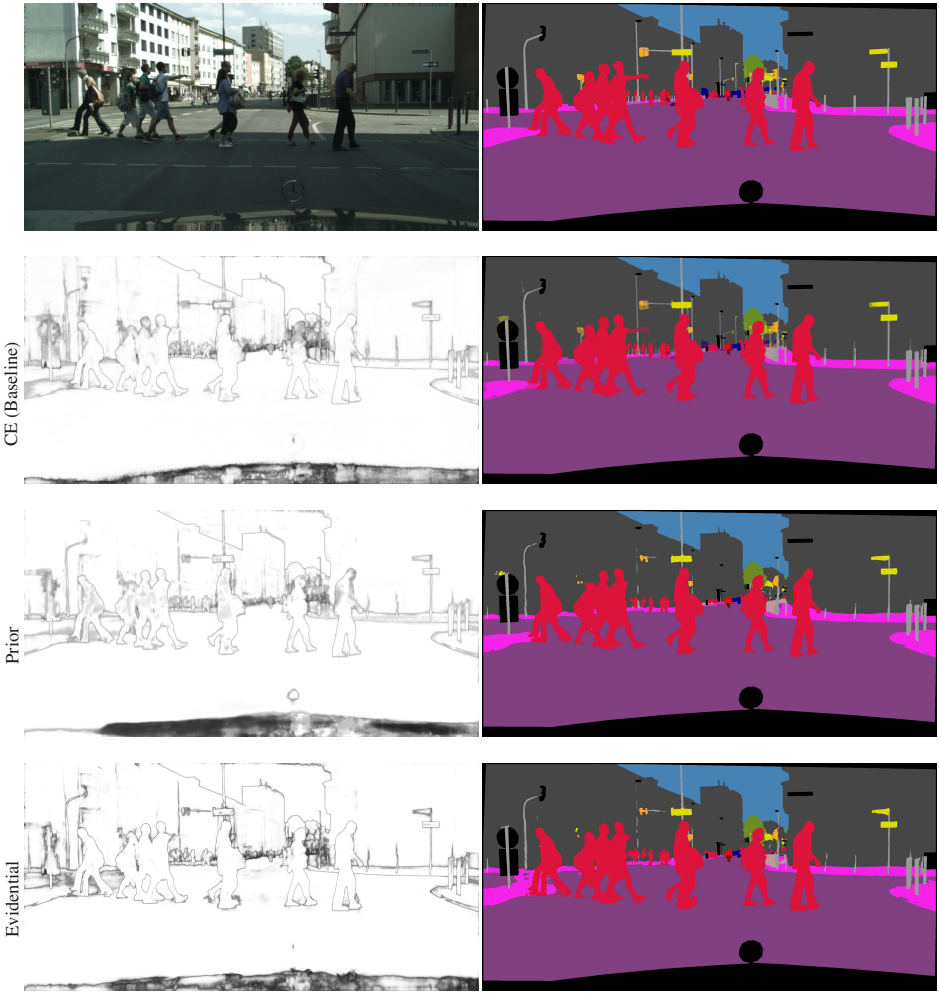


Figure 4.2: Sample visualization of pixel-wise segmentation and uncertainty estimation outputs of the baseline model and both SOTA models. The first row displays the input image (left) and its ground truth segmentation map (right). Subsequent rows show the uncertainty estimation maps on the left, and the corresponding semantic segmentation results on the right. The brightness of the pixels in the uncertainty estimation map indicates the level of certainty.

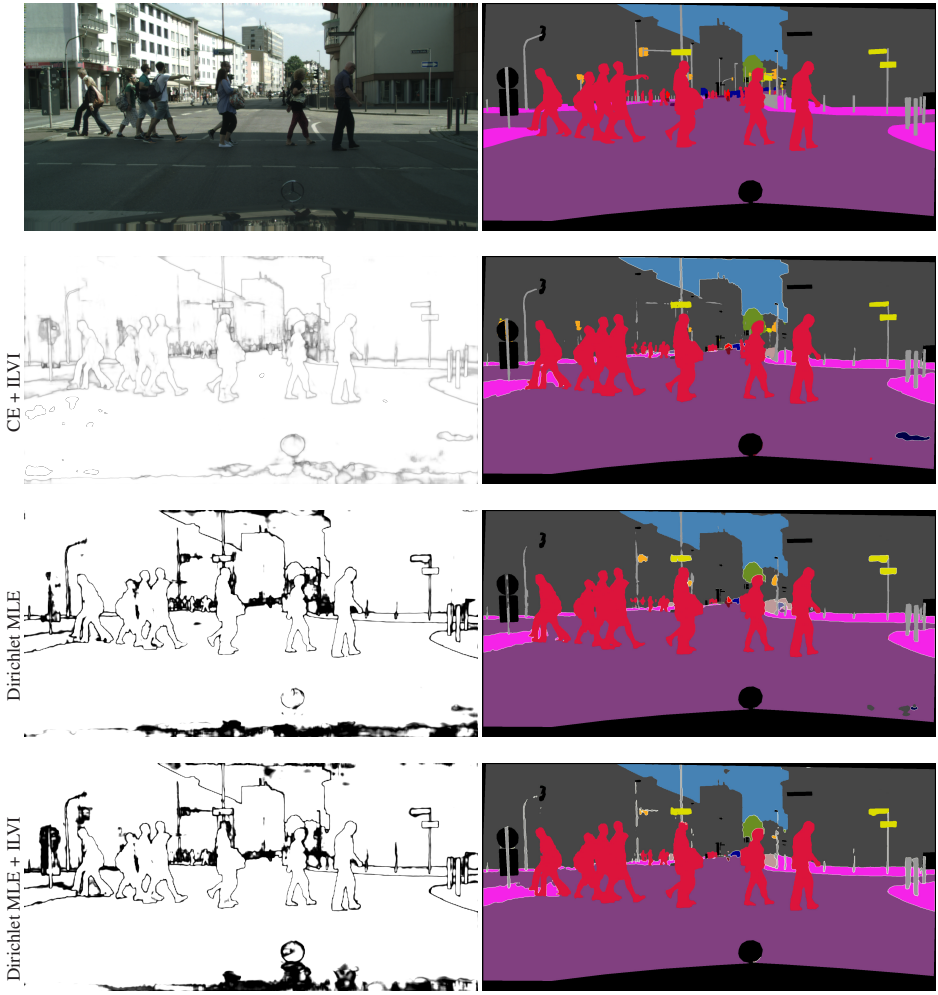


Figure 4.3: Sample visualization of pixel-wise segmentation and uncertainty estimation outputs of CE+ILVI, Dirichet MLE and Dirichlet MLE + ILVI. The first row displays the input image (left) and its ground truth segmentation map (right). Subsequent rows show the uncertainty estimation maps on the left, and the corresponding semantic segmentation results on the right. The brightness of the pixels in the uncertainty estimation map indicates the level of certainty.

4.2.3 Calibration

Model calibration is a crucial aspect of uncertainty estimation, ensuring that the predicted certainty levels correspond to the actual accuracy. A well-calibrated model produces outputs that accurately reflect the likelihood of correct predictions. This section evaluates the calibration performance of the models using ECE and MCE as key metrics. ECE measures the average discrepancy between predicted certainty and observed accuracy, while MCE captures the worst-case deviation, highlighting instances of extreme overconfidence [SHB+22].

The calibration error plots and calibration metrics, presented in Figure 4.4 and Table 4.4 respectively, provide insights into the reliability of the evaluated models. The reduction in ECE and MCE across the evaluated methods illustrates the incremental improvement in aligning predicted certainty with actual accuracy. The CE baseline exhibits the highest values for both ECE and MCE, indicating significant miscalibration where predictions are frequently overconfident, as confirmed by the sharp peaks in high-certainty bins of its respective histogram. This reflects the model's tendency to assign high certainty to incorrect predictions.

As models incorporate uncertainty-aware techniques, the calibration errors progressively decrease. The Prior model demonstrates some reduction in ECE and MCE compared to the CE baseline, but the histogram reveals residual clustering in high-certainty bins, suggesting that overconfidence remains an issue. Similarly, the MCE values for the Prior model.

The Evidential model further reduces calibration errors, with both ECE and MCE showing substantial improvement. The histogram for this model depicts a more uniform certainty distribution, with fewer predictions concentrated in the high-certainty bins. This indicates that the Evidential model achieves a better alignment between certainty and accuracy, reducing both average miscalibration and extreme cases of overconfidence. However, localized areas of error still remain.

The CE + ILVI model shows modest calibration improvements over the baseline but does not achieve the same level as the Evidential model nor Dirichlet-based methods. The histogram reveals similar characteristics to the Prior model, with some reduction in high-certainty peaks but without a significant redistribution of certainty values. This is consistent with its ECE and MCE values, suggesting

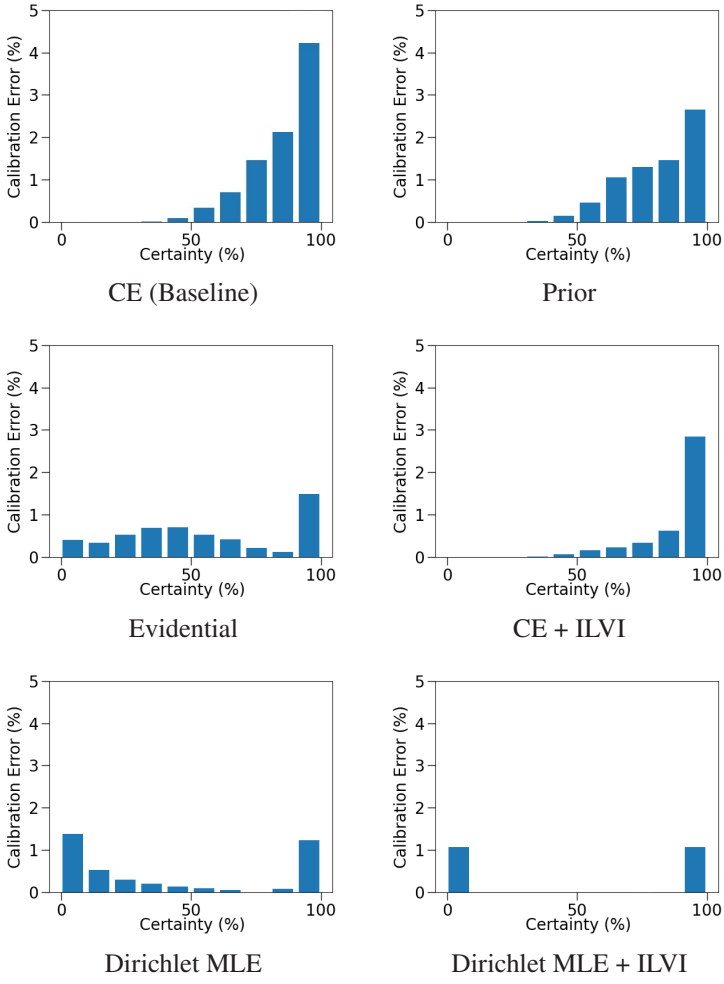


Figure 4.4: Calibration error plots illustrating the distribution of calibration error across different certainty levels for each evaluated model. The x-axis represents certainty values, while the y-axis shows the corresponding calibration error percentage.

Table 4.4: Calibration metrics comparing ECE and MCE across different models. Lower values indicate better calibration, meaning the model's predicted certainty more accurately reflects actual accuracy.

	ECE (↓)	MCE (↓)
CE (Baseline)	1.12	4.53
Prior	0.77	2.91
Evidential	0.54	1.48
CE + ILVI	0.55	2.89
Dirichlet MLE	0.48	1.45
Dirichlet MLE + ILVI	0.23	1.15

that while ILVI enhances calibration to some extent, its integration with CE lacks the robustness observed in the other models.

Dirichlet MLE exhibits notable calibration improvements, with its histogram showing a significant reduction in high-certainty peaks and a more even spread of certainty values. This is reflected in its ECE and MCE values, which are markedly lower than those of previous models. The histogram supports the conclusion that Dirichlet MLE improves calibration across the histogram-bins. This improvement highlights the impact of Dirichlet-based uncertainty modeling in aligning certainty and accuracy.

The Dirichlet MLE + ILVI model demonstrates the most significant calibration improvement, achieving the lowest ECE and MCE values among all evaluated models. When analyzing the calibration error plots for the models, it is evident that each model exhibits the highest calibration error at the points of most clustered prediction certainty. Notably, the Dirichlet MLE + ILVI demonstrates the lowest calibration error among its clustered predictions when compared to the other models. This indicates that the Dirichlet MLE + ILVI approach provides a more reliable estimation of prediction certainty, enhancing its performance in semantic segmentation tasks.

4.2.4 Accuracy vs. Certainty Analysis

The section presents a comparison of the models' performance in terms of their ability to align certainty with accuracy. Three metrics, P(Accurate|Certain), P(Uncertain|Inaccurate), and Accuracy vs. Uncertainty, are used to evaluate each approach, where they measure how often the model is accurate when it is confident, how often the model is uncertain when it makes an incorrect prediction and the overall alignment between prediction accuracy and model certainty respectively. The plots of all three metrics are shown in Figures 4.5 and 4.6, and the results are reported in Table 4.5

Across the evaluated models, the P(Accurate|Certain) metric shows relatively modest variations. The baseline, Prior, Evidential, and CE+ILVI models all achieve P(Accurate|Certain) values in a similar range, around 94% to 96%, whereas the Dirichlet MLE and Dirichlet MLE + ILVI within the range of 96% and 98%. This indicates that the models are highly confident when the predictions are correct.

In contrast, the P(Uncertain|Inaccurate) and Accuracy vs. Uncertainty metrics highlight more substantial differences among the models. The baseline's P(Uncertain|Inaccurate) of 18.28% and Accuracy vs. Uncertainty of 76.34% reflect limited uncertainty estimation performance, with the model often failing to correctly express uncertainty when it is incorrect. The Prior model improves both values to 27.06% and 86.18%, respectively, indicating an improvement in uncertainty estimates. The Evidential model remains in a similar range, with a higher P(Uncertain|Inaccurate) of 43.71% and Accuracy vs. Uncertainty of 88.75%, maintaining a balanced performance.

Table 4.5: Accuracy vs. Certainty Metrics

	P(A | C) in % (↑)	P(U | I) in % (↑)	Accuracy vs. Uncertainty in % (↑)
CE (Baseline)	94.51	18.28	76.34
Prior	95.15	27.06	86.18
Evidential	95.46	43.71	88.75
CE + ILVI	95.31	35.75	87.46
Dirichlet MLE	96.96	58.26	87.69
Dirichlet MLE + ILVI	97.91	70.15	89.86

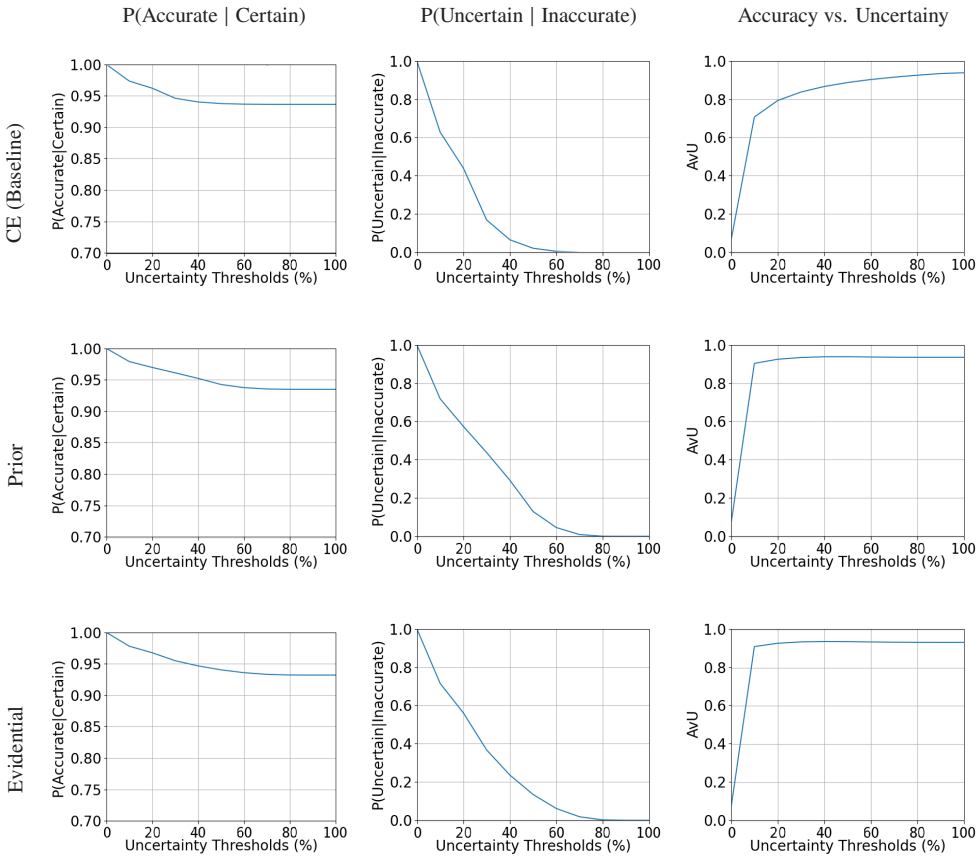


Figure 4.5: Accuracy vs. Certainty graphs showing the baseline (CE) model and SOTA models. The three columns represent the three metrics P(Accurate | Certain), P(Uncertain | Inaccurate) and AvU (Accurate vs. Uncertain) respectively.

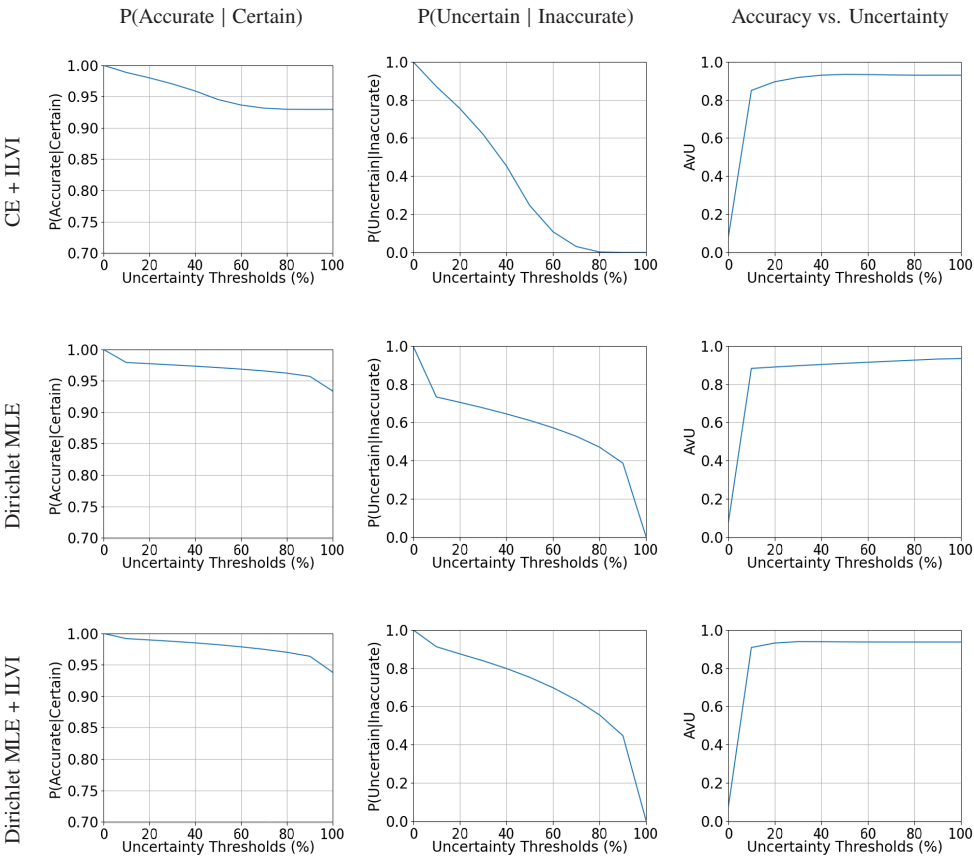


Figure 4.6: Accuracy vs. Certainty graphs showing the CE+ILVI and both Dirichlet-based models. The three columns represent the three metrics P(Accurate | Certain), P(Uncertain | Inaccurate) and AvU (Accurate vs. Uncertain) respectively.

The CE+ILVI model improves the baseline's P(Uncertain|Inaccurate) to 35.75% and achieves an Accuracy vs. Uncertainty score of 87.46%, bringing it closer to the performance of SOTA models. The Dirichlet MLE model further improves both metrics, reaching a P(Uncertain|Inaccurate) of 58.26% and an Accuracy vs. Uncertainty score of 87.69%. The Dirichlet MLE + ILVI model delivers the strongest performance overall, achieving the highest P(Uncertain|Inaccurate) of 70.15% and an Accuracy vs. Uncertainty score of 89.86%. These results demonstrate that combining Dirichlet-based modeling with ILVI not only enhances the alignment between certainty and accuracy but also significantly improves the model's trustworthiness on incorrect predictions, more effectively associating high certainty with correctness and high uncertainty with error than the other models.

4.2.5 Discussion

This section compares the performance of the evaluated models on ID data, focusing on the effects of integrating ILVI with the CE baseline model, the performance of the Dirichlet MLE model on its own, and the combined integration of Dirichlet MLE with ILVI. The analysis spans four key aspects: distributional separation, calibration, accuracy versus certainty, and segmentation performance.

The inclusion of ILVI in the CE baseline model demonstrates consistent improvements across all evaluated metrics on the baseline. ILVI enhances the separation between correct and incorrect distributions, as evidenced by a greater separation between the correct and incorrect distributions. Additionally, it improves calibration by reducing both ECE and MCE, and strengthens the correlation between accuracy and certainty. This integration significantly enhances the baseline model's performance, aligning it with SOTA approaches. Furthermore, ILVI provides a measurable boost to the segmentation performance.

The Dirichlet MLE model, even without ILVI, achieves notable performance and exhibits characteristics that are competitive with SOTA models. It exhibits distinct separation between correct and incorrect distributions, characterized by two prominent peaks, one at high certainty and another at low certainty. The accuracy versus certainty metrics further validate its effectiveness, outperforming other SOTA approaches in aligning certainty with correctness and uncertainty with error. The calibration performance and segmentation accuracy also improved when compared to the CE baseline.

Integrating ILVI into the Dirichlet MLE model results in the highest performance across all evaluation criteria. The model achieves great distributional separation, with minimal overlap between correct and incorrect predictions. Its accuracy versus certainty metrics also reflect substantial gains, showing superior alignment compared to the baseline and SOTA approaches. The calibration performance of the Dirichlet MLE + ILVI model is particularly notable, with calibration error plots demonstrating its ability to assign certainty levels that accurately reflect prediction correctness. Lastly, this integration provides a small yet measurable improvement in segmentation performance, achieving the highest accuracy among the evaluated models.

It is important to emphasize that maintaining segmentation performance is a critical requirement for any new uncertainty estimation approach. This criterion ensures that enhancements in uncertainty-related metrics do not degrade the model's fundamental segmentation capabilities, an objective that Dirichlet MLE + ILVI approach succeeds in. The analysis also excluded the application of ad-hoc or external calibration methods, allowing the models' intrinsic calibration capabilities to be evaluated. This choice was necessary to isolate and assess the inherent effectiveness of each approach without introducing confounding variables. The results validate this decision, as the Dirichlet MLE + ILVI model demonstrated superior calibration performance compared to the baseline and SOTA approaches.

4.3 Instance Segmentation and OOD Detection Results

This section provides an examination of the models' performance in instance segmentation, with particular attention to their ability to detect and identify OOD objects at the instance level. The evaluation encompasses four model types: CE, Prior networks, Evidential networks, and Dirichlet MLE combined with ILVI. Each model is assessed using two backbones: MobileNet and EfficientNet.

The first part of the analysis focuses on the models' instance segmentation capabilities. This includes an evaluation of generalization performance, measured by how effectively the models segment instances from both ID and shifted data domains. In addition, the analysis examines the separation between correct and incorrect predictions, the distributional separation performance. The analysis then extends to the effect of uncertainty thresholding, examining how applying uncertainty-based filters influences the reduction of false positive detections without significantly compromising true positive rates.

Subsequently, the depth estimation is also analyzed. This part evaluates the accuracy of per-pixel depth predictions associated with the segmented instances. The objective is to assess how well the models preserve spatial coherence and relative distance information.

The final part of this section presents two sets of experiments designed to evaluate the models' ability to detect and classify OOD instances. The first set focuses on the iOOD metric, assessing the models' capacity to differentiate novel objects not seen during training using segmentation outputs and uncertainty estimates. The second set involves a quantitative analysis based on ROC curves, which evaluates the discriminative performance of the models in separating OOD from ID instances based on uncertainty scores.

4.3.1 Instance Segmentation Performance

In this subsection, the instance segmentation performance of the models is compared across both backbone architectures. Tables summarizing these results include the metrics average precision (AP), AP@50%, and the distance-based metrics AP@50m and AP@100m, offering a comprehensive view of detection quality at different thresholds and distances.

The instance segmentation performance of the models is evaluated under two distinct settings. The first evaluation is performed on the validation split of the Cityscapes dataset, which shares the same distribution as the training data but consists of a disjoint subset of images. This setup allows for assessing the models' performance on the ID data before analyzing their ability to generalize beyond the training domain. The corresponding results are presented in Table 4.6. The second evaluation focuses on the generalization capabilities of the models using unseen datasets with differing characteristics, namely KITTI, BDD, and A2D2, with results reported in Table 4.7. Representative instance segmentation outputs from the Dirichlet MLE + ILVI EfficientNet model on Cityscapes samples are shown in Figure 4.7.

When evaluated on Cityscapes, all methods generally achieve their best performance, as this dataset reflects the distribution the models were trained on. In contrast, the performance declines observed on KITTI, BDD, and A2D2 highlight the challenges of generalization; however, the extent of this drop varies across methods.

Dirichlet MLE + ILVI stands out in all results as it consistently performs gains over the baseline and SOTA models across both backbone architectures. A key strength emerges from its superior performance in long-range detection tasks, particularly visible in distance-specific metrics like AP@50m and AP@100m. This robustness is crucial for reliable detection of far-off obstacles.

An equally important advantage of Dirichlet MLE + ILVI lies in its robustness under distribution shifts. While every model experiences some performance drop when transitioning from Cityscapes to the more diverse and differently distributed datasets (KITTI, BDD, and A2D2), Dirichlet MLE + ILVI still remains the best.

	Cityscapes			
	AP	AP 50%	AP 50m	AP 100m
MobileNet Variant				
CE	22.9	38.2	35.8	31.9
Prior	23.1	39.8	37.5	33.2
Evidential	23.8	42.9	39.8	35.6
Dirichlet MLE + ILVI	26.3	45.1	43.5	42.2
EfficientNet Variant				
CE	31.8	48.1	45.5	41.6
Prior	31.5	47.5	45.9	42.1
Evidential	32.3	50.7	47.5	43.9
Dirichlet MLE + ILVI	32.5	51.3	48.1	44.2

Table 4.6: Instance segmentation performance (%) of the evaluated methods on Cityscapes, the training dataset, for both MobileNet and EfficientNet architectures. The table reports AP, AP at 50%, and AP at 50m and 100m. All values are expressed as percentages (%), where higher values indicate better performance.

	KITTI		BDD		A2D2	
	AP	AP 50%	AP	AP 50%	AP	AP 50%
MobileNet Variant						
CE	23.5	39.1	19.5	37.5	17.5	32.7
Prior	22.1	38.7	20.2	38.1	17.2	31.9
Evidential	22.9	40.6	22.3	41.6	18.4	34.1
Dirichlet MLE + ILVI	23.8	42.1	23.8	43.2	21.5	37.3
EfficientNet Variant						
CE	22.4	43.2	24.7	46.6	19.2	36.3
Prior	22.2	43.5	25.4	47.4	20.3	37.1
Evidential	24.4	50.7	26.3	48.1	20.8	37.6
Dirichlet MLE + ILVI	25.6	51.3	28.5	50.8	23.4	39.4

Table 4.7: Instance segmentation performance (%) of the evaluated methods on generalization datasets (KITTI, BDD, and A2D2) for MobileNet and EfficientNet architectures. The table reports AP and AP at 50%. All values are expressed as percentages (%), where higher values indicate better performance.

Lastly, although upgrading to a more powerful backbone such as EfficientNet enhances the performance of all methods, Dirichlet MLE + ILVI consistently maintains a clear performance advantage. This indicates that its benefits scale effectively with increased feature extraction capacity and remain robust even as overall model complexity grows.

Figure 4.7: Sample instance segmentation results from the Cityscapes dataset using the Dirichlet MLE + ILVI model with the EfficientNet backbone. The visualization presents color-coded masks for pedestrians and vehicles across diverse urban scenes, demonstrating the model's segmentation performance. Different colors distinguish object instances, showcasing the model's ability to identify and separate individual objects within complex environments.

4.3.2 Distributional Separation Performance

This section utilizes distributional separation to evaluate two key aspects: how well the models differentiate between correct and incorrect predictions in the ID (Cityscapes) dataset, and how well they detect OOD objects in the OOD dataset (Lost and Found). This dual evaluation provides insights into the models' ability to leverage uncertainty estimation to address prediction errors in ID scenarios and adapt to distribution shifts in OOD scenarios. The separation plots are shown in Figure 4.8 and their respective Wasserstein distance values are reported in Table 4.8.

Dirichlet MLE + ILVI provides two outputs, an uncertainty-based metric and a Dirichlet strength metric. The uncertainty-based metric captures ID performance more effectively, while the Dirichlet strength metric provides clearer signals for OOD detection. Combining these two metrics offers additional insight into whether predictions are likely to be correct and whether an object may belong to an unknown class.

Figure 4.8 visualizes the distributional separation performance of various models across both backbone architectures. Each row corresponds to a model, while the columns are grouped by backbone. Within each backbone, the left column displays results on the ID dataset, Cityscapes, where predictions are classified as correct or incorrect. The right column shows results on the OOD dataset, Lost and Found, highlighting the separation between ID and OOD predictions.

An ideal scenario for maximizing distributional separation, and thus achieving a high Wasserstein distance, places correct and incorrect predictions at opposite ends of the certainty spectrum. In this case, correct predictions form a distinct high-certainty cluster, while incorrect predictions form a low-certainty cluster, resulting in a large Wasserstein distance and clear discrimination between the two categories. This principle similarly applies to the separation of ID and OOD data, where ID predictions should ideally exhibit high certainty, and OOD predictions low certainty. In the results, while most models consistently assign high certainty to correct predictions, not all models assign correspondingly low certainty to incorrect or OOD predictions. This leads to overlapping certainty distributions and consequently lower Wasserstein distance values.

Dirichlet MLE + ILVI demonstrates superior separation by producing distributions with reduced overlap in the ID setting, more clearly distinguishing correct

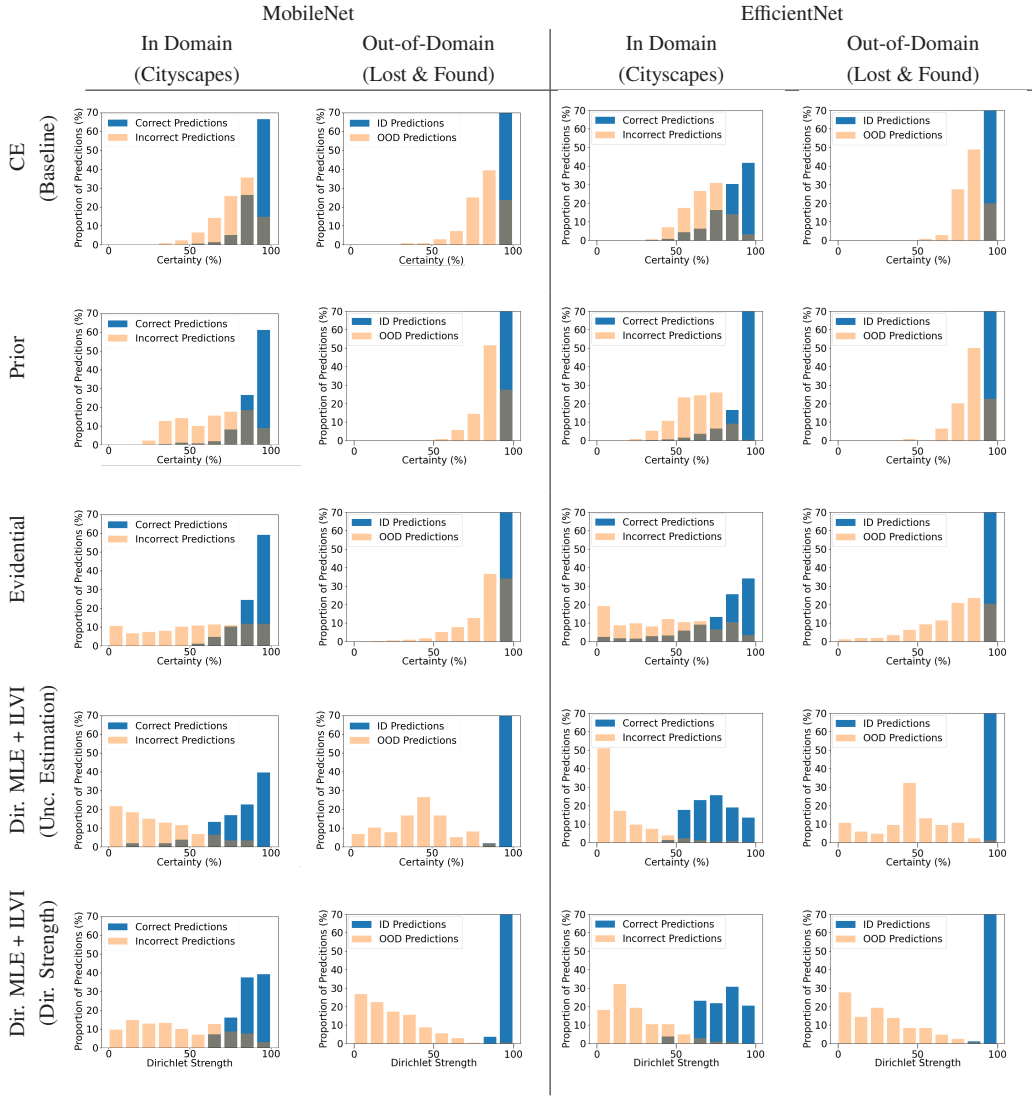


Figure 4.8: Visualization of distributional separation performance at the instance level across both backbones for all evaluated models on ID (Cityscapes) and OOD (Lost and Found) datasets. Within each backbone, the left set of plots compares the certainty distributions of correct versus incorrect predictions on the ID dataset, while the right set compares the certainty distributions of ID versus OOD predicted instances. This highlights each model's ability to separate these distributions under both evaluation settings.

	Wasserstein Distance (↑)	
	In-Domain (Cityscapes)	Out-of-Domain (Lost and Found)
MobileNet Variant		
CE	1.1	1.3
Prior	2.3	1.1
Evidential	3.3	0.9
Dirichlet MLE + ILVI (Unc. Estimation)	**4.9**	5.2
Dirichlet MLE + ILVI (Dir. Strength)	4.2	**7.1**
EfficientNet Variant		
CE	1.59	1.16
Prior	2.67	1.29
Evidential	3.36	2.27
Dirichlet MLE + ILVI (Unc. Estimation)	**5.81**	5.15
Dirichlet MLE + ILVI (Dir. Strength)	5.38	**6.78**

Table 4.8: Wasserstein distances representing the distributional separation at the instance level between correct and incorrect predictions across two architectures (MobileNet and EfficientNet) for all evaluated models on ID (Cityscapes) and OOD (Lost and Found) datasets.

from incorrect predictions. A similar observation is made in the OOD setting, where Dirichlet MLE + ILVI achieves more distinct separation between ID and OOD samples across both backbone architectures, outperforming other models in minimizing distributional overlap.

Table 4.8 shows that the Dirichlet MLE + ILVI model achieves the highest performance, as evidenced by its higher Wasserstein distances compared to all other approaches. This indicates that certainty estimated through this method aligns closely with the model's certainty in the ID setting, leading to a stronger separation of distributions for correct and incorrect predictions. In contrast, leveraging the Dirichlet strength of the Dirichlet MLE + ILVI approach shows significant effectiveness in detecting OOD objects, outperforming both the baseline and SOTA methods. This approach achieves higher Wasserstein distances, capturing the distributional differences between ID and OOD predictions under data distribution shifts. This underscores the role of Dirichlet strength as a reliable metric for detecting OOD objects.

The CE model fails to adequately separate the distributions, resulting in significantly lower Wasserstein distances across both backbones and for both ID and OOD data. The Prior and Evidential methods exhibit intermediate performance, with Evidential performing particularly poorly in the OOD scenario, as reflected by its low Wasserstein distance. These results underscore the limitations of these methods in generalizing effectively across ID and OOD datasets.

The findings emphasize the complementary strengths of utilizing uncertainty estimation and Dirichlet strength of the Dirichlet MLE + ILVI in addressing different challenges. While uncertainty estimation of Dirichlet MLE + ILVI excels in ID settings, the Dirichlet Strength demonstrates its effectiveness in detecting OOD objects. Together, both representations show a significant advancement over baseline and SOTA approaches, offering a versatile model functioning across diverse data distributions.

4.3.3 Reduction of False Positive Detections

Building on the strong distributional separation achieved by Dirichlet MLE + ILVI in distinguishing correct from incorrect predictions, this section explores reducing false positive detections in instance segmentation through uncertainty-based thresholding. By filtering out low-certainty predictions, thresholding improves precision by discarding uncertain instances, though potentially at the expense of recall. The key challenge lies in selecting an optimal threshold that effectively reduces false positives without significantly impacting true positive detections [HBGS23b].

This section examines the impact of varying uncertainty thresholds on precision and recall for MobileNet Dirichlet MLE + ILVI, particularly for detections within 50 meters. The analysis aims to identify an optimal threshold that improves precision while managing the trade-off with recall, ensuring the model's reliability in challenging detection scenarios.

To achieve a balanced and reliable performance, precision and recall are analyzed across different uncertainty levels. Figure 4.9 visualizes this relationship, with thresholds increasing in 10% increments. Additionally, sample results illustrating the reduction of false positives through thresholding are presented in Figure 4.10.

The results demonstrate that as the threshold increases, precision improves due to the effective removal of false positives. Recall remains stable up to a certain threshold but begins to decline thereafter, as some true positives are excluded. In this experiment, a threshold of 60% achieves the best accepted trade-off, significantly boosting precision while maintaining recall at an acceptable level.

The quantitative results in Table 4.9 highlight the benefits of applying thresholding. Dirichlet MLE + ILVI with thresholding achieves the highest AP of 57.2% and AP within 50 meters of 54.5%, representing substantial improvements over other methods and the non-thresholded Dirichlet MLE approach. This indicates the model's ability to assign high certainty to true positives and low certainty to false positives, effectively separating the two categories. However, as expected, recall slightly decreases after thresholding, with values of 37.5% (overall) and 54.2% (within 50 meters), which are slightly lower than the non-thresholded Dirichlet MLE (39.4% and 55.3%, respectively), but still

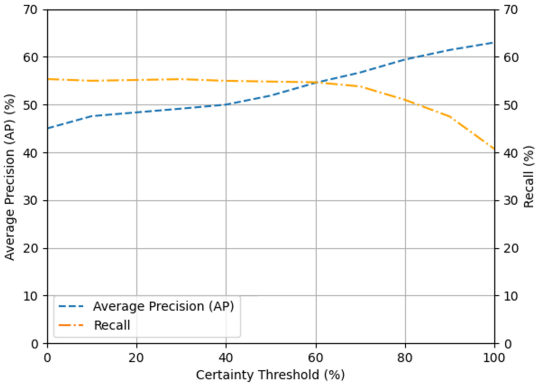


Figure 4.9: Average precision and recall values up to 50 meters are plotted with varying certainty thresholds.

	AP	AP 50m	Recall	Recall 50m
CE	38.2	35.8	31.8	45.6
Prior	39.8	37.5	34.5	48.9
Evidential	42.9	39.8	37.1	51.3
Dirichlet MLE + ILVI	45.1	43.5	**39.4**	**55.3**
Dirichlet MLE + ILVI$_{(Thresholded)}$	**57.2**	**54.5**	37.5	54.2

Table 4.9: Comparison of precision and recall performance (in % ↑) across different methods, with and without thresholding, for both overall detections (AP, Recall) and detections within 50 meters (AP 50m, Recall 50m). Thresholding is applied only to the Dirichlet MLE + ILVI.

higher than the SOTA models. This reflects the inherent trade-off introduced by thresholding.

Overall, the results emphasize the use of thresholding as a practical technique to enhance precision by reducing false positive detections. The improvement in AP demonstrates the model's capacity to differentiate between true and false positive detections effectively.

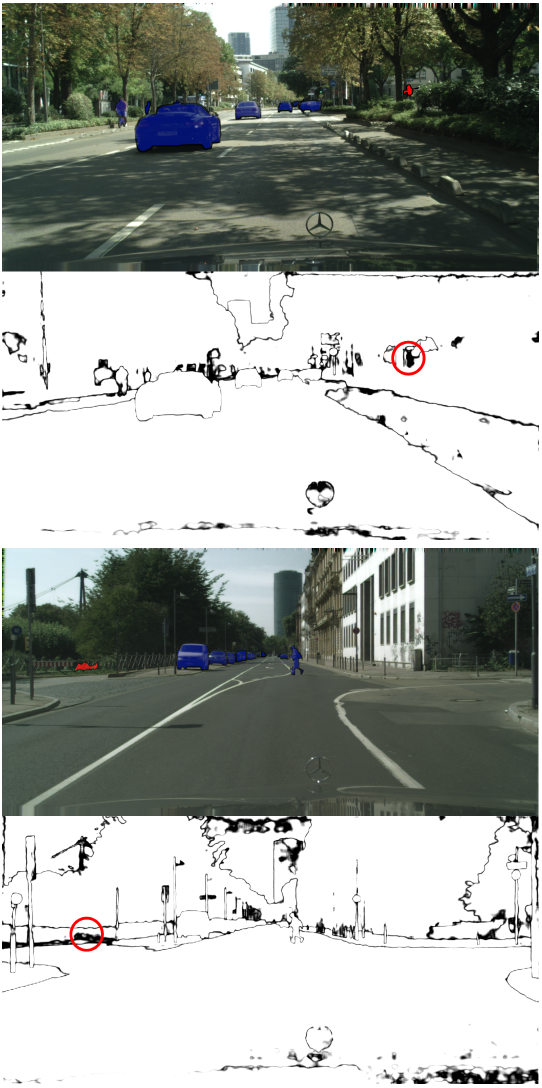


Figure 4.10: This figure presents two samples of instance segmentation results, where the top row displays the segmentation, and the bottom row illustrates the corresponding uncertainty estimates. True positive detections are highlighted in blue, representing high certainty, while false positive detections are marked in red, indicating low certainty. To emphasize false positive detections, a red circle is drawn around them in the uncertainty estimate. Adapted from Figure 5 in [HBGS23b] ©Uni-DAS.

4.3.4 Depth Estimation

In this section an investigation of the depth estimation performance of the models is presented. The evaluation metric, RSE, is computed only for detected instances, ensuring that missed detections do not influence the analysis. Table 4.10 presents the results, while Figure 4.11 provides sample visualizations.

The depth ranges analyzed include 0–15m, 15–25m, 25–50m, 50–100m, and an overall metric calculated for the entire range of 0–100m. This segmentation allows for a more granular evaluation of model performance across different distance intervals, capturing variations in depth estimation accuracy. Closer ranges (0–15m) typically provide higher confidence due to better visual resolution and richer feature details, while farther ranges (50–100m) introduce more uncertainty due to occlusions, reduced pixel density, and sensor limitations. By analyzing performance across these intervals, the investigation can assess how well the model generalizes across varying depths and whether uncertainty increases with distance.

The model is designed to estimate depth exclusively for objects within the scene, assigning a depth value to each detected and segmented instance. Figure 4.11 presents examples of depth estimation outputs. For visualization, predicted depth values are linearly scaled to the interval $[0, 150]$ m and mapped to a custom blue colormap, where darker blue corresponds to near distances and lighter blue indicates farther distances. Black denotes regions for which the model was not trained to estimate depth. The corresponding depth colormap is shown in Figure 4.12.

Notably, in the top sample image, the front-row pedestrians exhibit nearly uniform depth, as indicated by the consistent color shading. In contrast, the bottom sample image displays a gradual color transition, reflecting the model's ability to assign depth estimates based on the relative distance of objects from the ego vehicle.

Table 4.10 presents the RSE results across different depth ranges, evaluating the performance of the methods on both MobileNet and EfficientNet backbones. As expected, error values increase with distance reflecting the inherent challenges of long-range depth estimation due to reduced object resolution.

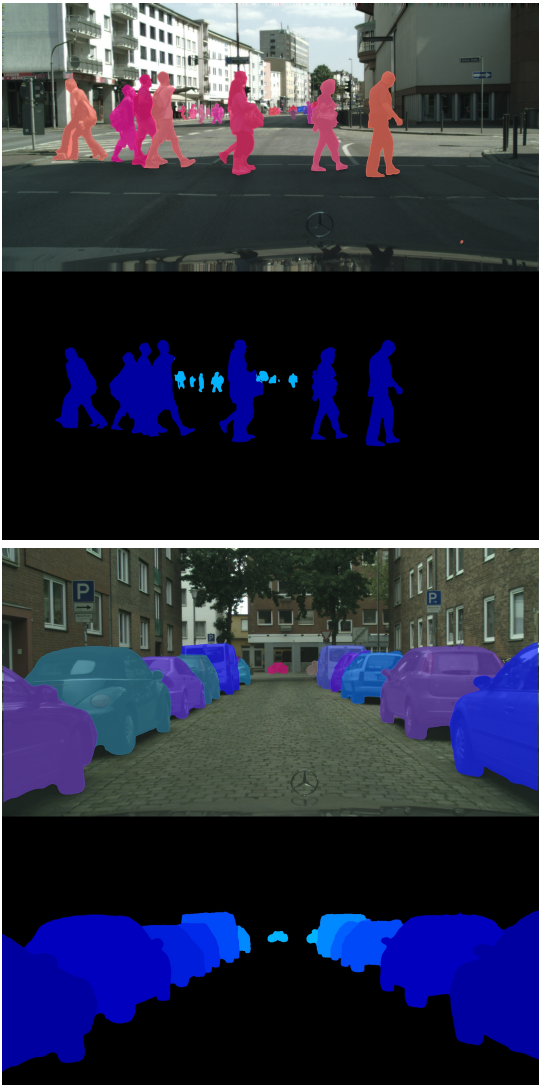


Figure 4.11: This figure shows two samples of instance segmentation and their respective depth estimation results obtained using the EfficientNet Dirichlet MLE + ILVI model. In each sample, the top image illustrates the instance segmentation, while the bottom image represents the corresponding depth estimates. The depth is color-coded such that lighter shades indicate objects farther from the ego vehicle. The first example highlights pedestrian segmentation and depth estimation, whereas the second example focuses on parked cars along a street.

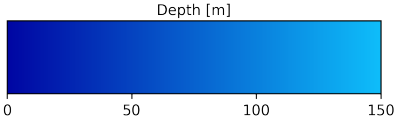


Figure 4.12: Depth is visualized using a custom blue colormap ranging from dark blue (near, 0 m) to light blue (far, 150 m). Black denotes regions for which the model was not trained to estimate depth.

	Relative Squared Error (%)(↓)				
	0-15m	15-25m	25m-50m	50-100m	0-100m
MobileNet Variant					
CE	2.90	3.17	4.49	6.28	4.60
Prior	3.11	3.19	3.17	4.17	3.49
Evidential	2.98	2.64	2.67	3.80	3.04
Dirichlet MLE + ILVI	2.37	2.48	2.53	3.39	3.01
EfficientNet Variant					
CE	2.79	2.86	3.89	5.38	4.06
Prior	2.80	2.63	3.12	3.15	2.82
Evidential	2.71	2.57	2.78	3.39	2.94
Dirichlet MLE + ILVI	2.15	2.25	2.42	3.24	2.78

Table 4.10: Relative squared error (%) for depth estimation across various methods and architectures (MobileNet and EfficientNet) evaluated over multiple depth ranges (0–15m, 15–25m, 25–50m, 50–100m, and 0–100m). Results are calculated only for detected instances, focusing on the accuracy of depth predictions. Lower values indicate better performance.

Among the methods, Dirichlet MLE + ILVI consistently achieves the lowest RSE values across all depth intervals, indicating superior depth prediction accuracy compared to other approaches. The Evidential model also performs well but slightly lags behind Dirichlet MLE + ILVI in most cases. Meanwhile, the Prior model generally exhibits higher errors than Dirichlet MLE + ILVI but remains competitive with the Evidential approach in certain scenarios.

A comparison between architectures reveals that EfficientNet consistently outperforms MobileNet across all methods and depth intervals, demonstrating a stronger ability to extract and represent features. This is particularly noticeable as the depth range increases the difference in error increases.

For the MobileNet architecture, Dirichlet MLE + ILVI demonstrates the most significant improvements over CE, with error reductions ranging from 18% to 46% across all depth ranges. At farther distances (50–100m), it reduces CE's peak error of 6.28% by 46%. Similarly, in the EfficientNet architecture, Dirichlet MLE + ILVI continues to outperform CE, achieving improvements of 21% to 40% across all depth intervals, with a notable 40% reduction in CE's highest error at 50–100m.

Overall, these results highlight the effectiveness of Dirichlet MLE + ILVI in providing more reliable depth estimates while demonstrating the benefits of using a stronger feature extraction backbone such as EfficientNet.

4.3.5 OOD Instance Segmentation and Identification Performance

In this section, the ability of the models to detect and identify OOD objects is evaluated using three metrics: AP $50\%_{\text{OOD}}$, $\text{Recall}_{\text{OOD}}$, and iOOD. AP $50\%_{\text{OOD}}$ measures the balance between precision and recall at an IoU threshold of 50% and $\text{Recall}_{\text{OOD}}$ quantifies the fraction of OOD objects correctly detected. The iOOD metric quantifies a model's ability in recognizing OOD objects by evaluating how confidently detected objects are identified as OOD. In each case, lower values indicate better identification of OOD objects [HBGS23a].

Figure 4.13 shows an example from the Lost and Found dataset that contains two boxes the model should identify as OOD. Each row corresponds to a model, where on the left the model's instance segmentation is presented, and on the right its corresponding uncertainty estimation. The last row of Dirichlet MLE + ILVI shows both the uncertainty estimate and the Dirichlet strength.

The baseline model fails to detect both boxes, producing no instance segmentation for these items and indicating low or no uncertainty in those regions. The Prior model partially detects both boxes, showing some uncertainty. The Evidential model detects one box completely and partially detects the other, with uncertainty levels higher than Prior.

In comparison, Dirichlet MLE + ILVI detects both boxes and assigns high uncertainty to them. The Dirichlet strength map reveals also low strength values for the boxes, indicating that the model interprets them as OOD rather than simply uncertain ID objects. This highlights how the uncertainty estimate and Dirichlet strength outputs complement each other, allowing Dirichlet MLE + ILVI to better distinguish between ID objects with high uncertainty and genuine OOD objects.

Table 4.11 presents the results for both MobileNet and EfficientNet, where the Dirichlet strength was used to calculate the iOOD for the Dirichlet MLE + ILVI models. Dirichlet MLE + ILVI provides large gains over CE, nearly doubling AP $50\%_{\text{OOD}}$ (from 16.2% to 35.2% on EfficientNet) and $\text{Recall}_{\text{OOD}}$ (from 17.1% to 35.8% on EfficientNet). Its iOOD scores increase by almost 6 folds on MobileNet (8.9% to 61.7%) and more than 5 folds on EfficientNet (11.2% to 69.1%). Dirichlet MLE + ILVI improves over Evidential by approximately

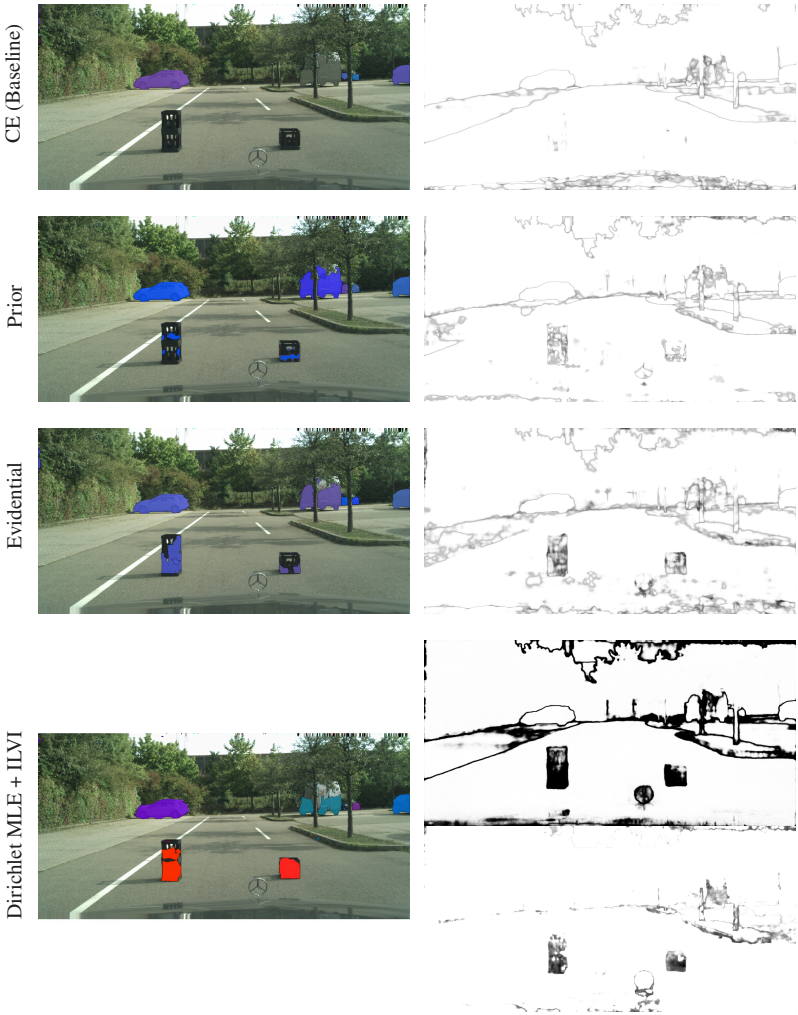


Figure 4.13: The figure illustrates the outputs of the models on a sample image with OOD objects for the Lost and Found dataset. For each model, the instance segmentation is shown (left) and its respective uncertainty estimation map (right). For Dirichlet MLE + ILVI (last row) shows the uncertainty estimate (top) and Dirichlet strength (bottom).

	AP 50%$_{\text{OOD}}$ (↑)	Recall$_{\text{OOD}}$ (↑)	iOOD (↑)
MobileNet			
CE	13.7	15.3	8.9
Prior	14.2	15.9	12.1
Evidential	17.6	28.2	16.2
Dirichlet MLE + ILVI	26.2	30.4	61.7
EfficientNet			
CE	16.2	17.1	11.2
Prior	16.9	17.3	14.1
Evidential	19.1	32.1	18.6
Dirichlet MLE + ILVI	35.2	35.8	69.1

Table 4.11: OOD detection metrics (AP 50%$_{\text{OOD}}$, Recall$_{\text{OOD}}$, and iOOD) for the methods evaluated on both architectures. AP 50%$_{\text{OOD}}$ represents the average precision at a 50% IoU threshold for OOD objects, Recall$_{\text{OOD}}$ measures the fraction of OOD instances correctly detected, and iOOD indicates the proportion of those detections assigned high uncertainty.

48% in AP 50%$_{\text{OOD}}$ (from 17.6% for Evidential to 26.2%) and by about 7.8% in Recall$_{\text{OOD}}$ (from 28.2% for Evidential to 30.4%). When compared to the Prior method, Dirichlet MLE + ILVI shows a significant increase of around 84% in AP 50%$_{\text{OOD}}$ (from 14.2% for Prior to 26.2%) and an improvement of approximately 14.5% in Recall$_{\text{OOD}}$ (from 15.9% for Prior to 30.4%). Additionally, Dirichlet MLE + ILVI achieves an iOOD score that surpasses both Evidential and Prior methods.

Across both architectures, Dirichlet MLE + ILVI demonstrates the most robust balance of precision, recall, and iOOD. Although EfficientNet improves baseline performance relative to MobileNet (CE's AP 50%$_{\text{OOD}}$ rises from 13.7% to 16.2%), Dirichlet MLE + ILVI's gains scale accordingly, resulting in the highest margins on stronger architectures.

A comparison of the methods shows that relying on recall alone does not fully describe OOD detection performance. Evidential matches or approaches Dirichlet MLE + ILVI in recall but falls short in AP 50%, indicating an elevated number of false positives.

In addition to AP 50%$_{\text{OOD}}$ and Recall$_{\text{OOD}}$, the iOOD metric offers insight into how confidently models label detected objects as OOD. The Dirichlet MLE

+ ILVI model consistently achieves the highest iOOD scores, reaching 61.7% on MobileNet and 69.1% on EfficientNet. These values show that Dirichlet strength used as a metric is efficient in detecting OOD objects. In contrast, CE yields iOOD values in the range of 8.9% to 11.2%, indicating less reliable certainty estimates for OOD objects.

4.3.6 Analysis of ROC Curves for OOD Detection

To assess the effectiveness of each model in detecting OOD inputs, ROC curves are generated and analyzed. These curves illustrate the trade-off between the TPR and the FPR across varying classification thresholds, providing a comprehensive view of each method's ability to discriminate between ID and OOD samples.

In the context of OOD detection, the TPR measures the model's ability to correctly identify OOD samples. It is defined as the proportion of actual OOD samples that are correctly detected as OOD. A higher TPR indicates that the model is more sensitive to detecting OOD inputs. However, the FPR measures the proportion of ID objects that are mistakenly classified as OOD. A lower FPR indicates the model's effectiveness in minimizing false alarms by correctly identifying ID objects.

Figure 4.14 presents the AUROC curves. The Dirichlet MLE + ILVI consistently outperforms the Evidential, Prior, and CE methods for both MobileNet and EfficientNet. While Evidential provides a notable improvement over CE and Prior, Dirichlet MLE + ILVI achieves the highest discrimination, showing a higher TPR and lower FPR simultaneously.

On both architectures, Dirichlet MLE + ILVI produces ROC curves that approach the top-left corner, indicating a high TPR with minimal FPR. This shows that Dirichlet MLE + ILVI is significantly more effective at differentiating between ID and OOD samples while avoiding false positives. Compared with all other methods, it also achieves a higher AUROC.

Moreover, each graph has a red line showing the FPR value at 95% TPR. It is a critical metric for evaluating how well a model can minimize false alarms while maintaining high OOD detection sensitivity. AUROC and FPR@95% TPR values are summarized in Table 4.12. A higher AUROC signifies stronger OOD discrimination, whereas a lower FPR at 95% TPR reflects fewer false positives when capturing 95% of all positives.

Dirichlet MLE + ILVI's greatest improvement is seen in the difference between CE and Dirichlet MLE + ILVI. With MobileNet, Dirichlet MLE + ILVI raises AUROC by approximately 11% while reducing the FPR@95% TPR from over 50% to 16.3%. Similar observations are seen with EfficientNet, where

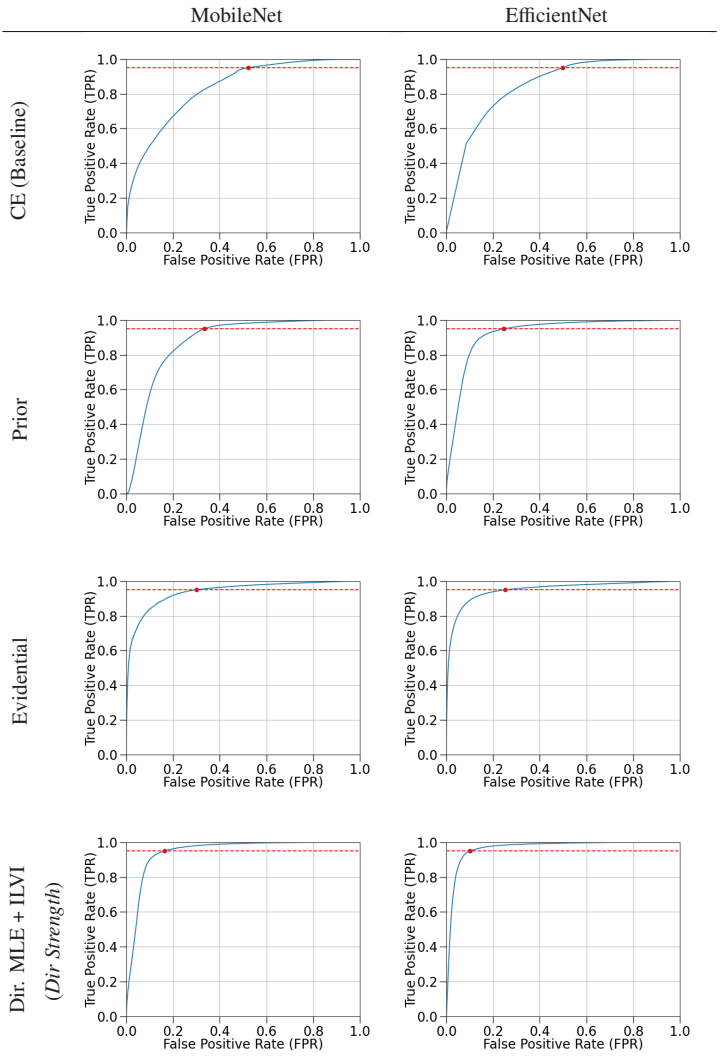


Figure 4.14: ROC curves for OOD detection performance across both backbones for all models are shown here. The blue curve represents the trade-off between the TPR and FPR at various thresholds. The red dashed line corresponds to a TPR of 95%, where the FPR is recorded (FPR @ 95% TPR) to assess the model's ability to minimize false positives while maintaining high sensitivity. A steeper blue curve and lower FPR at the red line indicate superior OOD detection capabilities.

	MobileNet		EfficientNet	
	AUROC (↑)	FPR @ 95% TPR (↓)	AUROC (↑)	FPR @ 95% TPR (↓)
CE	83.4	52.2	84.7	49.7
Prior	87.6	33.3	92.4	24.5
Evidential	93.4	30.1	94.9	25.3
Dirichlet MLE (Dir. Strength)	94.5	16.3	96.7	10.2

Table 4.12: The table reports the AUROC, indicating the overall OOD detection capability, and the FPR at a TPR of 95%. Higher AUROC and lower FPR values signify better OOD detection performance.

Dirichlet MLE + ILVI further lowers the FPR@95% TPR by approximately 15% compared to other methods.

Overall, Dirichlet MLE + ILVI achieves an optimal balance between high sensitivity (reflected by AUROC) and low false positive rates, making it the most robust model for OOD detection among the tested models. Its consistency across both MobileNet and EfficientNet further demonstrates its versatility for various backbone architectures.

4.3.7 Discussion

This section presented a comprehensive evaluation of the proposed methodology, centered around the Dirichlet MLE + ILVI method, for addressing critical challenges in automated driving perception: instance segmentation, OOD detection, and depth estimation. The experiments rigorously assessed the model's performance across a range of dimensions and datasets, comparing it against established baseline (CE) and state-of-the-art (Prior and Evidential) models, aiming to provide a clear understanding of the proposed architecture's strengths, limitations, and potential contributions to safer and more reliable automated driving systems.

From the perspective of automated driving systems, these findings have direct practical implications. In instance segmentation, Dirichlet MLE + ILVI achieved superior AP at various thresholds (e.g., AP@50m and AP@100m), enabling reliable detection of key objects, such as vehicles, pedestrians, and obstacles, even at extended distances. This robustness in long-range perception is essential for safe navigation, early obstacle recognition, and safe path planning. Moreover, the method's strong OOD detection capability allows it to identify unfamiliar or novel objects with high certainty, which is crucial for handling unexpected objects. By assigning higher uncertainty or lower Dirichlet strength to such objects, the model can flag potential risks for further inspection or cautious maneuvering.

The distributional separation analysis further highlighted Dirichlet MLE + ILVI's ability to distinguish between correct and incorrect predictions for ID data, and to separate ID from OOD instances. Compared to baseline and the SOTA methods, Dirichlet MLE + ILVI's uncertainty and Dirichlet strength metrics showed significantly reduced overlap between correct and incorrect (or ID and OOD) distributions.

While uncertainty-based thresholding has proven effective in reducing false positives (FPs), it is important to distinguish the factors influencing different types of detection errors. False negatives are primarily governed by the segmentation model's representational capacity and its ability to accurately localize and classify objects. Reducing FNs typically necessitates architectural enhancements or improved training procedures aimed at boosting segmentation performance. In contrast, the mitigation of FPs relies heavily on the availability of well-calibrated uncertainty estimates, which enable the model to assign low

certainty to ambiguous or spurious predictions. The strength of Dirichlet MLE + ILVI lies in its ability to address both error types: achieving high segmentation accuracy to reduce FNs and leveraging reliable uncertainty quantification to filter out FPs. This dual capability is particularly valuable in safety-critical applications such as automated driving, where both missed detections and erroneous activations can have significant operational consequences.

In addition to these advances in segmentation and OOD detection, Dirichlet MLE + ILVI also delivered consistently better depth estimation results. Although depth estimation is a separate regression task, training the model with Dirichlet MLE + ILVI evidently enhances feature extraction, as reflected by lower RSE values. This improvement translates into more accurate distance measurements.

The OOD detection and identification results demonstrate that Dirichlet MLE + ILVI effectively distinguishes ID objects from unexpected items, an essential requirement for safe automated driving. Across metrics such as AP_{OOD}, $Recall_{OOD}$, and iOOD, Dirichlet MLE + ILVI outperforms the other models, consistently attaining higher recall on truly unfamiliar objects. In particular, its iOOD scores confirm that it more accurately identifies objects lying outside the training distribution, underscoring its strength in separating novel instances from familiar classes.

For automated driving, failing to detect OOD objects, such as debris or novel objects, can lead to unsafe decisions, while misclassifying common objects as OOD may prompt unnecessary maneuvers. By striking a balance between true positives and false positives, Dirichlet MLE + ILVI reduces the risk of missing genuine unknown hazards while avoiding frequent false alarms. This balance minimizes both missed detections and excessive caution, ultimately supporting safer navigation.

5 Conclusion and Future Work

This thesis proposes a unified architectural framework for enhancing perception in ADS by integrating uncertainty estimation, OOD detection, and depth estimation into semantic and instance segmentation tasks. The approach addresses known limitations of conventional DNNs, which often produce overconfident predictions without providing a measure of uncertainty or the ability to recognize inputs outside the training distribution. Such limitations are especially significant in ADS, where incorrect predictions made with high confidence can negatively affect downstream modules and compromise system safety.

The proposed framework offers a principled and computationally efficient solution for incorporating uncertainty and OOD awareness into a single architecture. It improves both the predictive performance of segmentation and depth estimation tasks and the reliability of their associated confidence estimates. By jointly modeling uncertainty and output structure in a real-time-compatible design, this work contributes a step toward more dependable perception systems in open-world driving scenarios.

The proposed architecture is based on a shared feature extraction backbone, followed by three task-specific decoders for semantic segmentation, instance segmentation, and monocular depth estimation. This multi-task setup ensures computational efficiency and consistency across outputs while allowing specialization in each task. A core innovation in this work is the introduction of an ILVI module, which introduces stochasticity into the internal feature representations of the network. Sampling is performed exclusively from the ILVI layer, and only during inference, enabling the estimation of epistemic uncertainty without modifying the structure or execution path of the downstream decoders. This approach allows for variability in feature-level representations while preserving compatibility with real-time inference constraints.

To represent prediction uncertainty at the output level, the conventional softmax function is replaced with a Dirichlet output layer. This layer models output

class probabilities as parameters of a Dirichlet distribution, allowing uncertainty to be inferred directly from the distribution's concentration parameters. Unlike point-estimate confidence scores, the Dirichlet formulation captures the entire predictive distribution and provides a principled basis for uncertainty estimation and OOD detection. In particular, the Dirichlet strength serves as a quantitative measure of certainty, with lower values indicating potential OOD inputs.

Comprehensive experiments are conducted using publicly available datasets, and the performance of the proposed approach is evaluated across several dimensions: segmentation accuracy (mIoU, AP), depth estimation quality (RSE), calibration performance (ECE, MCE), and OOD detection capability (AUROC, FPR@95%TPR). The combined Dirichlet MLE + ILVI approach consistently outperforms standard baselines in both segmentation and uncertainty-aware metrics. Improved calibration indicates that the model's predicted confidence aligns more closely with its empirical accuracy, reducing the risk of overconfident misclassifications. Additionally, distributional separation analyses using the Wasserstein distance confirm that the model assigns distinguishable uncertainty values to correct versus incorrect predictions, enabling more informed downstream decisions.

A key result of this work is the demonstration that Dirichlet strength provides a reliable signal for distinguishing between ID and OOD inputs. Quantitative evaluations show that this method achieves higher AUROC scores and lower FPR values than existing approaches, including softmax-based entropy or confidence thresholds. Thresholding based on Dirichlet strength is also shown to enhance instance segmentation precision by filtering low-certainty predictions, particularly in close-range scenarios where high accuracy is critical for safety-critical actions such as braking or avoidance.

Another contribution lies in the analysis of how uncertainty estimation complements segmentation accuracy. High segmentation quality primarily reduces false negatives by improving localization and classification, while well-calibrated uncertainty estimates reduce false positives by suppressing low-confidence outputs. The joint application of ILVI and Dirichlet layers addresses both types of errors, resulting in more reliable and stable perception outputs. This combination enhances not only model performance in quantitative terms but also the operational robustness of the system under varying and unpredictable environmental conditions.

Finally, this thesis demonstrates that it is possible to incorporate probabilistic modeling techniques into real-time ADS perception pipelines without introducing significant inference-time computational overhead. The ILVI module enables uncertainty sampling from a compact latent space, while the Dirichlet output layer replaces the standard classifier without architectural complexity. The model is modular, task-agnostic, and compatible with embedded hardware constraints, offering a practical pathway toward uncertainty-aware perception in real-world autonomous systems.

Future Work

Future work may extend this research in several ways. One area involves multi-sensor fusion by integrating LiDAR, RADAR, or depth cameras into the shared backbone. This integration may combine confidence measures across sensors and improve uncertainty estimation and OOD detection. Another area involves temporal consistency by moving from single-frame to video-based segmentation to address issues arising from moving objects and dynamic environments.

Another direction is adaptive thresholding. Instead of using fixed thresholds to filter false positives, a learned or adaptive mechanism may adjust thresholds based on scene complexity and object classes. This approach may balance precision and recall more effectively in instance detection.

Further work should address the interpretability of uncertainty regions. It is necessary to determine the causes of uncertain predictions, such as sensor noise or occlusions. Integrating the architecture with explainability methods may offer insight into segmentation outputs and support oversight and certification processes in autonomous driving.

Bibliography

[AECR24] F. Ataeiasad, D. Elizondo, and S. Calderón Ramírez. Out-of-distribution detection with memory-augmented variational autoencoder. *Mathematics*, 2024.

[APH+21] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

[Bar19] Jonathan T Barron. A general and adaptive robust loss function. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4331–4339, 2019.

[BB16] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.

[BKC17] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[BKM17] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[BN06] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[BSD23] Kai Brach, Beate Sick, and Oliver Dürr. Single-shot bayesian approximation for neural networks. *arXiv preprint arXiv:2308.12785*, 2023.

[BSN+21] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *International Journal of Computer Vision*, 129(11):3064–3079, 2021.

[Buc23] S. Bucci. *Visual Domain Generalization via Self-Supervised Learning*. PhD thesis, University of Rome, 2023.

[CCZ+20] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020.

[CFT19] K. Ciosek, V. Fortuin, and R. Tomioka. Conservative uncertainty estimation by fitting prior networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[CKNH20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[CLT+21] Charles Corbière, Marc Lafon, Nicolas Thome, Matthieu Cord, and Patrick Pérez. Beyond first-order uncertainty estimation with evidential models for open-world recognition. In *ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning*, 2021.

[CMS+20] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[COR+16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[CPK+17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Rethinking atrous con-

volution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[Cra20] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020.

[CSG22] K. Chauhan, P. Shenoy, and M. Gupta. Robust outlier detection by de-biasing vae likelihoods. *CVPR*, 2022.

[CTS+21] Charles Corbiere, Nicolas Thome, Antoine Saporta, Tuan-Hung Vu, Matthieu Cord, and Patrick Perez. Confidence estimation via auxiliary models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6043–6055, 2021.

[CVC+22] Gabriela Csurka, Riccardo Volpi, Boris Chidlovskii, et al. Semantic image segmentation: Two decades of research. *Foundations and Trends® in Computer Graphics and Vision*, 14(1-2):1–162, 2022.

[CZG20] Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in Neural Information Processing Systems*, 33:1356–1367, 2020.

[CZP+18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[DCYL23] D. Deng, G. Chen, Y. Yu, and F. Liu. Uncertainty estimation by fisher information-based evidential deep learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[Doe16] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[EDHH24] D.M. El-Din, A.E. Hassanein, and E.E. Hassanien. An adaptive and late multifusion framework in contextual representation based on evidential deep learning and dempster-shafer theory. *Knowledge and Information Systems*, 2024.

[EEVG+15] Mark Everingham, SMAli Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual

object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015.

[FBLT20] Andrew Foong, David Burt, Yingzhen Li, and Richard Turner. On the expressiveness of approximate inference in bayesian neural networks. *Advances in Neural Information Processing Systems*, 33:15897–15908, 2020.

[FHSR+20] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020.

[FLR+19] Cheng-Yang Fu, Wei Liu, Aayush Ranga, Amit Tyagi, and Alexander C Berg. Retinamask: Learning to predict masks improves state-of-the-art single-shot detection for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4067–4076, 2019.

[Gal16] Yarin Gal. Uncertainty in deep learning. 2016.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[GDS20] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 318–319, 2020.

[GG16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016.

[GHC20] C. Geng, S. Huang, and S. Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[GHM+22] Mudasir A Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022.

[GKM+20] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020.

[GLSU13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[GPSW17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[GSK18] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. *International Conference on Learning Representations (ICLR)*, 2018.

[GTA+21] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

[GTA+23] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

[HAA+22] Sebastian Houben, Stephanie Abrecht, Maram Akila, Andreas Bär, Felix Brockherde, Patrick Feifel, Tim Fingscheidt, Sujan Sai Gannamaneni, Seyed Eghbal Ghobadi, Ahmed Hammam, et al. Inspect, understand, overcome: A survey of practical methods for ai safety. In *Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights Towards Safety*, pages 3–78. Springer International Publishing Cham, 2022.

[HBGS22a] Ahmed Hammam, Frank Bonarens, Seyed Eghbal Ghobadi, and Christoph Stiller. Predictive uncertainty quantification of deep neural networks using dirichlet distributions. In *Proceedings of*

the 6th ACM Computer Science in Cars Symposium, pages 1–10, 2022.

[HBGS22b] Ahmed Hammam, Frank Bonarens, Seyed Eghbal Ghobadi, and Christoph Stiller. Towards improved intermediate layer variational inference for uncertainty estimation. In *European Conference on Computer Vision*, pages 1–14. Springer, 2022.

[HBGS23a] Ahmed Hammam, Frank Bonarens, Seyed Eghbal Ghobadi, and Christoph Stiller. Identifying out-of-domain objects with dirichlet deep neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4560–4569, 2023.

[HBGS23b] Ahmed Hammam, Frank Bonarens, Seyed Eghbal Ghobadi, and Christoph Stiller. Reducing ghost detections through uncertainty modeling for automated driving. 2023.

[HBWP13] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

[HFW+20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[HGBS21] Ahmed Hammam, Seyed Eghbal Ghobadi, Frank Bonarens, and Christoph Stiller. Real-time uncertainty estimation based on intermediate layer variational inference. In *Computer Science in Cars Symposium*, pages 1–9, 2021.

[HGDG17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[HKH22] Marius Hobbhahn, Agustinus Kristiadi, and Philipp Hennig. Fast predictive uncertainty for classification with bayesian deep networks. In *Uncertainty in Artificial Intelligence*, pages 822–832. PMLR, 2022.

[HM19] D. Hendrycks and M. Mazeika. Using self-supervised learning can improve model robustness and uncertainty. *NeurIPS*, 2019.

[HMD18] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[HSC+19] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, October 2019.

[HTI+18] D. Hafner, D. Tran, A. Irpan, T. Lillicrap, and J. Davidson. Reliable uncertainty estimates in deep neural networks using noise contrastive priors. *ResearchGate Preprint*, 2018.

[Hub64] Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[HW21] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[JBZB20] A. Jaiswal, A.R. Babu, M.Z. Zadeh, and D. Banerjee. A survey on contrastive self-supervised learning. *Technologies*, 2020.

[JGB+20] Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and trends® in computer graphics and vision*, 12(1–3):1–308, 2020.

[JHP+22] Sunguk Jung, Hyeonbeom Heo, Sangheon Park, Sung-Uk Jung, and Kyungjae Lee. Benchmarking deep learning models for instance segmentation. *Applied Sciences*, 12(17):8856, 2022.

[KCZ+21] Anna-Kathrin Kopetzki, Bertrand Charpentier, Daniel Zügner, Sandhya Giri, and Stephan Günnemann. Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In *International Conference on Machine Learning*, pages 5707–5718. PMLR, 2021.

[KHN23] N. Kotelevskii, S. Horváth, and K. Nandakumar. Dirichlet-based uncertainty quantification for personalized federated learning. *arXiv Preprint*, 2023.

[KLM19] Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32:3792–3803, 2019.

[KPM+18] Soheil Kolouri, Philip Pope, Charles Martin, Gustavo K Rohde, et al. Sliced wasserstein distance: A lightweight statistical distance for machine learning. *arXiv preprint arXiv:1804.01947*, 2018.

[KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[KW+19] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LDG+17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017.

[LEvdLL23] Jasper Linmans, Stefan Elfwing, Jeroen van der Laak, and Geert Litjens. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. *Medical Image Analysis*, 83:102655, 2023.

[LLLS18] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[LPB16] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc., 2017.

[LQQ+18] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[LWOL20] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[LWX+22] Y. Li, C. Wang, X. Xia, T. Liu, and B. An. Out-of-distribution detection with an adaptive likelihood ratio on informative hierarchical vae. *NeurIPS*, 2022.

[Mac92] David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

[MBK21] Xuhui Meng, Hessam Babaee, and George Em Karniadakis. Multi-fidelity bayesian neural networks: Algorithms and applications. *Journal of Computational Physics*, 438:110361, 2021.

[MCL+21] Andrés R Masegosa, Rafael Cabañas, Helge Langseth, Thomas D Nielsen, and Antonio Salmerón. Probabilistic models with deep neural networks. *Entropy*, 23(1):117, 2021.

[MG18a] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[MG18b] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018.

[Min00] Thomas Minka. Estimating a dirichlet distribution, 2000.

[MMN19] John Mitros and Brian Mac Namee. On the validity of bayesian neural networks for uncertainty estimation. *arXiv preprint arXiv:1912.01530*, 2019.

[MPV21] J. Mena, O. Pujol, and J. Vitrià. A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. *ACM Computing Surveys*, 2021.

[MS24] Ibomoiye Domor Mienye and Theo G Swart. A comprehensive review of deep learning: Architectures, recent advances, and applications. *Information*, 15(12):755, 2024.

[Mur12] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[MV20] Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.

[MYII23] A. Miyai, Q. Yu, D. Ikami, and G. Irie. Rethinking rotation in self-supervised contrastive learning: Adaptive positive or negative data augmentation. *WACV*, 2023.

[OFR+19] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019.

[OSF+18] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[PBZ21] Tim Pearce, Alexandra Brintrup, and Jun Zhu. Understanding softmax confidence and uncertainty. *arXiv preprint arXiv:2106.04972*, 2021.

[PRG+16] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[PTR+21] Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-net transformer: Self and cross attention for medical image segmentation. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 267–276. Springer, 2021.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[RHGS15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[RJP+22] J. Ren, L. Jiang, H. Peng, Z. Liu, and J. Wu. Evidential temporal-aware graph-based social event detection via dempster-shafer theory. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[RLZ+22] Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[RMW14] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286. PMLR, 2014.

[RRK22] S. Ramakrishna, Z. Rahiminasab, and G. Karsai. Efficient out-of-distribution detection using latent space of ŏ3b2-vae for cyber-physical systems. *ACM Transactions on Cyber-Physical Systems*, 2022.

[RTG+19] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[SCM21] V. Sehwag, M. Chiang, and P. Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv Preprint*, 2021.

[SHB+22] Timo Sämann, Ahmed Mostafa Hammam, Andrei Bursuc, Christoph Stiller, and Horst-Michael Groß. Improving predictive performance and calibration by weight fusion in semantic segmentation. *arXiv preprint arXiv:2207.11211*, 2022.

[SKK18] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[SMH+21] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021.

[SS20] A. Schwaiger and P. Sinhamahapatra. Is uncertainty quantification in deep learning sufficient for out-of-distribution detection? *Fraunhofer Reports*, 2020.

[SY14] M. Sakurada and T. Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. *Machine Learning*, 2014.

[SZSS21] Yichen Shen, Zhilu Zhang, Mert R Sabuncu, and Lin Sun. Real-time uncertainty estimation in computer vision via uncertainty-aware distribution distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 707–716, 2021.

[Tha20] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 17(1):168–192, 2020.

[THD24] A.T. Tran, V.N. Huynh, and V.H. Dang. A novel privacy preserving framework for training dempster-shafer theory-based evidential deep neural network. In *International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, 2024.

[TL19] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019.

[Ton22] Z. Tong. *Evidential Deep Neural Network in the Framework of Dempster-Shafer Theory*. PhD thesis, Université de Technologie de Compiègne, 2022.

[TPL20] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10781–10790, 2020.

[Tsi21] Theodoros Tsiligkaridis. Information aware max-norm dirichlet networks for predictive uncertainty estimation. *Neural Networks*, 135:105–114, 2021.

[TXD21] Z. Tong, P. Xu, and T. Denoeux. An evidential classifier based on dempster-shafer theory and deep learning. *Neurocomputing*, 2021.

[TZMF25] Paschalis Tsirtsakis, Georgios Zacharis, George S Maraslidis, and George F Fragulis. Deep learning for object recognition: A comprehensive review of models and algorithms. *International Journal of Cognitive Computing in Engineering*, 2025.

[UHF21] Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *arXiv preprint arXiv:2110.03051*, 2021.

[VGVG+21] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3614–3633, 2021.

[VHVZ21] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman. Open-set recognition: A good closed-set classifier is all you need? *NeurIPS*, 2021.

[WGW23] Dongdong Wang, Boqing Gong, and Liqiang Wang. On calibrating semantic segmentation models: analyses and an algorithm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23652–23662, 2023.

[WJ21] H. Wang and Q. Ji. Beyond dirichlet-based models: When bayesian neural networks meet evidential deep learning. *OpenReview Preprint*, 2021.

[WKM+19] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2, 2019.

[WLC+21] Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, and Dongsheng Li. Energy-based open-world uncertainty modeling

for confidence calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9302–9311, 2021.

[WVH24] Hongjun Wang, Sagar Vaze, and Kai Han. Dissecting out-of-distribution detection and open-set recognition: A critical analysis of methods and benchmarks. *International Journal of Computer Vision*, pages 1–26, 2024.

[WZH+22] Yanan Wu, Zhiyuan Zeng, Keqing He, Yutao Mou, Pei Wang, and Weiran Xu. Distribution calibration for out-of-domain detection with bayesian approximation. *arXiv preprint arXiv:2209.06612*, 2022.

[XLZL23] Mixue Xie, Shuang Li, Rui Zhang, and Chi Harold Liu. Dirichlet-based uncertainty calibration for active domain adaptation. *arXiv preprint arXiv:2302.13824*, 2023.

[XSG21] F. Xia, J. Snell, and T.L. Griffiths. Early exiting in deep neural networks via dirichlet-based uncertainty quantification. *OpenReview Preprint*, 2021.

[XWY+21] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090, 2021.

[YCW+20] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[YZ18] Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. In *International conference on machine learning*, pages 5660–5669. PMLR, 2018.

[ZL23] Z. Zeng and B. Liu. Unsupervised out-of-distribution detection by restoring lossy inputs with variational autoencoder. *ResearchGate Preprint*, 2023.

[ZS20] Christian Zimmermann and Didier Stricker. Efficientpose: Efficient human pose estimation with neural architecture search. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020.

[ZWXH24] C.C. Zong, Y.W. Wang, M.K. Xie, and S.J. Huang. Dirichlet-based prediction calibration for learning with noisy labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.