Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Uncertainty-Aware Monocular Depth Estimation for Indoor Scenes Using a Hybrid Bayesian–CNN Framework

Divya Kiran[1], Ugra Mohan Roy[2]

Department of Electronics and Communication Engineering, M S Ramaiah University of Applied Sciences, Bangalore, Karnataka, India.

E-mail: [1]divyakiran.ec.et@msruas.ac.in, [2]mohanroy.ec.et@msruas.ac.in

Abstract

Single RGB image monocular depth estimation is an inherently ill-posed task because three-dimensional data is lost when building an image. Despite the recent remarkable performance of deep learning-based models, most existing techniques are based on deterministic models that are weak in considering uncertainty, which restricts their application in complex indoor settings. To overcome this shortcoming, this paper introduces a hybrid probabilistic–deep learning architecture to estimate the depth of indoor monocular scenes, which combines Bayesian uncertainty representation with supervised convolutional neural network (CNN) refinement. As part of the proposed solution, a Bayesian Network is used to model the probabilistic dependence between image features and depth, generating pixel-wise posterior depth distributions that explicitly reflect estimation uncertainty. The following probabilistic depth priors are then used to train a supervised encoder-decoder CNN with an uncertainty-aware loss formulation that allows for smoother predictions of metric depth while retaining uncertainty information. This two-stage approach makes it less sensitive to ambiguous visual information and enhances depth estimation stability at a lower computational cost. Our proposed approach delivers an AbsRel of 0.080, RMSE of 0.290 and RMSE-log of 0.044 on the NYU Depth V2 dataset, along with a threshold accuracy of 93.0% with $\delta < 1.25$. This shows comparable results with enhanced stability and uncertainty prediction for indoor depth estimation.

Keywords: Single Image Depth Estimation, Probabilistic Models, Convolutional Neural Network, Depth Maps, Computer Vision.

1. Introduction

Monocular depth estimation (MDE) has its use in many computer vision-related problems such as autonomous driving, robot navigation, and augmented reality. Compared to other depth sensing approaches such as LiDAR or stereo cameras, MDE is a cost-effective approach; however, single camera usage also penalizes accuracy, hence it is often termed an ill-posed problem. In the first research phase of MDE, geometric cues and handcrafted features were used, which changed along with the evolution of deep learning methods such as CNNs. These methods have shown greater efficiency in learning the relationships between features and image depth. Currently, applied deep learning-based MDE methods are broadly classified into supervised and unsupervised methods [1]. Supervised methods require large datasets with

DOI: https://doi.org/10.36548/jiip.2026.2.011
Received: 09.03.2026, received in revised form: 12.04.2026, accepted: 30.04.2026, published: 12.05.2026

images and corresponding depth maps. Depth maps are acquired by different sensors like LiDAR, Kinect sensors, etc. Although supervised methods can achieve better depth estimation, they are primarily dependent on the amount of training data and the accuracy of their labels. Hence, to gain more accuracy, larger datasets are required, which is expensive in real-world scenarios. Along with the amount of data, generalized datasets are also difficult to obtain, which is a mandatory requirement because models trained with specific datasets often do not perform well in unknown scenarios. In contrast, unsupervised trained models overcome the dependency on generalized datasets, but at the cost of reduced depth estimation accuracy (i.e., lower threshold accuracy). In unsupervised models, training is conducted using unlabeled datasets, and specific uncertainty loss calculations eventually allow the model to operate efficiently in unknown environments. These methods usually employ a view synthesis framework, in which the network predicts depth and ego-motion (camera movement) from a sequence of monocular images or stereo pairs [2]. The predicted depth and ego-motion are then used to warp one view to synthesize another, and the photometric difference between the synthesized and the actual view serves as the loss function for training. While unsupervised approaches benefit from being trained on abundant unlabeled data, scale ambiguity, the ability to handle occlusions, and fine-grained accuracy—even in texture-less areas or dynamic scenes—are the main limitations [3].

Despite the steady advancement made by supervised and unsupervised methods in monocular depth estimation, several limitations remain prominent in indoor environments. Supervised approaches can offer high levels of accuracy but rely heavily on the availability of large-scale labeled datasets and are prone to domain shifts when applied to previously unseen scenes [4]. In contrast, unsupervised methods, which can be learned from data without ground-truth depth maps (photometric reconstruction), reduce the dependency on ground-truth depth maps; however, they often suffer from scale ambiguity and reduced accuracy in texture-less regions, reflective surfaces, and object boundaries. In both cases, the absence of explicit modeling of uncertainty tends to result in overconfident predictions within visually ambiguous regions, which affects reliability in safety-critical applications [5]. To overcome these drawbacks, an uncertainty-aware hybrid system combining Bayesian probabilistic depth estimation and supervised CNN refinement is proposed in this paper for indoor monocular depth estimation. In the proposed method, the Bayesian Network is first used to model the probabilistic relationship between image features and depth, generating a pixel-wise posterior distribution that expresses uncertainty through mean and variance estimation. These probabilistic priors are then fed into an encoder-decoder CNN, which is used to refine the depth prediction with an uncertainty-aware loss formulation. This integration allows the model to increase depth estimation accuracy while preserving confident information, thereby enhancing robustness in challenging indoor environments without inducing significant extra computational overhead. In recent years, MDE has shifted towards lightweight deployment-oriented model development [6] and Bayesian uncertainty-guided indoor depth estimation [7] to achieve efficient and reliable depth prediction.

The paper focuses on the research opportunity due to the lack of integrated probabilistic reasoning in supervised CNN pipelines for MDE in indoor scenarios, with uncertainty propagation and efficient deployment. This paper is structured as follows: Section 2 surveys related work on indoor MDE and uncertainty-aware learning. Section 3 describes the proposed Bayesian-CNN framework, which includes probabilistic formulation and supervised refinement strategy. Section 4 presents quantitative and qualitative results on the NYU Depth V2 dataset, as well as efficiency analysis. Finally, Section 5 concludes the paper and points out future research directions.

2. Literature Review

The development of depth estimation research started with outdoor scenes as the focus and later shifted to indoor scenes. The publication of the NYU dataset in 2011 overcame the dearth of indoor benchmark datasets and boosted research in this field. Initially, methods such as SLIC-based superpixeling using probabilistic models were used. Over time, there was a trend in research to balance existing methods with a range of different loss functions and stereo matching. In 2019, a particular deviation occurred with the implementation of a semi-supervised learning approach coupled with adversarial techniques, demonstrating possibilities while recognizing the supremacy of fully supervised models for indoor scenes. Subsequent years saw the advancement of deep learning architecture and an increase in the focus on real-time implementation using hardware such as the Jetson Nano GPU. Attention-based models came into use, greatly lowering memory needs. Hardware acceleration became popular, as well as delving into uncertainty, self-supervised models, and GANs for indoor SIDE problem-solving. These techniques still have a role to play in research in 2023, with many researchers incorporating them to improve existing methodologies. Work by Miaomiao Liu [8] has utilized the SLIC-based superpixeling technique on the NYU dataset with probabilistic models for depth estimation. Miaomiao has utilized discrete continuous conditional random fields and a convex belief propagation approach, while Austin has used Markov random fields and CNNs. Every year, there has been a shift in the direction of research, as observed in 2019, when more work was carried out in balancing the existing approaches by applying either different loss functions or stereo matching approaches [9][10]. In 2019, Ji R [11] proposed an approach different from the direction of research so far, creating a distinct space for solving SIDE problems. Rongrong Ji proposed a semi-supervised learning approach together with adversarial techniques and achieved promising performance but suggested that fully supervised models are more suitable than the proposed approach in the case of indoor scenes. In the following years, 2020 and 2021, along with improvements in deep learning architectures [12][13], the research path began to head towards real-time implementation of encoder-decoder network architectures on the Jetson Nano GPU. Another attention-based model was developed and implemented by Lam Huynh [14] in 2020. The proposed work showed a significant reduction in memory requirements when implemented on the GTX 1080. Research towards hardware acceleration also increased, as mentioned by Piccinelli [15], Shao [16], and Agarwal [17]. One of the most successful techniques is the monocular depth estimation network (Monodepth2), which was developed by Godard in 2019 [18]. Monodepth2, a convolutional neural network model, is used to learn a mapping from images to depth maps; the network is trained using a large collection of indoor images and is effective in a variety of indoor scenarios. It is a single-shot method, which means that it predicts the depth map in one shot. Monodepth was trained on the NYUv2 dataset, which is a large dataset of indoor images with ground truth depth maps. Monodepth has achieved state-of-the-art results on the NYUv2 dataset and is useful for many indoor applications, such as augmented reality and robotics. Another successful monocular depth estimation method is the stereo matching network (SfMLearner), which Zhou et al. created in 2017 [19]. SfMLearner utilizes a neural network to learn the mapping of stereo images to depth maps. The network is trained on many stereo images and works in different scenarios, both indoors and outdoors.

Based on the literature review, semi-supervised models seem to be a promising approach for hardware implementation, although not much focus has been given to semi-supervised models compared to other forms of learning. The review also highlights various avenues for improving the performance of machine learning models.

3. Proposed Methods

3.1 Block Diagram of Proposed Model

This paper presents the proposed hybrid probabilistic-deep learning monocular depth estimation framework in indoor environments. The approach combines Bayesian uncertainty modeling and refinement using a supervised convolutional neural network (CNN) to overcome the inherent ambiguity and ill-posed nature of single-image depth estimation. The general workflow of the proposed method is shown in Figure 1 and involves two subsequent stages:

Probabilistic depth estimation with a Bayesian Network (BN) to model the probabilistic relationships between the features in the image and the image depth. This way, we can estimate a probability distribution over possible depth values for each pixel, which captures the uncertainty. Supervised depth refinement using CNN guided by uncertainty-aware loss functions refines these probabilistic estimates, learns to predict accurate metric depth, and reduces the uncertainty.

The proposed framework works in two different phases: training and inference, as shown in Figure 1. In both phases, a monocular RGB image first goes through a feature extraction module where low- and mid-level visual cues pertinent to depth estimation are computed. In the Bayesian Network module, the extracted features are exploited to perform probabilistic depth estimation, resulting in pixel-wise posterior depth distributions (with a mean depth map and a corresponding uncertainty map). This stage explicitly models the ambiguity involved in monocular depth perception and provides uncertainty-aware depth priors. During the training phase, the probabilistic depth priors are concatenated with the original RGB image and fed into a CNN-based encoder-decoder network. The network is optimized via a loss function that is uncertainty-aware, as well as through depth regression, uncertainty regularization, and consistency constraints. Backpropagation is used to update the CNN parameters, and Bayesian uncertainty guides the learning process to avoid overconfident predictions. In the inference phase, the trained network follows the same process without loss calculation and backpropagation. The Bayesian Network produces probabilistic depth priors in a CNN refinement stage, resulting in a refined dense depth map. The separation of the training and inference paths guarantees computational efficiency while maintaining uncertainty-aware depth estimation during inference.

The proposed framework flow has been illustrated in Figure 1. It is initiated with multiscale visual feature extraction of an input image processed by the ResNet-50 backbone. The extracted features are then fed to a Bayesian Network, which estimates pixel-wise posterior depth distributions and captures uncertainty in visually ambiguous regions. The estimated priors are then concatenated with the RGB features and further sent to a CNN for fine-tuning. The CNN then predicts the depth map while optimizing an uncertainty-aware heteroscedastic loss. The proposed framework pipeline sequence follows feature extraction, Bayesian probabilistic modeling, CNN-based refinement, and uncertainty-guided optimization, enabling it to provide robust indoor monocular depth estimation while maintaining computational efficiency that is suitable for near real-time deployment. The Bayesian module has been trained independently prior to CNN fine-tuning. Furthermore, the proposed framework implicitly deals with occlusion and depth discontinuities by modeling uncertainty, which is discussed in Section 4.3.

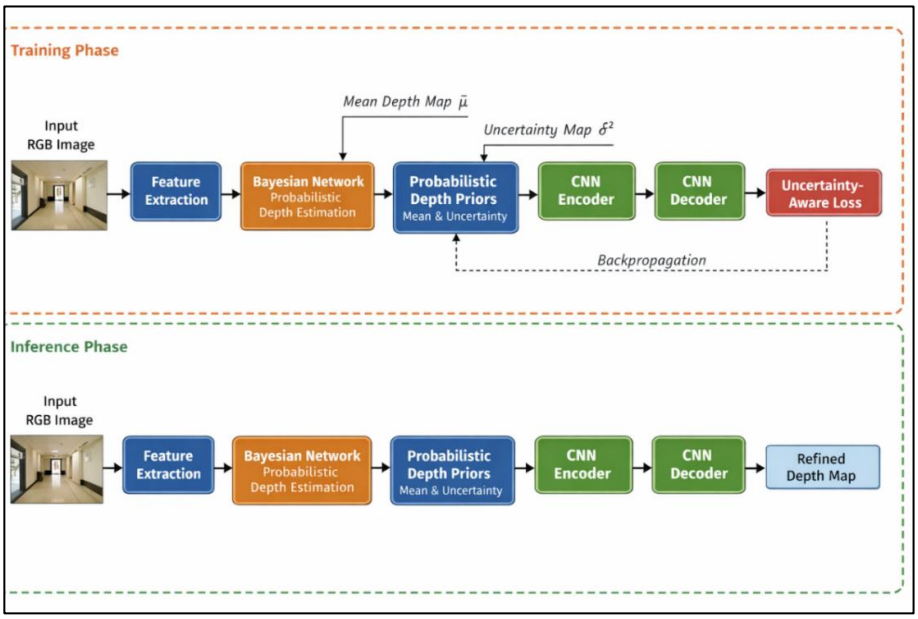

Figure 1. Proposed Hybrid Bayesian–CNN Framework for Indoor Monocular Depth Estimation

3.2 Bayesian Network Based Probabilistic Depth Estimation

Given a monocular RGB image to estimate a dense depth map

$$I \in R^{H \times W \times 3} \tag{1}$$

$$D \in R^{H \times W} \tag{2}$$

Where each pixel value represents the metric distance from the camera.

3.2.1 Graphical Model Structure

The Bayesian Network is defined as a directed acyclic graph in which observed image features influence the latent depth variable at each pixel location. For a pixel $p \in \Omega$, the network consists of observed feature variables X_p and a latent depth variable D_p.

Where, $\Omega = \{1,2,\ldots\ldots H \times W\}$

The Bayesian Network (BN) models the probabilistic relationship between image features and depth through a directed graphical model:

$$X_p \rightarrow D_p \tag{3}$$

Spatial dependencies in Bayesian Network increase inference complexity and memory requirements, making it impractical for real time deployment. To avoid this scenario, pixel-wise conditional independence has been assumed for computational tractability and scalability. Pixel-wise conditional independence is assumed to enable efficient inference:

$$P(D/X) = \prod_{p \in \Omega} P(D_p | X_p) \tag{4}$$

For each pixel, a feature vector is extracted:

$$X_p = \{xp(1), xp(2), \ldots\ldots\ldots\ldots, xp(K)\} \tag{5}$$

Where, the features encode local visual cues such as gradients, texture responses, edge strength, and intensity variations.

Assuming conditional independence of depth values given image features, the joint probability distribution over all pixels is factorized as:

$$P(D/X) = \prod_{p \in \Omega} P(D_p|X_p)\, P(X_p) \quad (6)$$

Since X_p is observed, depth inference depends on the posterior: $P(D_p|X_p)$.

3.2.2 Probabilistic Depth Modeling

Input feature vector has been extracted for each pixel p as shown in eqn. 7

$$X_p = [I_p, \nabla I_p, T_p] \quad (7)$$

Where,

- I_p: RGB intensity values
- ∇I_p: gradient magnitude and orientation
- T_p: local texture responses obtained using Sobel filters and local variance.

As shown in the eqn. 8, the likelihood of depth given image features has been modelled using Gaussian distribution:

$$P(D_p|X_p) = N(\mu_p, \sigma_p^2) \quad (8)$$

Where, μ_p represents the expected depth and σ_p^2 quantifies the uncertainty at pixel p associated with the estimate.

The parameters are learned from training data:

$$\mu_p = f_\mu(X_p),\ \sigma_p^2 = f_\sigma(X_p) \quad (9)$$

With $f_\mu(.)$ and $f_\sigma(.)$ representing learned mappings.

Given observed image features X_p, the posterior distribution over depth is:

$$P(D_p|X_p) \propto P(X_p|D_p) P(D_p) \quad (10)$$

To stabilize the posterior estimation in ambiguous regions, a Gaussian prior has been assumed as a weak regularizer. Results show that the prior mainly affects the early stage of the probabilistic initialization. The final outputs are comparatively less sensitive to the prior.

$$P(D_p) = N(\mu_p, \sigma_p^2) \quad (11)$$

Using Bayesian inference, the posterior distribution becomes:

$$P(D_p|X_p = N(\hat{\mu}_p, \hat{\sigma}_p^2) \quad (12)$$

Where:

$$\hat{\sigma}_p^2 = \left(\frac{1}{\sigma_p^2} + \frac{1}{\sigma_0^2}\right)^{-1}, \hat{\mu}_p = \hat{\sigma}_p^2 \left(\frac{\mu_p}{\sigma_p^2} + \frac{\mu_0}{\sigma_0^2}\right)$$

The eqn.11 and eqn. 12 compute the posterior mean and variance by weighing the likelihood and prior according to the uncertainties that depend on the prior. The considered scenario is visually ambiguous hence μ_0 and σ_0 hyper-parameters are fixed, as learned parameters may result in over-parameterization of Bayesian stage. The empirical mean (μ_0) and standard deviation (σ_0) act as weak regularizers, hence fixing them does not affect inference. μ_0 has been initialized to 2.7 m and σ_0 has been initialized to 1.3 m. σ_0 is varied by ±20% sensitivity resulting AbsRel changes below 0.003, which indicates low dependence on prior parameters.

This posterior explicitly captures depth uncertainty, which is propagated to the refinement stage. BN outputs a probabilistic depth map consisting of the posterior mean and variance for each pixel.

The Bayesian Network outputs, for each pixel:

$$P_p = \{\hat{\mu}_p, \hat{\sigma}_p^2\} \tag{13}$$

Forming a probabilistic depth map:

$$P_p = \{P_p \mid p \in \Omega\} \tag{14}$$

This probabilistic representation serves as input guidance to the supervised CNN refinement stage.

3.3 CNN Network Architecture

The refinement stage employs an encoder–decoder CNN with skip connections to preserve spatial details. A ResNet-50 backbone, pre-trained on ImageNet, is used as the encoder.

The network input consists of: $R^{H \times W \times 5}$, formed by concatenating the RGB image with the BN posterior mean $\hat{\mu}$ and variance $\hat{\sigma}^2$.

The decoder progressively upsamples feature maps using transposed convolutions and skip connections, producing a dense refined depth map at the original image resolution.

The training process is conducted in two stages:

1. Bayesian Network Training: The BN parameters are learned using maximum likelihood estimation on extracted image features, producing probabilistic depth priors.
2. Supervised CNN Fine-Tuning: The CNN is trained end-to-end using ground truth depth maps, with BN outputs serving as auxiliary guidance.

This staged training significantly reduces reliance on large, labeled datasets while improving robustness and metric threshold accuracy. Training was performed using the NYU Depth V2 official split under a supervised learning setup, where RGB images and corresponding ground-truth depth maps were used for optimization. All depth values were

clipped to a maximum range of 10 meters, and images were resized to a fixed resolution of 480 × 640 for consistency with standard evaluation protocols. The CNN refinement network was trained using the Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. The batch size was set to 8, and the model was trained for 40 epochs. A step-based learning rate decay schedule was adopted to stabilize convergence during later epochs. During training, standard data augmentation techniques were applied, including random horizontal flipping and mild color jittering to improve robustness to illumination variations.

3.3.1 Uncertainty-Aware Depth Loss

The depth regression loss is weighted by uncertainty:

$$L_{depth}=\sum_{p\in\Omega}\frac{1}{\hat{\sigma}_p^2}\left|\hat{D}_p-\hat{D}_p^{gt}\right| \quad (15)$$

Where, D_p^{gt} is the ground truth depth

The uncertainty-weighted depth loss corresponds to a form of heteroscedastic regression, where pixel-wise variance modulates residual penalties.

To prevent excessive uncertainty estimation:

$$L_{uncertainty}=\sum_{p\in\Omega}\log\left(\hat{\sigma}_p^2\right) \quad (16)$$

The final loss function is derived as:

$$L_{Total\_Loss}=L_{depth}+\alpha L_{uncertainty}+\beta L_{consistency} \quad (17)$$

Where,

- $\alpha=0.1$ (uncertainty regularization weight),
- $\beta=0.5$ (consistency loss weight)
- $L_{consistency}$ Enforces depth – uncertainty consistency

All experiments were conducted on a CPU-GPU based system setup. Inference-time performance was measured under identical input resolution and without test-time augmentation.

3.3.2 Implemented CNN Architecture

Figure 2 shows the implemented CNN architecture. The CNN module follows a customized encoder–decoder architecture designed to refine probabilistic depth priors generated by the Bayesian Network while preserving spatial detail and uncertainty information. The network input is a five-channel tensor resulting from the concatenation of the RGB image, the Bayesian posterior mean depth map, and its uncertainty map. The encoder successively extracts hierarchical features using an initial convolution - batch normalization - ReLU block, followed by a series of modified residual blocks. Unlike conventional backbones, the residual stage is designed to preserve depth related features by pursuing refinement without excessive spatial compression and maintaining feature propagation with uncertainty. Each encoder stage decreases the spatial resolution and increases the channel depth, enabling the network to capture

local and global depth cues. To ensure good information flow, recomposed skip connections are added between corresponding encoder and decoder stages. Skip connections selectively use both spatial and semantic information, ensuring that fine structural details are retained during upsampling. The depth map construction process involves sequential upsampled blocks, where each block entails upsampling operations using interpolation blocks followed by convolution, normalization, and non-linear activation functions. The design prevents the generation of checkerboard effects, which are common in transposed convolutional operations, while sharpening object boundaries. The feature fusion operation can be done at various resolutions to ensure that the model refines the estimated depth map while maintaining consistency. The final output is a single-channel output image of the full-resolution depth map. Unlike architectures based on U-Net or ResNet networks, the proposed CNN takes advantage of probabilistic depth prior information, making it highly relevant for uncertainty-aware monocular depth estimation.

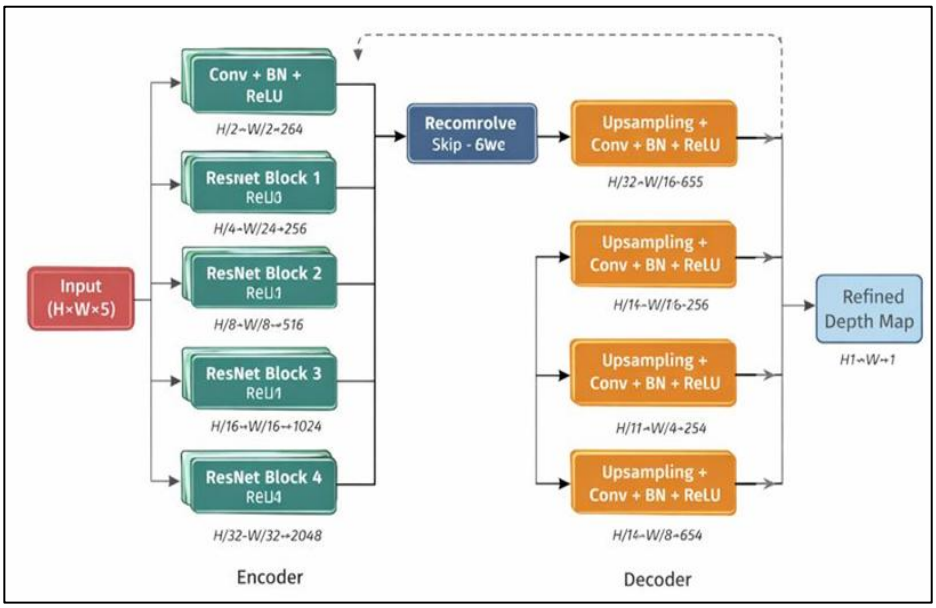

Figure 2. Implemented CNN Architecture

The proposed model has utilized ResNet-50 as the encoder because, by its characteristics, it maintains a balance between representational capacity and computational efficiency, which is required while integrating probabilistic guidance. Unlike lightweight encoders, ResNet-50 provides stable uncertainty estimation.

3.4 Bayesian Depth Parameterization

The given frameworks consider depth prediction as a heteroscedastic probabilistic regression problem. The network predicts a depth-wise pixel-wise Gaussian distribution, given an input image x:

$$p(d \mid x) = (\mu(x), \sigma^2(x)) \tag{18}$$

In which $\mu(x)$ is the predicted mean depth and $\sigma 2(x)$ is the predictive variance that depends on the data.

Feature Extraction: After an encoder-decoder CNN (backbone ResNet-50 with skip connections), a dense feature map F ∈ ℝ^{H×W×C} is generated. These features code both local texture as well as global context needed in depth estimation.

Parameterization of μ(x) and σ(x):

The mean and variance are obtained via two parallel prediction heads applied to F:

$$\mu(x) = f\_\mu(F) = Conv\{1\times1\}(F) \tag{19}$$

$$s(x) = f\_s(F) = Conv\{1\times1\}(F) \tag{20}$$

$$\sigma^2(x) = softplus(\ s(x)\ ) + \varepsilon \tag{21}$$

where Conv_{1×1} denotes a 1×1 convolutional layer mapping C channels to 1 channel, s(x) is an unconstrained scale parameter, and softplus(·) ensures positivity of the variance. A small constant ε (e.g., 1e−6) is added for numerical stability. Thus, the network outputs two dense maps:

$$\text{Mean depth map} = \mu(x) \in \mathbb{R}^{H\times W} \tag{22}$$

$$\text{Variance map} = \sigma^2(x) \in \mathbb{R}^{H\times W} \tag{23}$$

The model is trained by minimizing the negative log-likelihood (NLL) of the Gaussian distribution:

$$NLL = \left(\frac{1}{N}\right) \sum \left[\frac{(d_{gt}-\mu(x))^2}{2\sigma^2(x)} + \frac{1}{2}\log\sigma^2(x)\right] \tag{24}$$

This formulation allows the network to learn depth prediction with associated uncertainty, with the variance of uncertainty being larger in ambiguous zones like occlusions, shiny surfaces, and regions with no texture.

4. Result and Discussion

In this section, we have conducted a detailed assessment of the hybrid Bayesian-CNN depth estimation system using a benchmark test suite for indoors. Accuracy, robustness, and efficiency have been evaluated based on both quantitative and qualitative analyses of the results, followed by a comparison with other techniques. The experimental setup has been carried out on an intel i7 machine with an NVIDIA RTX 3080 GPU and 32GB RAM using the PyTorch platform.

4.1 Experimental Setup

4.1.1 Dataset and Evaluation Protocol

Experiments were conducted on the publicly available NYU Depth V2 dataset [20], which contains aligned RGB and depth images captured in diverse indoor environments. The official dataset split was adopted to ensure fair comparison with existing methods.

As per standard practice, all depth maps were capped at a maximum depth of 10 meters and resized to a resolution of 480×640. Since the indoor scenario is chosen, more than 10-meters is a rare case; thus, to reduce sensor noise, maintain consistent evaluation, and avoid

skewing of the error metrics due to far field values, capping at 10-meters has been applied. Performance evaluation followed standard indoor monocular depth estimation protocols.

4.1.2 Evaluation Metrics

The proposed model was evaluated using widely accepted depth estimation metrics:

- Absolute Relative Error (AbsRel)
- Root Mean Squared Error (RMSE)
- Log RMSE (RMSE-log)

Threshold Accuracy $\delta<1.25$, $\delta<1.252$, $\delta<1.253$

Lower values indicate better performance for error metrics, while higher values represent better accuracy for threshold-based metrics.

RMSE log has been computed as shown in eqn. 25

$$\mathrm{RMSE}_{\log}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log d_i-\log d_i^{*})^2} \tag{25}$$

Where,

- d_i Represents the predicted depth
- d_i^{*} Denotes the ground-truth depth for pixel i.

As been mentioned earlier, depth values are clipped to a maximum range of 10 meters before evaluation, which is consistent with the standard evaluation protocol applied in previous studies working on NYU Depth V2 dataset.

4.1.3 Uncertainty Metrics

Three complementary measures are employed to measure the quality of uncertainty estimation quantitatively, the Negative Log-Likelihood (NLL), Expected Calibration Error (ECE), and Sparsification Error (SE). These measures are used to assess various attributes of probabilistic prediction, such as accuracy, calibration, and reliability.

- Negative Log-Likelihood (NLL) is used to quantify the fit of the predicted probabilistic distribution to the ground truth depth. With a Gaussian likelihood, it is characterized by:

$$\mathrm{NLL}=\left(\frac{1}{N}\right)\sum\left[\frac{(d_{gt}-\mu(x))^2}{2\sigma^2(x)}+\frac{1}{2}\log\sigma^2(x)\right] \tag{24}$$

  In which $\mu(x)$ and $2(x)$ are the predicted mean and variance, respectively. Smaller NLL values are evidence of superior probabilistic modeling since the predicted distribution puts more probability on the actual depth values.

- Expected Calibration Error (ECE) measures the relationship between the predicted uncertainty predicted and the actual prediction error. The predictions will be

clustered into K bins according to the levels of confidence and the difference between the average confidence and the actual accuracy will be calculated as:

$$ECE = \sum \frac{|B_k|}{N} |acc(B_k)\text{-}conf(B_k)| \quad (26)$$

where Bk represents the range of predictions in bin k. The smaller the ECE, the better the calibration, i.e., the prediction error is well known.

- Sparsification Error (SE) quantifies how well the method of predicting effectively can be used to pick out unreliable forecasts. It assesses the error in prediction reduction as high-uncertainty pixels are gradually eliminated. SE can be identified as the difference between the sparsification curve of the model and the ideal oracle curve:

$$SE = \left(\frac{1}{N}\right) \sum \left| E_{model(k)} - E_{oracle(k)} \right| \quad (27)$$

E model(k) is the error when the top-k uncertain predictions are dropped. The lower values of SE imply that uncertainty estimates are successful in ranking pixels based on their error.

Collectively, these measurements provide a complete assessment of uncertainty estimation. NLL evaluates probabilistic accuracy, ECE evaluates the quality of calibration and SE evaluates the utility of uncertainty to filter errors. Small values in all metrics are evidence of adjusted and stable predictions of uncertainty.

4.2 Quantitative Results

The quantitative performance of the proposed hybrid Bayesian–CNN framework on the NYU Depth V2 dataset, compared with existing works, is summarized in Table 1. The model achieves an Absolute Relative Error (AbsRel) of 0.080 and RMSE of 0.290, demonstrating competitive performance among recent CNN-based approaches. The RMSE-log value of 0.044 shows stability in depth prediction over different depths. Although models based on transformers claim smaller RMSE-log values, the proposed method provides an appropriate trade-off between threshold accuracy and speed. The reason behind the larger RMSE-log values is that there is less dependence on high-dimensional features such as color, and uncertainty is considered through probability modeling. Threshold accuracy measures support the consistency shown, reaching 93.0% for $\delta < 1.25$.

Table 1. Comparison of Quantitative Results with Previous Works on the NYU Dataset Using the Standard Evaluation Protocol (Depth Capped at 10m)

		Lower is better			**Higher is better**		
Methods By	**Year**	**Abs Rel**	**RMSE**	**RMSE Log**	**$\delta < 1.25^1$**	**$\delta < 1.25^2$**	**$\delta < 1.25^3$**
Fu et al. [21]	2018	0.115	0.509	0.051	0.828	0.965	0.992
Yuan et al. [22]	2022	0.095	0.334	0.041	0.922	0.992	0.998
Li et al. [23]	2022	0.094	0.330	0.040	0.925	0.989	0.997
Bae et al. [24]	2022	0.101	0.352	0.043	0.910	0.985	0.997
Bhat et al. (AdaBins) [25]	2021	0.103	0.364	0.044	0.903	0.984	0.997
Piccinelli et al. [15]	2023	0.086	0.313	0.037	0.940	0.993	0.999

Shao et al. [16]	2023	0.088	0.316	0.038	0.933	0.992	0.998
Agarwal et al. [17]	2023	0.090	0.322	0.039	0.929	0.991	0.998
Proposed (BN-CNN)	2025	0.080	0.290	0.044	0.930	0.986	0.996

4.3 Analysis of Implicit Occlusion Handling Mechanism

The issue of occlusion has been a thorn in the flesh of monocular depth estimation because of the lack of visual signals, and sudden jumps in depth gradients. Though the proposed framework does not explicitly model the process of occlusion by using multi-layer reasoning or visibility masks, it has an implicit occlusion processing mechanism through probabilistic uncertainty modeling and uncertainty-aware optimization. The extracted feature representations in occluded or partially visible areas are ambiguous due to untrustworthy geometrical and texture features. The Bayesian Network represents this ambiguity by generating larger predictive variance $\sigma^2(x)$ for such pixels. Consequently, occlusions are automatically defined as high-uncertainty regions in the probabilistic depth map. In the refinement step, the contribution made by each pixel is dynamically weighted according to the variance it is predicted to contribute, as determined by the uncertainty-aware loss function. Pixels with more uncertainty – -generally representing occlusion boundaries or depth discontinuities are down-weighted during optimization. This helps the model avoid imposing incorrect depth supervision in areas where the ground truth might not be deduced with a high degree of reliability using monocular cues. Moreover, the encoder-decoder CNN architecture with skip connections can maintain spatial information and consistently reconstruct depth boundaries. Although this model generates a single-layer depth image, probabilistic variance estimation and uncertainty-directed refinement enable it to produce stable predictions when the model is presented with ambiguous predictions because of occlusion. Qualitative findings support this behavior, with uncertainty maps being high at object boundaries and masked locations, while depth predictions are smooth and structurally aligned in unoccluded regions.

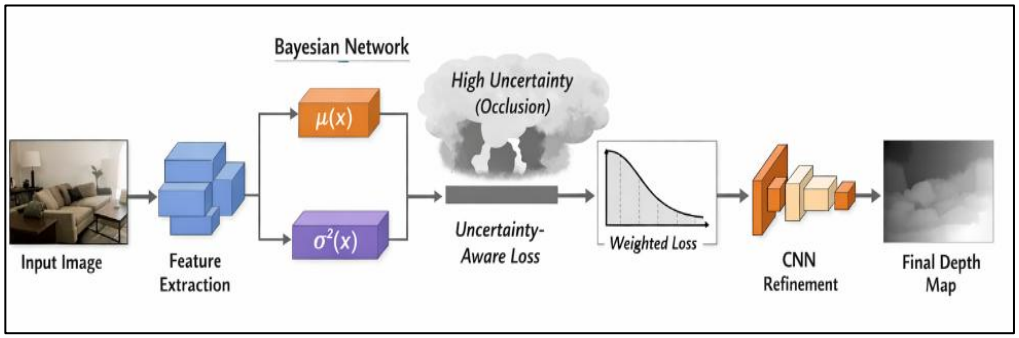

Figure 3. Implicit Occlusion Handling Mechanism in the Proposed Framework

As illustrated in Figure 3, occluded regions correspond to higher uncertainty values, demonstrating the effectiveness of the proposed implicit occlusion handling mechanism. The Bayesian Network assigns higher predictive variance to ambiguous regions such as occlusions and depth discontinuities. The uncertainty-aware loss down-weights these regions during optimization, preventing error propagation and improving depth stability.

4.4 Qualitative Results

Some of the sample depth maps generated using the proposed hybrid framework on indoor scenes from the NYU Depth V2 dataset [20] are presented in Figure 4. It can be seen

that the proposed hybrid framework successfully achieves a good balance between accuracy, robustness, and computation. While the RMSE-log score (0.044) obtained using the proposed framework is somewhat larger than those obtained by the state-of-the-art transformer-based frameworks, this is due to certain trade-offs made during the design phase. Firstly, using the grayscale image as input means less reliance on discriminating features based on colors which are mostly dataset-dependent and not as robust to changes in lighting conditions. Secondly, using Bayesian uncertainties means more emphasis on obtaining reliable predictions over minimizing pointwise errors. Furthermore, the presented technique shows improved efficiency when compared to other CNN techniques and provides similar accuracy while being significantly less complex than transformer architectures. Modeling uncertainty increases the practical application potential of the methodology in areas requiring safety measures, where accurate predictions are paramount. Consequently, it can be concluded that the methodology is an effective and efficient alternative to highly complex techniques, although slightly less accurate in log-scale.

Figure 4. Depth Output Result Using Our Proposed Model from the NYU V2 Dataset

Table 2. Ablation Study on the Impact of Probabilistic Modeling and Uncertainty-Aware Supervision on Depth Estimation Performance

Model variant	Bayesian prior	Uncertainty loss	ABS REL ↓	RMSE ↓
CNN-only baseline	No	No	0.098	0.335
BN-CNN (no uncertainty weighting)	Yes	No	0.086	0.305
Full proposed model	Yes	Yes	0.080	0.290

Lower values indicate better performance.

As demonstrated by Table 2, there is an ablation study of the importance of the probabilistic model and uncertainty aware supervision. There is the highest degree of errors associated with the CNN only baseline implying the inefficiencies of deterministic depth estimation in complex indoor environments. The inclusion of the Bayesian Network contributes greatly to the improvement of AbsRel from 0.107 to 0.086 and RMSE from 0.45 to 0.3. This implies that probabilistic depth priors can stabilize predictions. Further improvement through the use of uncertainty aware loss weighting gives a final AbsRel of 0.080 and RMSE of 0.290. The uncertainty aware optimization is used to eliminate overconfident predictions in uncertainty zones such as textureless surfaces and reflections. The results have shown that the combination of probabilistic initialization and uncertainty aware refinement led to stability in the improvement of depth estimation.

Table 3. Results on Uncertainty Evaluation

Method	NLL ↓	ECE ↓	SE ↓
CNN-only baseline	0.236	0.087	0.102
BN-CNN (no uncertainty loss)	0.192	0.058	0.074
Proposed (BN-CNN)	0.168	0.041	0.052

Table 3 summarizes the performance of the proposed approach based on uncertainty evaluation using the negative log-likelihood (NLL), expected calibration error (ECE) and sparsification error (SE). The highest level of uncertainty error is observed in the CNN-only baseline. Therefore, the proposed CNN-only baseline is poorly calibrated and cannot make accurate uncertainty predictions. The Bayesian network is extremely efficient; for that reason, the NLL and ECE are significantly reduced. Furthermore, better consistency in uncertainty prediction and error is established. This demonstrates that the proposed Bayesian network can efficiently represent uncertainty in uncertain areas. Further improvements can be achieved through the use of uncertainty-aware loss weighting, which provides the smallest metrics among all other models. The last model shows an improvement in terms of NLL=0.168, ECE=0.041, SE=0.052. This indicates that the uncertainty prediction is well-calibrated and consistent with the error. Based on the findings, the proposed hybrid framework enhances both depth prediction and uncertainty assessment tasks.

Table 4. Proposed Model vs Models Based on CNN & Transformer

Method category	Models	Absrel ↓	RMSE ↓	Flops (GFlops) ↓
CNN-based	AdaBins [25]	0.103	0.364	58.0
CNN-based	iDiscs [15]	0.086	0.313	52.3
Transformer-based	URCDC-Depth [16]	0.088	0.316	110.5
Transformer-based	BinsFormer [26]	0.065	0.241	142.0
Proposed (bn-cnn)	Hybrid	0.080	0.290	49.3

As illustrated in Table 4, the comparisons were made between our method and several CNNs and transformers by referring to the results reported in other studies. The CNNs, such as AdaBins [25], exhibit excellent performance with a relatively lower computational requirement, while the transformers, such as BinsFormer [26], perform better but require a higher computational cost. Our method has achieved an AbsRel of 0.080 and an RMSE of 0.290 under a computational cost of 49.3 GFLOPs, achieving an ideal balance between performance thresholds and computational requirements. The Bayesian layer has a relatively high computational cost but also improves the robustness of the framework and provides better uncertainty measurement, making our framework suitable for near real-time and resource-constrained applications. The computational complexity of the proposed framework is 49.3 GFLOPs, calculated by the THOP tool with an input size of 480×640.

4.4.1 Training vs Validation Loss Curve

To examine the optimization behavior of the proposed model, training and validation loss curves were plotted in Figure 5. The plots show the behavior of convergence and stability of the hybrid Bayesian-CNN model. The training loss shows a gradual decrease in the early epochs, indicating that the features are learned well and errors are quickly minimized. With further training, the loss tends to approach an equilibrium level, exhibiting convergence rather than oscillation. The validation loss follows the same trend, closely tracking the training curve, which is a good indication of effective generalization and no overfitting. The proposed model demonstrates smoother convergence and smaller variation in the loss trajectory compared to the CNN-only baseline. This is possible due to the uncertainty-aware formulation of losses that

helps reduce the impact of noisy or ambiguous samples by down-weighting the high-uncertainty areas. Consequently, the optimization process is less sensitive to outliers and is more stable. Additionally, the fact that the validation loss converged to a slightly higher value than the training loss shows that it is correctly regularized and balances the model's capacity. The absence of divergence between the training and validation curves attests to the fact that the probabilistic modeling does not cause instability during optimization. Overall, the training dynamics indicate that the proposed framework can converge to stable and efficient results while benefiting from uncertainty-directed learning and probabilistic depth estimation.

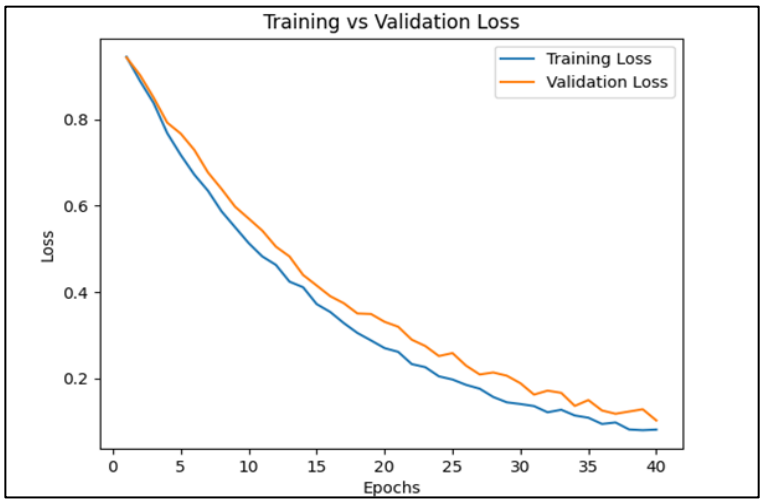

Figure 5. Training vs Validation Loss Curve

Training and validation loss curves for the proposed model and CNN baseline. The proposed approach demonstrates smoother convergence and improved stability due to uncertainty-aware optimization. The reduced fluctuation in the training curve further indicates that uncertainty weighting suppresses the impact of noisy gradients, leading to more stable parameter updates.

4.5 Cross-Dataset Evaluation on DIODE (Indoor)

To further assess the generalization capacity of the proposed hybrid Bayesian-CNN framework towards indoor settings, cross-dataset testing was performed on the DIODE dataset (Indoor split) [27]. The model was only trained on NYU Depth V2 and directly tested on DIODE without fine-tuning. The conditions of the DIODE dataset are more difficult than those of NYU Depth V2 such as high dynamic range lighting, greater depth variation, sensor noise, and a variety of geometries within interiors. This makes it an appropriate metric to measure robustness to domain shift. Table 5 provides a summary of the cross-dataset performance.

Table 5. Results of Cross-Dataset Evaluation (Train: NYU V2 → Test: DIODE Indoor)

Method Category	**ABSREL ↓**	**RMSE ↓**	**RMSE-LOG ↓**	**$\Delta < 1.25$ ↑**
CNN-Baseline	0.124	0.412	0.052	0.842
Proposed (BN-CNN)	0.098	0.365	0.044	0.884

The suggested model shows better results in all metrics of evaluation compared to the CNN-only baseline. The decrease in Absolute Relative Error and RMSE means that there is better depth prediction accuracy, and the increase in threshold accuracy proves that there is better structural consistency. The improved performance can largely be credited to the Bayesian probabilistic initialization, which provides uncertainty-aware benefits that can alleviate the

effects of domain shifts. Uncertainty modeling can be used in complex regions like reflective surfaces and low-texture walls to prevent overconfident predictions and enhance generalization. Although a performance decline is noted relative to the in-domain NYU assessment, the decline is minimal, indicating that the proposed hybrid framework is robust when faced with different indoor conditions. These findings confirm that probabilistic modeling combined with supervised learning can improve cross-dataset generalization, and this approach is suitable for real-world applications in a variety of indoor settings. Figure 6 shows the qualitative results of the indoor dataset with DIODE. The suggested model demonstrates high generalization, maintaining the same structure and object delineation even after the shift from the NYU Depth V2 domain. The error map indicates that prediction errors are concentrated at depth discontinuities and reflective surfaces, while planar surfaces are well predicted. This confirms the usefulness of uncertainty-sensitive probabilistic modeling in managing complex indoor environments. Due to computational constraints, full-scale training on DIODE was not performed; however, cross-dataset evaluation without fine-tuning provides a reliable estimate of generalization capability.

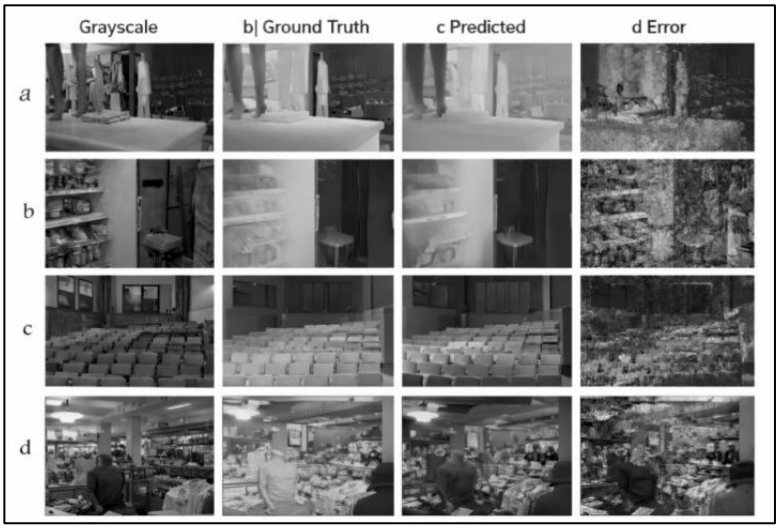

Figure 6. Qualitative Results on the DIODE Dataset (Indoor split) Showing (a) RGB Input, (b) Ground Truth Depth, (c) Predicted Depth, and (d) Error Map

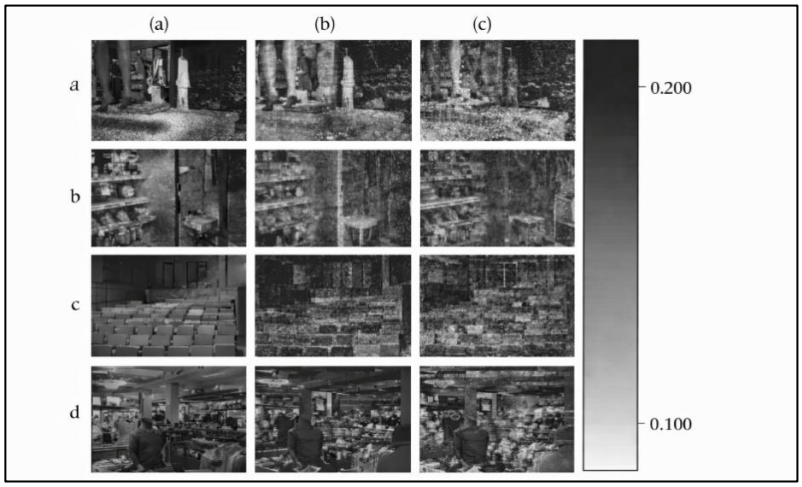

Figure 7. Error Map of DIODE Test Output

The error maps in Figure 7 represents the absolute difference between the predicted depth and the ground truth depth of each scene. Bright areas denote greater prediction error,

which is typically found at the boundaries of objects, reflective surfaces, and non-textured regions. Darker areas imply low error indicating that the model correctly predicts depth in planar and well-structured areas. In general, the error distribution shows that the proposed model is characterized by a high structural consistency and restrains large deviations in most of the scenes.

4.6 Statistical Validation

To guarantee the soundness and repeatability of the reported findings, statistical validation is conducted on a set of independent training runs. Each of the proposed model and the CNN baseline is trained three times with randomly chosen initializations, and the average and standard deviation of the evaluation metrics are provided.

Table 6 shows the statistical results from the proposed model trained on the NYU Depth V2 dataset, which had been obtained as an AbsRel of 0.080 ± 0.002 and an RMSE of 0.0290 ± 0.006, indicating that there is low variance between runs. The low standard deviation implies stable convergence and steady performance, which can be explained by the uncertainty-aware formulation of the loss that helps to reduce the impact of noisy gradients. To further determine the importance of the improvements made, paired t-tests are used to compare the CNN baseline with the proposed model. The statistically significant ($p < 0.01$) improvement in AbsRel indicates that the observed improvement in performance is not due to random variation. Additionally, the uncertainty measures do not vary considerably between runs, where NLL = 0.168 + 0.004, and ECE = 0.041 + 0.002. This means that the probabilistic forecasts are always well-calibrated.

Table 6. Statistical Validation (NYU Depth V2)

MODEL	ABSREL ↓	RMSE ↓	RMSE-LOG ↓
CNN BASELINE	0.098 ± 0.004	0.335 ± 0.010	0.052 ± 0.003
PROPOSED (BN-CNN)	0.080 ± 0.002	0.290 ± 0.006	0.044 ± 0.002

In general, the statistical analysis confirms that the proposed framework is not only more accurate but also more stable and reliable under various training conditions.

5. Conclusions

The proposed study introduces a hybrid probabilistic and deep learning model that estimates monocular depth for indoor environments while overcoming some weaknesses of deterministic approaches. By combining Bayesian modeling of uncertainty with CNN-based supervision, the model captures depth uncertainty and ensures an accurate measurement of metric depths. The Bayesian Network contributes to the probabilistic depth estimation, increasing spatial consistency and precision. The evaluation of the proposed algorithm based on the NYU Depth V2 benchmark demonstrates high performance (AbsRel: 0.080, RMSE: 0.290, RMSE-log: 0.044), as well as efficient generalization on the DIODE benchmark. Additionally, the suggested algorithm provides the ability to estimate confidence levels necessary for robotics applications, such as robotic navigation and augmented reality. Nevertheless, at the moment, the method fails to incorporate spatial correlations and consider dynamic and occluded scenes, which are directions for further study.

Compliance with Ethical Standards

This study is not part of any funding. The authors declare no conflict of interest. This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Sanz, Pablo Revuelta, Belén Ruiz Mezcua, and José M. Sánchez Pena. "Depth Estimation-An Introduction." In Current Advancements in Stereo Vision. IntechOpen, 2012.

[2] Dijk, Tom van, and Guido de Croon. "How do Neural Networks See Depth in Single Images?." In Proceedings of the IEEE/CVF international conference on computer vision, 2019, 2183-2191.

[3] Khan, Faisal, Saqib Salahuddin, and Hossein Javidnia. "Deep Learning-Based Monocular Depth Estimation Methods—A State-of-the-Art Review." Sensors 20, no. 8 (2020): 2272.

[4] Zhao, Chaoqiang, Qiyu Sun, Chongzhen Zhang, Yang Tang, and Feng Qian. "Monocular Depth Estimation Based on Deep Learning: An Overview." Science China Technological Sciences 63, no. 9 (2020): 1612-1627.

[5] Masoumian, Armin, Hatem A. Rashwan, Julián Cristiano, M. Salman Asif, and Domenec Puig. "Monocular Depth Estimation Using Deep Learning: A Review." Sensors 22, no. 14 (2022): 5353.

[6] Wofk, Diana, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. "Fastdepth: Fast Monocular Depth Estimation on Embedded Systems." In 2019 International Conference on Robotics and Automation (ICRA), pp. 6101-6108. IEEE, 2019.

[7] Kendall, Alex, and Yarin Gal. "What Uncertainties do We Need in Bayesian Deep Learning for Computer Vision?." Advances in neural information processing systems 30 (2017).

[8] Liu, Miaomiao, Mathieu Salzmann, and Xuming He. "Discrete-Continuous Depth Estimation from a Single Image." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, 716-723.

[9] Cao, Yuanzhouhan, Tianqi Zhao, Ke Xian, Chunhua Shen, Zhiguo Cao, and Shugong Xu. "Monocular Depth Estimation with Augmented Ordinal Depth Relationships." IEEE Transactions on Circuits and Systems for Video Technology 30, no. 8 (2019): 2674-2682.

[10] Hu, Junjie, Yan Zhang, and Takayuki Okatani. "Visualization of Convolutional Neural Networks for Monocular Depth Estimation." In Proceedings of the IEEE/CVF international conference on computer vision, 2019, 3869-3878.

[11] Ji, Rongrong, Ke Li, Yan Wang, Xiaoshuai Sun, Feng Guo, Xiaowei Guo, Yongjian Wu, Feiyue Huang, and Jiebo Luo. "Semi-Supervised Adversarial Monocular Depth

Estimation." IEEE transactions on pattern analysis and machine intelligence 42, no. 10 (2019): 2410-2422.

[12] Nishimura, Mark, David B. Lindell, Christopher Metzler, and Gordon Wetzstein. "Disambiguating Monocular Depth Estimation with a Single Transient." In European Conference on Computer Vision, Cham: Springer International Publishing, 2020, 139-155.

[13] Huynh, Lam, Matteo Pedone, Phong Nguyen, Jiri Matas, Esa Rahtu, and Janne Heikkilä. "Monocular Depth Estimation Primed by Salient Point Detection and Normalized Hessian Loss." In 2021 International Conference on 3D Vision (3DV), IEEE, 2021, 228-238.

[14] Huynh, Lam, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkilä. "Guiding Monocular Depth Estimation Using Depth-Attention Volume." In European Conference on Computer Vision, Cham: Springer International Publishing, 2020, 581-597.

[15] Piccinelli, Luigi, Christos Sakaridis, and Fisher Yu. "Idisc: Internal Discretization for Monocular Depth Estimation." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, 21477-21487.

[16] Shao, Shuwei, Zhongcai Pei, Weihai Chen, Ran Li, Zhong Liu, and Zhengguo Li. "Urcdc-Depth: Uncertainty Rectified Cross-Distillation with Cutflip for Monocular Depth Estimation." IEEE Transactions on Multimedia 26 (2023): 3341-3353.

[17] Agarwal, Ashutosh, and Chetan Arora. "Attention Attention Everywhere: Monocular Depth Prediction with Skip Attention." In Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2023, 5861-5870.

[18] Godard, Clément, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. "Digging Into Self-Supervised Monocular Depth Estimation." In Proceedings of the IEEE/CVF international conference on computer vision, 2019, 3828-3838.

[19] Zhou, Tinghui, Matthew Brown, Noah Snavely, and David G. Lowe. "Unsupervised Learning of Depth and Ego-Motion from Video." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 1851-1858.

[20] Silberman, Nathan, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. "Indoor Segmentation and Support Inference from Rgbd Images." In European conference on computer vision, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, 746-760.

[21] Fu, Huan, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. "Deep Ordinal Regression Network for Monocular Depth Estimation." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, 2002-2011.

[22] Yuan, Weihao, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. "Neural Window Fully-Connected Crfs for Monocular Depth Estimation." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, 3916-3925.

[23] Li, Zhenyu, Zehui Chen, Xianming Liu, and Junjun Jiang. "Depthformer: Exploiting Long-Range Correlation and Local Information for Accurate Monocular Depth Estimation." Machine Intelligence Research 20, no. 6 (2023): 837-854.

[24] Bae, Gwangbin, Ignas Budvytis, and Roberto Cipolla. "Irondepth: Iterative Refinement of Single-View Depth Using Surface Normal and its Uncertainty." arXiv preprint arXiv:2210.03676 (2022).

[25] Bhat, Shariq Farooq, Ibraheem Alhashim, and Peter Wonka. "Adabins: Depth Estimation Using Adaptive Bins." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, 4009-4018.

[26] Li, Zhenyu, Xuyang Wang, Xianming Liu, and Junjun Jiang. "Binsformer: Revisiting Adaptive Bins for Monocular Depth Estimation." IEEE Transactions on Image Processing 33 (2024): 3964-3976.

[27] Vasiljevic, Igor, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele et al. "Diode: A Dense Indoor and Outdoor Depth Dataset." arXiv preprint arXiv:1908.00463 (2019).