EMVSNet: Evidential Multi-View Stereo Reconstruction for Sampling-free Depth and Uncertainty Estimation

Christian Grannemann, Max Mehltretter

Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, Germany
(grannemann, mehltretter)@ipi.uni-hannover.de

Keywords: Evidential Deep Learning, Deep Evidential Regression, 3D Reconstruction, Single-Pass Inference

Abstract

We present EMVSNet, a sampling-free Multi-View Stereo (MVS) method that, to the best of our knowledge, is the first to integrate Evidential Deep Learning into MVS. Given a set of overlapping images, our method predicts a depth value together with its associated uncertainty per pixel of a reference image, incorporating uncertainty from aleatoric and epistemic sources. Specifically, we use an existing convolutional neural network architecture designed for MVS as backbone and extend it to regress evidential parameters per pixel, describing the probability distribution over the depth corresponding to this pixel. In contrast to existing MVS methods that often neglect epistemic uncertainty or obtain it via sampling at inference, our evidential formulation does not require sampling, but enables single-pass inference. We evaluate the uncertainty estimation capabilities of our method using two publicly available datasets and compare the depth predictions against a deterministic variant. The experimental results demonstrate that EMVSNet achieves competitive depth accuracy while, at the same time, providing uncertainty estimates that enable us to reliably rank depth estimates according to their risk of being incorrect and to automatically identify out of distribution data. Our model shows only slightly increased inference time compared to a deterministic baseline while giving comparable uncertainty estimates to an computationally expensive sampling based approach, marking a first step towards real-time capable uncertainty estimation for image-based 3D reconstruction. Our code is available at: `https://github.com/BuTTerK3ks/EMVSNet`.

1. Introduction

Learning-based approaches using deep neural networks have positively impacted many tasks in photogrammetry and computer vision. Multi-View Stereo reconstruction, referring to the reconstruction of per-pixel depth for a reference image by exploiting multiple, overlapping source images, has likewise benefited from recent advances, e.g., via learned cost volume generation, regularization and view aggregation (Wang et al., 2024). While these advancements have mainly improved accuracy, information on the uncertainty of depth predictions is often a central requirement, in particular, for safety-critical downstream use in real world scenarios, such as automated driving and medical imaging. In such use cases, wrong depth estimates can lead to catastrophic outcomes and having information about their uncertainty might prevent wrong decisions. More generally, knowledge on the uncertainty commonly improves the accuracy of a 3D reconstruction if depth estimates from multiple images are combined with each other or with observations from other types of sensors.

For our work, we adopt the standard decomposition of predictive uncertainty into aleatoric and epistemic components (Kendall and Gal, 2017): Aleatoric uncertainty captures random variability inherent in the observations, given as data noise. Typical sources include sensor noise, radiometric effects like poor texture and view-dependent reflections and weak geometric configurations (e.g., small baselines between the projection centers). This kind of uncertainty cannot be reduced by using more training data from the same data distribution, as the source of the uncertainty lies in the data itself. Epistemic uncertainty, in contrast, reflects uncertainty about the model and its parameters due to an unsuitable model structure or due to limited, biased, or non-representative training data. Thus, epistemic uncertainty typically increases under Out-Of-Distribution (OOD) inputs, for which the learned priors and image matching heuristics become unreliable. This kind of uncertainty is often reducible using more diverse training data that better represents the operational domain, by increasing the model’s capacity, by adjusting its layout or by optimizing the training process (e.g., by changing hyperparameters).

While multiple deep learning-based MVS methods exist in the literature that estimate uncertainty alongside the corresponding depth (Xu et al., 2021; Su et al., 2022), their limitations can be summarized in two main aspects: (i) Some methods solely model the aleatoric uncertainty and neglect the epistemic type, which limits the expressiveness of the estimated uncertainty. (ii) Other methods model both, aleatoric and epistemic uncertainty, but rely on sampling from the posterior distribution to approximate the latter (e.g., via Monte Carlo Dropout or deep ensembles). Such sampling requires multiple predictions during inference, making it computationally expensive.

In the present work, we tackle these limitations by adapting Evidential Deep Learning (EDL) to MVS. In EDL (Sensoy et al., 2018), a neural network predicts the parameters of a prior over the likelihood’s parameters rather than over the target (here, the depth) directly. In other words, for each input, the network outputs the parameters of a higher-level distribution that encodes its belief about the mean and the second central moment of the assumed likelihood. The final prediction is then given by the marginal distribution obtained by integrating the likelihood over this prior - yielding a closed-form expression in our setting. Specifically, we present the Evidential Multi-View Stereo Network (EMVSNet): Given a set of stereo images, EMVSNet estimates a depth alongside the corresponding aleatoric and epistemic uncertainty for each pixel in a reference image, without the need for sampling at inference. This network is trained end-to-end and only needs an assump-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-259-2026

tion regarding the shape of the prior distribution to model the uncertainty of predictions. The uncertainty is learned from the predictions within the training process, so it includes the uncertainty of all preliminary steps (like sensor noise or the estimation of relative orientation parameters). The main contributions of the present work are:

- We present **EMVSNet**, the first sampling-free MVS method to jointly predict the depth and its corresponding uncertainty, considering aleatoric and epistemic components.
- We show that EMVSNet delivers more accurate depth estimates than a deterministic baseline while additionally providing uncertainty information with only marginally increased runtime.
- We carry out experimental evaluations on a publicly available dataset analyzing the capability of our method to identify erroneous depth estimates based on the uncertainty, and compare the results against a network variant which uses stochastic sampling. On a second dataset we perform detection of OOD samples, indicated by increased epistemic uncertainty.

2. Related Work

2.1 Uncertainty in Depth Estimation

The reconstruction of depth information from a single, two or a collection of overlapping images is tackled by various methods as shown in related survey articles (Zhang, 2025; Tosi et al., 2025; Feng et al., 2025). While a wide variety of approaches have emerged that focus on improving the accuracy of a 2.5D reconstruction, incorporating uncertainty estimation has been neglected in many cases. Nevertheless, some exceptions exist:

Mono- and Binocular In the context of monocular depth estimation, some methods consider the uncertainty of their results, e.g., by predicting the mean and variance of a Gaussian distribution assumed over depth and by fusing these prediction with respect to the camera movement across multiple frames (Yang et al., 2019). Wang et al. (2019) model aleatoric uncertainty via test-time augmentation, while using Monte Carlo Dropout (MCD) for estimating epistemic uncertainty. While also applying MCD for epistemic uncertainty estimation, Chen et al. (2021) predict per-pixel aleatoric uncertainty by regressing the standard deviations of the reprojected 2D coordinates. Poggi et al. (2020) present an approach to estimate depth in a self-supervised manner using a photometric loss function. They derive a per-pixel mean and log-variance describing the aleatoric uncertainty of the depth via a Gaussian negative log-likelihood loss. Additionally they compare epistemic uncertainty estimation based on MCD, flip-consistency and ensembles. Hornauer and Belagiannis (2022) predict an uncertainty value for every pixel which is learned using a separate loss term that builds on the gradients of a depth-consistency loss term, instead of using depth errors directly. In (Marsal et al., 2024), the prediction of a per-pixel depth distribution is learned by optimizing a probabilistic reconstruction likelihood over the depth, which is calibrated by self-distillation.

For the binocular stereo case, Mehltretter (2022) trains a Bayesian neural network with variational inference to jointly predict the depth-related disparity and both, aleatoric and epistemic uncertainty. Chen et al. (2023b) concentrate on aleatoric uncertainty and propose a loss function that minimizes the difference between the distribution of the predicted uncertainty and the distribution of disparity errors via a differentiable soft-histogram and Kullback-Leibler term. Jing et al. (2023) compute a variance-based (aleatoric) uncertainty map from a computed cost volume, which encodes the matching costs between potentially homologous points in stereo images.

Multi-View Stereo For multi-view stereo, Song et al. (2023) propose a Bayesian probability volume to estimate the aleatoric uncertainty: For consecutive frames, they project the previous depth estimation and uncertainty, represented by the variance within the cost volume, to the current frame to obtain a depth prior. Combining this prior with the current likelihood in a Bayesian filtering step, they build a refined cost volume. Chen et al. (2023a) derive information about aleatoric uncertainty from the distribution across depth hypothesis within the probability volume (an alternative formulation of the cost volume, transferring matching costs into matching probabilities). Beyond modeling uncertainty in image space, Liao and Waslander (2024) model aleatoric uncertainty directly in object space: for each object, they predict a Gaussian latent code and pass samples of this code through a Signed Distance Function decoder to obtain per-vertex uncertainties on the reconstructed 3D mesh. The uncertainties derived from multiple images are fused via precision-weighted Gaussian updates in latent space. Lu et al. (2025) derive aleatoric uncertainty from the variation of the probability volume and introduce an uncertainty-aware cost–volume aggregation: pairwise uncertainty maps, given by the cost volume, guide the adaptive aggregation of group-wise correlation volumes within a coarse-to-fine cascade, supervised by an additional uncertainty loss. Zhao et al. (2024) integrate an epipolar Transformer to better encode three-dimensional relations along epipolar lines and add an uncertainty-guided sampling module that measures the dispersion of the per-pixel probability volume to dynamically narrow the depth hypotheses in the next stage, enabling single-pass inference without sampling.

In contrast to our EDL-based approach, none of the mentioned methods provide a sampling-free approximation of epistemic uncertainty. In most cases, the distribution of probabilities within a cost volume is taken as basis to estimate uncertainty. As the network's activations fluctuate more for unseen data, the so-estimated uncertainty may include an epistemic part (Postels et al., 2020), but explicitly describing epistemic uncertainty is not possible this way. The only work considering epistemic uncertainty for multi-view stereo reconstruction has been presented by Xu et al. (2021); however, their method is based on MCD, requiring sampling the posterior distribution during inference.

2.2 Evidential Deep Learning

In contrast to the approaches discussed so far, EDL allows to predict epistemic uncertainty directly, and does not build on the concept of sampling from the posterior distribution during inference to approximate it. Building on the theoretical concepts developed by Dempster (2008) and Shafer (1976), early works solely focus on modeling epistemic uncertainty (Malinin and Gales, 2018; Zhao et al., 2020; Nandy et al., 2020): These methods incorporate prior knowledge about the training distribution, keep the prior distribution fixed during training and only learn a mapping of individual predictions to this prior. This approach

has the advantage of fine control over the confidence behavior and results in an optimization objective with favorable convergence behavior. Later work further developed this idea into so-called posterior EDL networks (Sensoy et al., 2018, 2020): The basic idea is to learn input-dependent distribution parameters, using task-dependent ground truth for supervision (e.g., class labels or regression targets). This enables a decomposition into aleatoric uncertainty (captured by the likelihood) and epistemic uncertainty (captured by the prior). For classification tasks, the Dirichlet distribution is often used as prior over class probabilities, while the Normal-inverse Gaussian (NiG) distribution is commonly employed for regression tasks. As our defined task is the prediction of depth and its associated aleatoric and epistemic uncertainty, in the following, we concentrate on this posterior type of networks: Amini et al. (2020) leverage a NiG prior over a Gaussian likelihood, yielding a Student-*t* predictive distribution whose closed-form moments decompose variance into aleatoric and epistemic terms in a single pass; a regularizer penalizes confidence errors. Other regression based methods incorporate a Normal–Inverse–Wishart (NIW) prior (Meinert and Lavin, 2021) over a multivariate Gaussian distribution (Malinin et al., 2020) or an asymmetric Laplace distribution for quantiles (Hüttel et al., 2023).

From an application point of view, EDL is already used in a broad field of research areas (e.g., computer vision, natural language processing, life and natural sciences) as shown in recent surveys by Ulmer (2021) and Gao et al. (2024). For monocular depth estimation, Ye et al. (2024) use a NiG prior for single-pass depth and uncertainty estimation with a decomposition into aleatoric and epistemic uncertainty. An uncertainty regularizer is added that prevents gradient vanishing in high-uncertainty regions, preventing the network to generally predict high uncertainties. In (Menon et al., 2025), two evidential heads are used to predict semantic labels alongside depth, using both, regression and classification priors within the same architecture. Wang et al. (2022) adapt EDL to stereo matching by predicting NiG parameters for each disparity level in a cost volume and aggregating them using the matching probability. In training, they use two regularizers for error-aware evidence suppression and uncertainty-smoothness. The fusion of two evidential heads, a cost volume-based and a transformer-based, is presented by Lou et al. (2023). They use a Mixture of NiG (MoNIG) to fuse predicted distribution parameters. In contrast, Liu et al. (2024) adapt a Mixture-of-Gaussians model which replicates multiple disparity hypothesis per pixel. This approach shows trustworthy uncertainty predictions, especially under domain shift.

2.3 Discussion

The prediction of aleatoric and epistemic uncertainty associated with depth derived from images has been studied in multiple recent works. Especially for the aleatoric type, various approaches have been presented, which in most cases analyze the variability of the feature representation from different views. Estimating the inherent epistemic uncertainty is neglected in most cases or is build on sampling-based approaches such as ensembles or MCD, characterizing them as computationally expensive. In contrast, EDL offers a sampling-free approach to jointly estimate depth and the associated aleatoric and epistemic uncertainty. However, for the MVS case, to the best of our knowledge, there is no existing work building on EDL. This states a research gap that is - due to the outlined advantages of EDL - worth investigating.

3. Methodology

Given N overlapping color images $I \in \mathbb{R}^{3\times H_I\times W_I}$ of height H_I and width W_I, divided into one reference image I_r and $N-1$ source images I_s, all with known interior and exterior orientation parameters, our network predicts a depth $\hat{d}(p)$ and aleatoric $\hat{u}_a(p)$ and epistemic $\hat{u}_e(p)$ uncertainties for each pixel $p \in I_r$. To predict these quantities, our method is divided into two main parts: a backbone network and an evidential parameter estimation module. Given D depth hypotheses $z_i \in \{z_1,\dots,z_D\}$ resembling a depth range $[d_{\min}, d_{\max}]$, the backbone network takes the reference and the source images and constructs a probability volume $Q \in \mathbb{R}^{D\times H\times W}$ (see Sec. 3.1). We use H and W to denote the spatial size of the feature maps, corresponding to a four-times downsampling of the original image dimensions H_I and W_I. Within the subsequent evidential module, Q is processed at three scales k, incorporating backbone features through skip connections. For each scale, per-pixel NiG parameters $(\hat{\gamma}_k, \hat{\nu}_k, \hat{\alpha}_k, \hat{\beta}_k)$ are predicted and fused by MoNiG into a single per-pixel set $(\hat{\gamma}, \hat{\nu}, \hat{\alpha}, \hat{\beta})$ (see Sec. 3.2). During training, these parameters are optimized to yield accurate per-pixel depth and corresponding uncertainty estimates following the concept of EDL (see Sec. 3.3). An overview of our complete method is given in Figure 1.

3.1 Backbone network

According to (Wei et al., 2021), the backbone network consists of a feature extraction performed per image, a subsequent warping and a per-image cost volume creation, resembling matching costs from each source image to the reference image at all disparity levels. Finally, these cost volumes are refined and concatenated, which results in the combined probability volume Q.

Feature extraction All images I are processed by a shared-weight feature extractor f_θ to obtain feature maps $F_n = f_\theta(I_n) \in \mathbb{R}^{C\times H\times W}$, for $n \in \{1,..,N\}$ with $C = 4\times b$, where b is a chosen amount of base filters. These feature maps are refined by an intra-view adaptive-aggregation (AA) module (Wei et al., 2021), resulting in feature maps $\tilde{F}_n$ of the same size as F. This module is based on modulated deformable convolutions which learn spatial kernels, yielding a content-adaptive receptive field that expands in low-texture regions while preserving details near edges.

Warping and per-image cost volume creation For each depth hypothesis z and source image I_s, $\tilde{F}_s$ is warped from the coordinate system of the respective source image into the coordinate system of the reference image via a differentiable homography H, and a cost volume $C_s \in \mathbb{R}^{C\times D\times H\times W}$ is computed by squared feature differences:

$$C_s(z,p) \;=\; \left\| \tilde{F}_r(p) - H\big(\tilde{F}_s(p), z\big) \right\|_2^2 . \tag{1}$$

Cost volume combination The per-image cost volumes C_s are fused by an inter-view AA module to obtain an aggregated cost volume C_{agg}. This module applies learned, pixel-wise normalized attention weights w to adaptively combine information from all source images, emphasizing consistent matches and suppressing outliers:

$$C_{\mathrm{agg}}(z,p) = \frac{1}{N-1}\sum_{s=1}^{N-1}\big(1 + w_s(z,p)\big)\, C_s(z,p)\,. \tag{2}$$

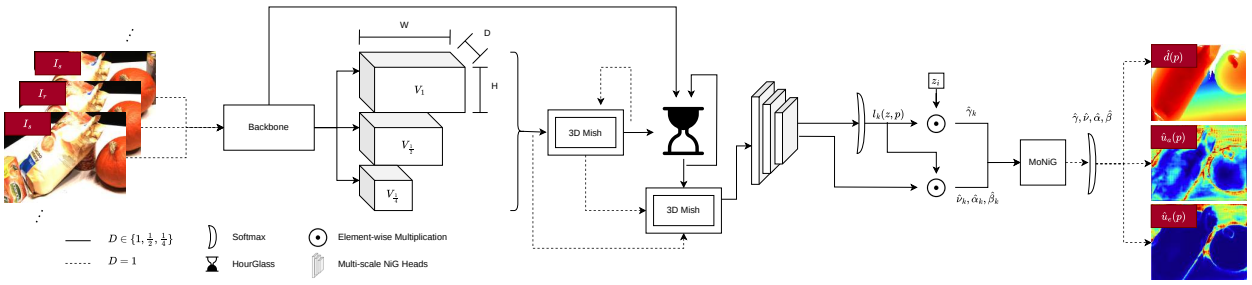

Figure 1. Network architecture of EMVSNet. Three volumes V_k with $k \in \{1, \frac{1}{2}, \frac{1}{4}\}$ are derived from backbone features and processed individually. Multi-scale NiG Heads retrieve the feature maps from different stages within the Hourglass network, incorporating skip connections from the backbone. Each scale is trilinearly upscaled to determine a probability $l_k(z,p)$ and NiG shape parameters $\hat{\gamma}_k, \hat{\nu}_k, \hat{\alpha}_k, \hat{\beta}_k$ per scale. These parameters are fused using MoNiG to obtain shape parameters of a combined NiG distribution, from which per pixel depth $\hat{d}(p)$ is predicted along corresponding aleatoric $\hat{u}_a(p)$ and epistemic $\hat{u}_e(p)$ uncertainty.

Cost volume regularization The aggregated cost volume is treated as a sequence of 2D slices along the depth dimension, which are processed one after another by a convolutional recurrent unit. In this way, information from neighboring depth hypotheses is aggregated while preserving the spatial structure through 2D convolutions: For each depth hypothesis z, the hidden state h_z of the convolutional recurrent unit combines the current matching evidence with the context from previous depth hypotheses, resulting in regularized multi-scale feature maps. These feature maps are passed to the evidential parameter estimation module via skip connections (see Sec. 3.2). Within the cost volume regularization module, a lightweight prediction head transforms these feature maps into per-depth logits $S(z,p)$, representing the matching cost of each pixel for each depth hypothesis. Applying a softmax across all hypotheses per pixel, yields the normalized probability volume Q:

$$Q(z,p) = \mathrm{softmax}_z\left(-\frac{S(z,p)}{\tau}\right), \tag{3}$$

which represents, for each pixel p, the probability distribution over all depth hypotheses and where τ is a hyperparameter used for normalization.

3.2 Evidential Parameter Estimation

The second part of our method aims at estimating the parameters of the higher-order distribution from which we finally derive the depth and its associated uncertainty. For the estimation of meaningful parameters of this distribution, we use information from different scales and apply a sophisticated network layout focusing on the fusion of the evidential parameters across scales: We first process the probability volume Q by 3D convolutions, extracting volumes of different scale. Each of these volumes is subsequently processed by an Hourglass network, enriched by skip connections, before estimating distribution parameters per scale in NiG heads. Finally, the estimated parameters per head are fused across scales. This part has been adapted for the MVS case from the work by Lou et al. (2023), who have originally presented a related network structure for the binocular stereo case. The multi-scale skip connections provided by the backbone encoder are received in this module and integrated into the volume representation at each scale before scale-wise fusion.

3D Convolutions Given the probability volume Q, downscaled volumes $V_k \in \mathbb{R}^{k(D\times H\times W)}$ are retrieved from Q by a factor $k \in \{1, \frac{1}{2}, \frac{1}{4}\}$ using trilinear interpolation, followed by softmax normalization along the depth. Further processing is done in sequential blocks consisting of 3D convolutions with a mish activation function (Misra, 2019), yielding volumes $W_k \in \mathbb{R}^{C\times k(D\times H\times W)}$. At each scale k, the corresponding skip features are trilinearly interpolated to match the volume's dimension and added elemetwise. Using the Mish activation function, compared to the Rectified Linear Unit, has the advantage that no strict negative cutoff is applied, thus it should stabilize the training process.

Hourglass network Each volume W_k is further processed by an Hourglass network (Newell et al., 2016) which combines information from different feature map regions via convolution-based down and up-sampling, reducing and increase spatial resolution. The global relation of feature information plays an important role for the estimation of depth as well as aleatoric and epistemic uncertainty, allowing to consider long-range dependencies in the 3D structure of the depicted scene. Upsampling of all volumes to input size yields to $X_k \in \mathbb{R}^{32\times D\times H\times W}$ feature maps within the NiG heads.

Evidential parameter estimation To estimate the four evidential parameters, 3D convolutions reduce the number of feature channels in X_k from 32 to 4, resulting for each scale k in a volume $Y_k \in \mathbb{R}^{4\times D\times H\times W}$. The first channel is normalized by the Softmax function along the depth dimension to obtain a probability $l_k(z,p)$. The depth estimation $\hat{\gamma}_k(p)$ for each pixel p is then obtained by regression over the depth hypotheses z_i, using $l_k(z,p)$:

$$\hat{\gamma}_k(p) = \sum_{i=1}^{D} z_i \, l_k(z_i, p)\,. \tag{4}$$

Additionally, the remaining channels are used to obtain the NiG parameters. The probability $l_k(z,p)$ is element-wise multiplied with each channel $m \in \{2,3,4\}$ and mapped into the valid NiG domain via the softplus function σ to obtain the shape parameters $\hat{y}_{k,m}(p) \in \{\hat{\nu}_k, \hat{\alpha}_k, \hat{\beta}_k\}$:

$$\hat{y}_{k,m}(p) = \sigma\left(\sum_z l_k(z,p)\, Y_k^{(m)}(z,p)\right). \tag{5}$$

Evidential parameter fusion The fusion of the parameters across all scales k is done using the *Mixture of Normal-inverse Gamma* method presented by Ma et al. (2021):

$$\forall\, p \in I_r: \quad \mathrm{NiG}_p(\hat{\gamma}, \hat{\nu}, \hat{\alpha}, \hat{\beta}) = \bigoplus_k \mathrm{NiG}_p(\hat{\gamma}_k, \hat{\nu}_k, \hat{\alpha}_k, \hat{\beta}_k)\,,$$

where $\oplus$ represents the NiG summation: a closed-form rule that fuses two NiG posteriors into a single NiG distribution while preserving conjugacy. Finally, we add one to $\hat{\alpha}$ to keep this quantity strictly positive.

3.3 Loss function

A key aspect of EDL is a properly tailored optimization objective that yields accurate predictions for the quantity of interest (here: depth) while simultaneously estimating the associated uncertainty. In our work, we adopt the Student-t marginal of Amini et al. (2020), obtained by integrating out the likelihood parameters (μ, σ^2) under a NiG prior:

$$\begin{aligned}\mathcal{L}_{\mathrm{NLL}} =& \frac{1}{2}\log\left(\frac{\pi}{\hat{\nu}}\right) - \hat{\alpha}\log\Omega \\ &+ (\hat{\alpha}+\frac{1}{2})\log(\hat{\nu}(\hat{\gamma}-d)^2+\Omega) \\ &+ \log\Gamma(\hat{\alpha}) - \log\Gamma(\hat{\alpha}+\frac{1}{2})\end{aligned} \tag{6}$$

with

$$\Omega = 2\hat{\beta}(1+\hat{\nu}) \qquad \Gamma(\hat{\alpha}) = \int_0^\infty x^{\hat{\alpha}-1}e^{-x}\,dx,$$

denoting d as reference depth. An additional regularization term penalizes high confidence (high $\hat{\nu}$, $\hat{\alpha}$) in case of a large prediction error $|d-\hat{\gamma}|$, preventing the network from becoming overconfident and stabilizing training given the non-identifiability of the Student-t marginal, resulting in our final loss function:

$$\mathcal{L} = \mathcal{L}_{\mathrm{NLL}} + \lambda_{reg}\,|d-\hat{\gamma}|\,(2\hat{\nu}+\hat{\alpha}), \tag{7}$$

with chosen λ_{reg} as weighting factor.

3.4 Uncertainty representation

The aleatoric ($\hat{u}_{\mathrm{a}}$) and epistemic ($\hat{u}_{\mathrm{e}}$) uncertainty is derived by:

$$\hat{u}_{\mathrm{a}}^2 = \frac{\hat{\beta}}{\hat{\alpha}-1}, \quad \hat{u}_{\mathrm{e}}^2 = \frac{\hat{\beta}}{\hat{\nu}(\hat{\alpha}-1)}, \tag{8}$$

where $(\hat{\alpha}, \hat{\beta}, \hat{\nu})$ jointly encode per-pixel confidence, with $\hat{\beta}$ and $\hat{\alpha}$ determining the total variance and $\hat{\nu}$ partitioning it between aleatoric and epistemic components.

4. Experiments

4.1 Datasets

For our experiments, we mainly use the DTU dataset (Aanæs et al., 2016), providing images of small objects under seven different lighting conditions, captured using a camera mounted on a robot arm, and ground truth depth gathered with a structured-light scanner. Each scene shows a single object from 49 or 64 viewing directions, with a high overlap between images; the images have a size of 1600×1200 pixels. Precise exterior orientation parameters and a controlled environment make this dataset especially suitable for MVS tasks. We use a custom split of the dataset into training, validation and test scenes (72/13/12) to have ground truth information for every split.

We use the Tanks and Temples (TnT) dataset (Knapitsch et al., 2017) to test our method's ability to identify a domain shift between training and inference by providing real world outdoor stereo images. Specifically, we use three scenes for testing (no training is done on this dataset), with $151 - 314$ images per scene and an image size of 1920×1080 pixels. The images are captured as a sequence with an object centric view and are oriented relatively to each other via Structure-from-Motion. The object size, method of capturing (handheld) and outdoor lighting mark a clear domain shift to the DTU dataset.

4.2 Implementation and training details

The Adam optimizer with an initial learning rate of 10^{-3} and a decay of 0.9 is used. Images are scaled to a resolution of 128×160 pixels for training and inference to reduce the computational effort. In each training iteration, we use 7 consecutively captured images from the DTU dataset, with the middle image serving as the reference image I_r and the remaining six as source images I_s. The considered depth range is defined as $d_{min} = 425\ mm$ to $d_{max} = 935\ mm$. We set $\lambda_{\mathrm{reg}} = 0.1$, number of base filters $b = 8$ and train for 10 epochs, using the parameter values of the checkpoint having the lowest validation loss for testing.

AA-RMVSNet developed by Wei et al. (2021) has been used as baseline for EMVSNet and is compared against it to verify the depth prediction capabilities. Additionally, we introduce a variant of this baseline, in which MCD is applied to multiple layers, following a common sampling-based uncertainty estimation approach. Specifically, one layer within the intra-view Adaptive Aggregation module and both forward branches of the convolutional recurrent unit within AA-RMVSNet are changed to integrate dropout with a rate of $10\,\%$ during training and inference. According to Mosquera Rojas et al. (2025), a dropout rate of $25\,\%$ provides a good balance between depth prediction accuracy and uncertainty estimation capabilities. In our experiment, however, a dropout rate of $10\,\%$ provided better results, but we kept the suggested 30 samples during inference.

4.3 Quality Metrics

To evaluate the effectiveness of our uncertainty prediction quantitatively, we analyze the correlation of the predicted uncertainty and the depth prediction error. For this purpose, we define a threshold τ_d, which sets the upper limit of the acceptable depth prediction error e_p; if this threshold is exceeded, this prediction is defined as erroneous and should be detected by our method. In addition, we use the Area Under the Curve of a Receiver Operating Curve (AUC-ROC), where the uncertainty predictions u_a and u_e for a given pixel are relatively ranked against the other uncertainty predictions within the same image. By setting a variable threshold within the range of u_a and u_e, respectively, and sweeping through this range, the ROC is derived. Furthermore, we report the Mean Absolute Error (MAE) of the depth predictions, which measures the average absolute deviation between the predicted and the ground-truth depth values over all evaluated pixels. The MAE provides an absolute measure of the overall depth prediction accuracy and complements the uncertainty-based detection metrics. Additionally, we report the F1-Score and the accuracy. For this purpose, we first perform a relative calibration of our uncertainty estimates: For predefined error thresholds of 2, 4 and 8 mm, we determine uncertainty thresholds τ_{max}, with the aim that uncertainty estimates exceeding the uncertainty threshold correspond to depth

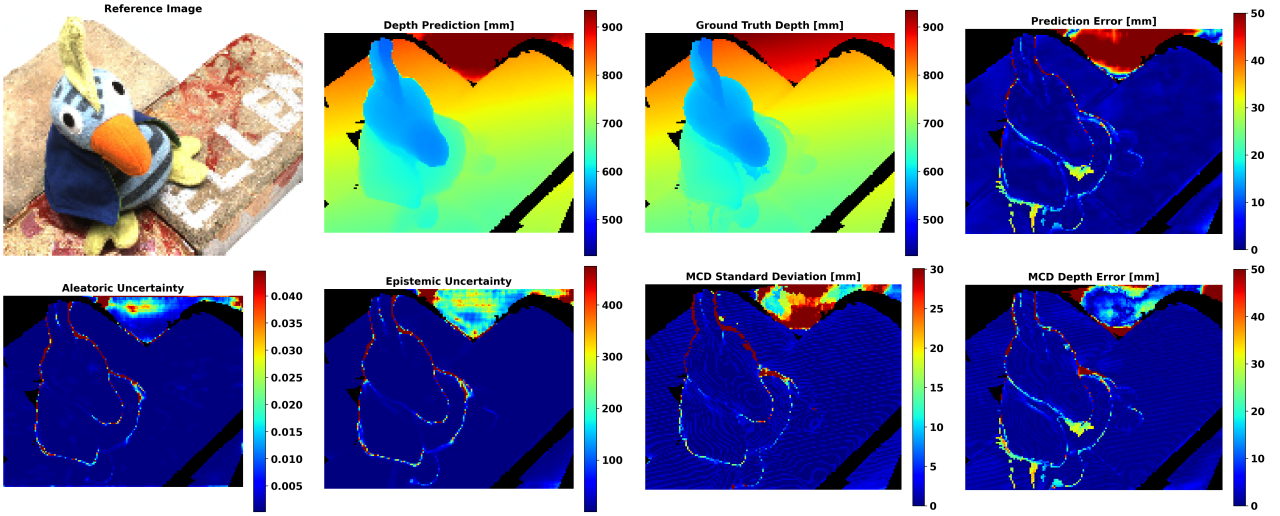

Figure 2. Qualitative results of EMVSNet’s depth and uncertainty prediction. We show depth and uncertainty estimates in comparison to the ground truth depth. For the MCD variant, the illustrations show the mean predicted depth with associated variance between predictions. For all illustrations, the highest and lowest two percent of values are clipped to improve visibility. Pixels for which no valid ground truth depth is available are masked out and shown in black.

Model	N	z	Time ↓	VRAM ↓	MAE [mm] ↓	2mm ↓	4mm ↓	8mm ↓	16mm ↓	32mm ↓
EMVSNet	7	128	892 ms	3232 MB	5.1	19.2%	9.9%	5.9%	4.1%	2.9%
EMVSNet	7	64	479 ms	1687 MB	5.4	22.7%	11.6%	6.6%	4.3%	3.1%
AA-RMVSNet	7	128	545 ms	1619 MB	5.9	27.7%	9.9%	5.6%	4.1%	3.1%
MCD	7	128	17.04 s	1619 MB	6.6	30.8%	11.7%	6.3%	4.5%	3.5%

Table 1. Depth prediction performance comparison of EMVSNet and AA-RMVSNet, including an MCD based stochastic sampling variant (showing the average results over 30 samples), on the DTU test set with a selected numbers of views N and depth hypothesis z. Each image of a scene is taken once as reference image, and depth estimates are evaluated on foreground object pixels. We report the MAE for the whole set as well as the percentage of pixels exceeding the defined error thresholds. We report two EMVSNet variants configured to match AA-RMVSNet in computational cost and in the number of depth hypotheses, respectively. Performance is reported for inference on a Nvidia RTX 3090.

estimates with an error larger than the error threshold. For this purpose, we chose τ_{max} so that it maximizes the F1-score on the validation set. Based on this uncertainty threshold we then classify depth estimates into potentially correct and incorrect estimates and compute the F1-score as well as the accuracy.

4.4 Depth prediction results

In our first experiment, we investigate the accuracy of the depth estimates of EMVSNet against both other methods. Additionally, we show results for a variant of this base method incorporating MCD, which we use to compare uncertainty estimates of our evidential deep learning based network to. The goal is to equip our MVS method with the capability of estimating uncertainty without deteriorating its depth estimates. As shown in Table 1, this goal has been achieved, with EMVSNet actually leading to higher accuracy of the depth estimation: For both, the mean absolute error (MAE) and the rate of erroneous estimates under a small threshold of 2 mm, an improvement of 0.81 mm in MAE and 8.5 percent points for the smallest threshold can be observed. We achieve MAE improvements also when we reduce the number of depth hypotheses z to meet computational resource requirements compared to the base method.

4.5 Uncertainty prediction results

In our second experiment, we evaluate the quality of the predicted aleatoric and epistemic uncertainties. In addition, we consider a combined uncertainty measure, reflecting practical scenarios in which unreliable predictions are filtered based on total uncertainty, irrespective of its decomposition. To this end, we define the combined uncertainty as the total predictive standard deviation applying quadratic error propagation, i.e., computing the sum of aleatoric and epistemic variances.

As shown in Figure 2, EMVSNet is generally able to correctly detect areas of increased depth error: Qualitatively, we see an increased error in depth estimation for regions where higher uncertainty is predicted. This can, in particular, be seen in image regions that are generally challenging in the context of image matching, such as at the borders of objects where depth abruptly changes significantly and occlusion for certain viewpoints occurs. Compared to the qualitative results of the MCD model, we see increased error in depth prediction for individual pixels in the same, challenging regions. Also pixels with high variance of depth predictions across samples matches regions of increased uncertainty predicted by EMVSNet. The massively reduced runtime of EMVSNet compared to the MCD-based variant (28 times using 30 samples for MCD) underlines the effectiveness of our evidential single-pass approach. Interestingly, aleatoric and epistemic uncertainty tend to highlight the same areas in the image for having increased uncertainty, although increased epistemic uncertainty would only be expected for unfamiliar objects. This seems to indicate a remaining issue in the decomposition of uncertainty. Additionally note that the

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-259-2026 | © Author(s) 2026. CC BY 4.0 License.

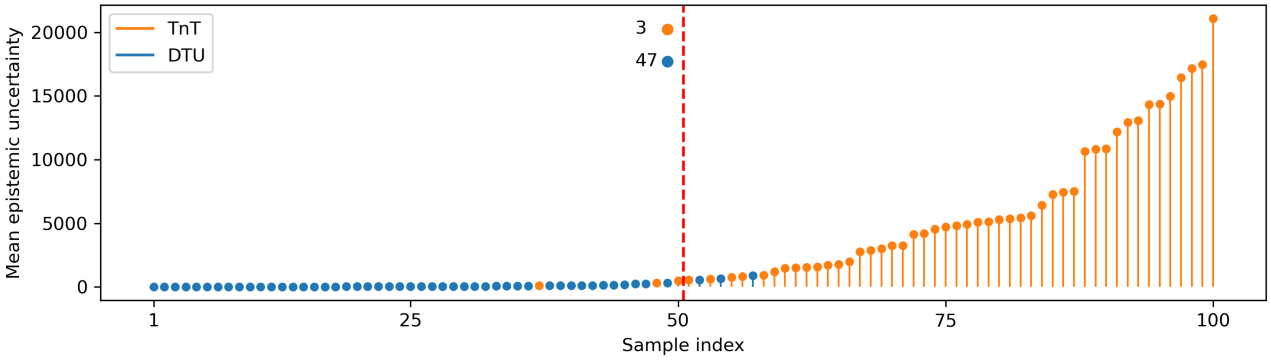

Figure 3. Mean epistemic uncertainty results of the domain shift experiment, with our model trained on DTU and tested on DTU and three TnT scenes (Lighthouse, M60, Panther). Data from TnT, indicating a domain shift, shows the tendency to produce increased mean epistemic uncertainty. The red line indicates split of the dataset in half, additionally indicating how many samples of which dataset can be found within the lower 50 percent of values. Perfect separation of data under domain shift would be achieved if all samples from DTU were on the left and all samples from TnT to the right of this line.

ground truth depth maps are not perfect but contain errors (e.g., visible in the given example at the beak), which has a negative impact on the training and leads to a false negative assessment of depth estimates.

The quantitative results, shown in Table 2, reveal that EMVS-Net is able to achieve comparable, in some cases even superior, results compared to MCD. Additionally, our method is able to consider aleatoric and epistemic uncertainty, while the uncertainty obtained via MCD is commonly assumed to be purely epistemic. For the the two uncalibrated metrics (ROC-AUC and Accuracy) and for the calibrated F1-score a strong correlation of predicted uncertainty and real error can be observed. For interpretation of the F1-score, the low amount of high-error pixels shown in Table 1 needs to be considered, as naive random guessing would lead to very low scores. It can be observed that epistemic uncertainty seems to correlate strongly with the depth estimation error. Additionally, the combined uncertainty measures are mainly influenced by the epistemic uncertainty. These observations support the claim that the applied decomposition into aleatoric and epistemic components is of limited practical use. Note, however, that our method takes both sources of uncertainty into consideration, which itself is crucial to determine the overall uncertainty.

Generally it can be observed that, in absolute terms, the aleatoric and epistemic uncertainty predictions differ by multiple orders of magnitude and are therefore not directly comparable. This shows that the estimated uncertainties are only interpretable as relative values; the same observation is also made in the literature: Meinert et al. (2023) show that the relative classification of uncertainty can be highly effective using EDL methods, but the approach lacks especially (i) the capability of distinguishing between aleatoric and epistemic uncertainty and (ii) does not provide metric uncertainty values that can be interpreted as such, for example, to estimate the standard deviation of depth predictions in pixels or metric units.

Even with this limitation, relative uncertainty measures should enable the network to identify a domain shift between training and test. To test this ability, we apply our method trained on the DTU dataset to a total of 100 randomly selected test samples, each consisting of 7 images. Half of these samples is taken from our DTU test set, the other half is taken from TnT. We calculate the mean epistemic uncertainty across all pixels for the training domain, thus pixels on the objects for DTU, and across all pixels for TnT. We sort the mean epistemic uncertainty per pixel for the reference image in ascending order and indicate which dataset the sample belongs to, as shown in Figure 3. In this case, we see only three samples misclassified. In general, the samples from a dataset different than the training samples show a clearly increased mean epistemic uncertainty compared to the samples from the same dataset. This indicates that our method is generally able to detect out-of-distribution cases.

While our method is able to identify erroneous depth estimates by assigning them a high uncertainty, limitations can be seen with respect to the ability to estimate uncertainty in terms of metric units and the decomposition into aleatoric and epistemic uncertainty. Thus, future work should focus on (i) improving the stochastic model encoded in the optimization objective, (ii) the calibration on the validation set, and (iii) the decomposition of aleatoric and epistemic uncertainty, which has also been identified as limitation in the literature. From a technical perspective, we additionally aim to further optimize memory usage to enable the processing of larger input images.

5. Conclusions

We introduced EMVSNet, a sampling-free multi-view stereo reconstruction method that predicts per-pixel depth together with the associated aleatoric and epistemic uncertainty via EDL. Our neural network architecture extends an existing MVS network used as backbone with an evidential head incorporating skip connections. Our architecture estimates NiG parameters at multiple scales and fuses them in closed form. EMVSNet achieves an comparable depth prediction accuracy in relation to a deterministic baseline, while providing informative uncertainty estimates; the correlation between the observed depth error and the estimated uncertainty allows for a reliable identification of erroneous depth estimates based on the estimated uncertainty. These results are achieved by single-pass inference, i.e., only a minor computational overhead is caused compared to the baseline and a clear reduction in inference time is achieved compared to sampling-based approaches, such as Monte Carlo Dropout.

Thus, EMVSNet demonstrates that evidential modeling can provide useful uncertainty estimates, encompassing aleatoric and epistemic sources, for MVS without the need for sampling

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-259-2026 | © Author(s) 2026. CC BY 4.0 License.

Metric	τ_d [mm]	Method	A	E	C
ROC AUC	2	EMVSNet	71%	87%	87%
		MCD	–	76%	–
	4	EMVSNet	82%	93%	93%
		MCD	–	87%	–
	8	EMVSNet	86%	93%	93%
		MCD	–	87%	–
F1	2	EMVSNet	54%	70%	70%
		MCD	–	55%	–
	4	EMVSNet	50%	64%	64%
		MCD	–	61%	–
	8	EMVSNet	53%	59%	59%
		MCD	–	62%	–
Acc	2	EMVSNet	64%	78%	78%
		MCD	–	72%	–
	4	EMVSNet	89%	91%	91%
		MCD	–	92%	–
	8	EMVSNet	94%	95%	95%
		MCD	–	96%	–

Table 2. ROC, F1-Score and Accuracy for thresholds $\tau_d \in \{2, 4, 8\}$ mm, reported for aleatoric (A), epistemic (E) and combined (C) uncertainty for EMVSNet. Monte Carlo Dropout (MCD) uncertainties are interpreted as purely epistemic.

at inference, potentially enabling systems to take more informed decisions in real-world applications. Decision-making algorithms in applications such as autonomous driving may modulate their responses to predictions based on the potential harmfulness of an action, while explicitly accounting for associated uncertainty estimates. If properly working, the delineation into aleatoric and epistemic uncertainty can be useful in active learning frameworks, where increased epistemic uncertainty shows underrepresented data while the combined uncertainty can be analyzed to only learn on reliable information.

However, and in line with recent analyses in the literature, our evidential uncertainty estimates behave primarily as relative proxies and require sophisticated calibration to transfer them into metric units. Moreover, as recent research points out, the precision of the decomposition into aleatoric and epistemic components needs to be questioned. Addressing these limitations, in future work, we will investigate alternative uncertainty formulations and optimization objectives in the evidential framework. Moreover, we will investigate the integration of the derived uncertainty into further processing steps to derive a complete 3D reconstruction as well as the use of evidential methods in alternative 3D reconstruction approaches, such as Gaussian Splatting and NeRF.

References

Aanæs, H., Jensen, R. R., Vogiatzis, G., Tola, E., Dahl, A. B., 2016. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2), 153–168.

Amini, A., Schwarting, W., Soleimany, A., Rus, D., 2020. Deep evidential regression. *Advances in neural information processing systems*, 33, 14927–14937.

Chen, H., Huang, Y., Tian, W., Gao, Z., Xiong, L., 2021. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10379–10388.

Chen, J., Yu, Z., Ma, L., Zhang, K., 2023a. Uncertainty awareness with adaptive propagation for multi-view stereo. *Applied Intelligence*, 53(21), 26230–26239.

Chen, L., Wang, W., Mordohai, P., 2023b. Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17235–17244.

Dempster, A. P., 2008. Upper and lower probabilities induced by a multivalued mapping. *Classic works of the Dempster-Shafer theory of belief functions*, Springer, 57–72.

Feng, S., Wu, X., Cao, J., 2025. A survey of multi-view stereo 3D reconstruction algorithms based on deep learning. *Digital Signal Processing*, 105291.

Gao, J., Chen, M., Xiang, L., Xu, C., 2024. A comprehensive survey on evidential deep learning and its applications. *arXiv preprint arXiv:2409.04720*.

Hornauer, J., Belagiannis, V., 2022. Gradient-based uncertainty for monocular depth estimation. *European Conference on Computer Vision*, Springer, 613–630.

Hüttel, F. B., Rodrigues, F., Pereira, F. C., 2023. Deep Evidential Learning for Bayesian Quantile Regression. *arXiv preprint arXiv:2308.10650*.

Jing, J., Li, J., Xiong, P., Liu, J., Liu, S., Guo, Y., Deng, X., Xu, M., Jiang, L., Sigal, L., 2023. Uncertainty guided adaptive warping for robust and efficient stereo matching. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3318–3327.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Knapitsch, A., Park, J., Zhou, Q.-Y., Koltun, V., 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4), 1–13.

Liao, Z., Waslander, S. L., 2024. Multi-view 3d object reconstruction and uncertainty modelling with neural shape prior. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3098–3107.

Liu, W., Wang, X., Wang, L., Cheng, J., Liu, F., Yang, X., 2024. Gaussian Mixture based Evidential Learning for Stereo Matching. *arXiv preprint arXiv:2408.02796*.

Lou, J., Liu, W., Chen, Z., Liu, F., Cheng, J., 2023. Elfnet: Evidential local-global fusion for stereo matching. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17784–17793.

Lu, P., Cai, Y., Yang, J., Wang, D., Wu, T., 2025. Uanet: uncertainty-aware cost volume aggregation-based multi-view stereo for 3D reconstruction. *The Visual Computer*, 41(7), 4567–4580.

Ma, H., Han, Z., Zhang, C., Fu, H., Zhou, J. T., Hu, Q., 2021. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. *Advances in Neural Information Processing Systems*, 34, 6881–6893.

Malinin, A., Chervontsev, S., Provilkov, I., Gales, M., 2020. Regression prior networks. *arXiv preprint arXiv:2006.11590*.

Malinin, A., Gales, M., 2018. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31.

Marsal, R., Chabot, F., Loesch, A., Grolleau, W., Sahbi, H., 2024. Monoprob: Self-supervised monocular depth estimation with interpretable uncertainty. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3637–3646.

Mehltretter, M., 2022. Joint estimation of depth and its uncertainty from stereo images using bayesian deep learning. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2, 69–78.

Meinert, N., Gawlikowski, J., Lavin, A., 2023. The unreasonable effectiveness of deep evidential regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, 9134–9142.

Meinert, N., Lavin, A., 2021. Multivariate deep evidential regression. *arXiv preprint arXiv:2104.06135*.

Menon, R., Dengler, N., Pan, S., Chenchani, G. K., Bennewitz, M., 2025. EvidMTL: Evidential Multi-Task Learning for Uncertainty-Aware Semantic Surface Mapping from Monocular RGB Images. *arXiv preprint arXiv:2503.04441*.

Misra, D., 2019. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*.

Mosquera Rojas, G. E., van der Voort, S., Pirkl, C. M., Kaushik, S., Smits, M., Klein, S., 2025. Evaluation of monte carlo dropout for uncertainty quantification in multi-task deep learning-based glioma subtyping. *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, Springer, 180–190.

Nandy, J., Hsu, W., Lee, M. L., 2020. Towards maximizing the representation gap between in-domain & out-of-distribution examples. *Advances in neural information processing systems*, 33, 9239–9250.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation. *European conference on computer vision*, Springer, 483–499.

Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S., 2020. On the uncertainty of self-supervised monocular depth estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3227–3237.

Postels, J., Blum, H., Strümpler, Y., Cadena, C., Siegwart, R., Van Gool, L., Tombari, F., 2020. The hidden uncertainty in a neural networks activations. *arXiv preprint arXiv:2012.03082*.

Sensoy, M., Kaplan, L., Cerutti, F., Saleki, M., 2020. Uncertainty-aware deep classifiers using generative models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 5620–5627.

Sensoy, M., Kaplan, L., Kandemir, M., 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Shafer, G., 1976. *A mathematical theory of evidence*. 42, Princeton university press.

Song, S., Truong, K. G., Kim, D., Jo, S., 2023. Prior depth-based multi-view stereo network for online 3D model reconstruction. *Pattern Recognition*, 136, 109198.

Su, W., Xu, Q., Tao, W., 2022. Uncertainty guided multi-view stereo network for depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11), 7796–7808.

Tosi, F., Bartolomei, L., Poggi, M., 2025. A survey on deep stereo matching in the twenties. *International Journal of Computer Vision*, 133(7), 4245–4276.

Ulmer, D. T., 2021. A survey on evidential deep learning for single-pass uncertainty estimation.

Wang, C., Wang, X., Zhang, J., Zhang, L., Bai, X., Ning, X., Zhou, J., Hancock, E., 2022. Uncertainty estimation for stereo matching based on evidential deep learning. *Pattern Recognition*, 124, 108498.

Wang, F., Zhu, Q., Chang, D., Gao, Q., Han, J., Zhang, T., Hartley, R., Pollefeys, M., 2024. Learning-based multi-view stereo: A survey. *arXiv preprint arXiv:2408.15235*.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338, 34–45.

Wei, Z., Zhu, Q., Min, C., Chen, Y., Wang, G., 2021. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. *Proceedings of the IEEE/CVF international conference on computer vision*, 6187–6196.

Xu, H., Zhou, Z., Wang, Y., Kang, W., Sun, B., Li, H., Qiao, Y., 2021. Digging into uncertainty in self-supervised multi-view stereo. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6078–6087.

Yang, X., Gao, Y., Luo, H., Liao, C., Cheng, K.-T., 2019. Bayesian denet: Monocular depth prediction and frame-wise fusion with synchronized uncertainty. *IEEE Transactions on Multimedia*, 21(11), 2701–2713.

Ye, K., Chen, T., Wei, H., Zhan, L., 2024. Uncertainty regularized evidential regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38, 16460–16468.

Zhang, J., 2025. Survey on Monocular Metric Depth Estimation. *arXiv preprint arXiv:2501.11841*.

Zhao, N., Wang, H., Cui, Q., Wu, L., 2024. U-ETMVSNet: Uncertainty-Epipolar Transformer Multi-View Stereo Network for Object Stereo Reconstruction. *Applied Sciences*, 14(6), 2223.

Zhao, X., Chen, F., Hu, S., Cho, J.-H., 2020. Uncertainty aware semi-supervised learning on graph data. *Advances in Neural Information Processing Systems*, 33, 12827–12836.