Holistic and Contextual Evidential Stereo-LiDAR Fusion for Depth Estimation

Jiayuan Fan, Haixiang Chen, Weide Liu, Xun Xu, Jun Cheng, *Senior Member, IEEE*

***Abstract*—Stereo-LiDAR fusion is often used for autonomous systems such as self-driving cars as the two modalities are complementary to each other. Existing stereo-LiDAR fusion methods are mostly at feature level or outcome level, without considering the uncertainty of the depth estimation in each modality. To this end, we propose a holistic and contextual evidential stereo-LiDAR fusion network (HCENet) for depth estimation, which considers both intra-modality and inter-modality uncertainties from stereo matching and LiDAR point cloud depth completion. We design a dual network structure that consists of a stereo matching branch and a LiDAR depth completion branch with new introduced uncertainty estimation modules for both two branches. Specifically, a multi-scale depth guided feature aggregation module is first developed to enable information propagation at early input stage, and then followed by fusing the predicted depths from two branches based on evidential uncertainties to generate the final output. Extensive experimental results on KITTI depth completion and Virtual KITTI2 datasets achieve RMSE of 599.3 and 2253.1, and show that our method outperforms state-of-the-art SLFNet by 6.52% and 20.7%, respectively.**

***Index Terms*—Stereo-LiDAR fusion, uncertainty estimation, depth completion, stereo matching.**

I. Introduction

Nowadays perceiving 3D information of the scenery becomes an essential task in many applications, such as autonomous vehicles [1], obstacle detection and avoidance [2], [3], mobile robotics [4], etc. Ensuring the high-precision and reliability of depth maps becomes critical especially in safety-critical applications. With the rapid development of active and passive sensors, the LiDAR and stereo depth sensors are often used to measure the depth information. As LiDAR often acquires sparse point cloud, depth completion is often conducted to compute dense depth maps [5]–[8]. However, their performance are still limited to the sparsity of LiDAR point clouds and properties of captured objects. Alternatively, the stereo-based depth estimation methods [9]–[12] that estimate the depth maps from a pair of left and right RGB images are studied in recent years. Different from LiDAR that often captures spare but accurate depth, stereo matching methods compute dense but less accurate depth [13]. The performance of stereo matching methods are mainly restricted by the error of pixel matching between left and right images, especially for texture-free and repetitive areas. Studies have shown that the LiDAR and stereo matching are complementary to each other. In order to explore the LiDAR and stereo images jointly, many of existing fusion methods integrate the two modalities at the feature level [14], [15]. These stereo-LiDAR fusion methods achieve some progress on the complementary information exploration. However, the uncertainty of the data has not been fully considered, leading to untrustworthy depth estimation.

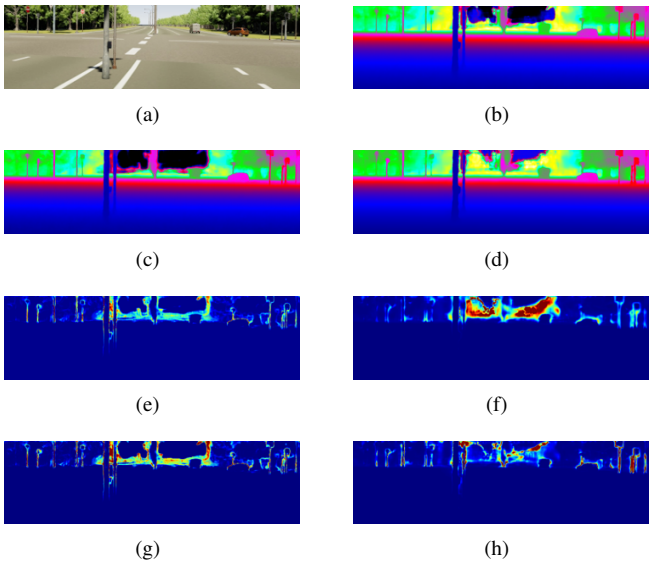


Fig. 1. The input of the model consists of the left image, left sparse depth, right image, and right sparse depth. (a) The left image, (b) final fusion depth, (c) depth map from depth completion, (d) depth map from stereo matching, (e) data uncertainty in depth completion, (f) data uncertainty in stereo matching, (g) model uncertainty in depth completion, (h) model uncertainty in stereo matching. Our proposed method estimates the data uncertainty and model uncertainty for stereo matching and LiDAR depth completion. The predicted uncertainty maps have large values at the boundaries of the object. We obtain the fusion depth by combining the results from LiDAR depth completion and stereo matching based evidential fusion.

Both the RGB and the point cloud data may suffer from some distortions or corruptions during data acquisition, such as saturation, sensor interference, adversarial attacks, etc. Such distortion or corruption often leads to uncertainty and the untrustworthy depth estimation results. There are two main types of uncertainties including the data uncertainty and model uncertainty. The data uncertainty is also called aleatoric uncertainty and it reflects the degree to which the data is noisy, corrupted and inaccurate, compared with underlying accurate

Manuscript received xxxx xx, 2024; revised xxxx, 2024. This research is supported in part by National Natural Science Foundation of China Grant No. 62101137, Shanghai Natural Science Foundation Grant No. 23ZR1402900 and the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funds Grant No. M23L7b0021. *(Corresponding author: Jun Cheng)*.

J. Fan and H. Chen are with Academy for Engineering and Technology, Fudan University, Shanghai, China (email: jyfan@fudan.edu.cn, 21210860038@m.fudan.edu.cn).

W. Liu, X. Xun and J. Cheng are with Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research (A*STAR), Singapore 138632 (email: {liu_weide, xu_xun, cheng_jun}@i2r.a-star.edu.sg).

Digital Object Identifier xx.xxxx/xxxx.xxxx.xxxxxxx

data. The model uncertainty is also named the epistemic uncertainty and it refers to the degree of predicted results of trained models compared to real results, describing the confidence of the model output [16], [17]. As shown in Fig. 1, following evidential regression [18], we estimate data and model uncertainties for both depth completion and stereo matching branches as well as the depth maps for depth fusion. The model uncertainty in depth completion represents the uncertainty generated by the depth completion branch, while model uncertainty in stereo matching represents the uncertainty from the stereo matching branch. The data uncertainty in depth completion or stereo matching reflects the noise present in corresponding data, respectively. This uncertainty can arise due to various factors such as sensor noise, or due to inherent characteristics of objects, like reflective objects. Quantifying the data and model uncertainty is practical for enhancing the reliability of depth estimation.

With these motivations in mind, we propose a holistic and contextual evidential stereo-LiDAR fusion network (HCENet) to kill two birds with one stone (see Fig. 2). It employs a dual network structure including a stereo matching branch and a LiDAR depth completion branch. The framework enables uncertainty estimations in the two branches for fusion and elegantly integrate multi-modal information for evidential based depth fusion. To enable information propagation at early stage, we also propose a multimodal feature aggregation module at the early input stage.

We propose an evidential based fusion to combine the depth completion with stereo matching. The approach is able to provide more accurate depth estimation and the uncertainties at pixel level simultaneously. There are two benefits. Firstly, the improved depth estimation could provide more accurate distance estimation for surrounding environment, e.g., pedestrians, motorcycles, vehicles. Secondly, the uncertainty is another important outcome. As shown from the results, the approach is able to compute model uncertainty which is correlated with errors in depth estimation. Therefore, it can be used to identify predictions that might have large errors. This can be used for high-level decision-making or an evidence to prompt for human intervention.

Our main contributions are summarized as follows:

- We propose a novel holistic and contextual evidential based stereo-LiDAR fusion for depth estimation, which considers the uncertainties and enables both intra-evidential and inter-evidential fusion.
- We propose a feature aggregation module to propagate the information from LiDAR point cloud to camera image for cost volume construction. It exploits complimentary information from two modalities to improve depth estimation.
- Extensive experiments demonstrate that our proposed method outperforms the state-of-the-art stereo-LiDAR fusion methods on KITTI depth completion and Virtual KITTI2 datasets.

II. Related work

A. Stereo Matching

With the advent and maturation of deep learning, contemporary approaches to stereo matching have predominantly embraced a learning-based foundation [19]–[23]. These methods aggregate the features of the left and right images to construct the cost volume and then calculate the disparity value through 3D convolution [24]–[27]. The disparity can then be converted into depth value through the focal length and baseline of the camera. Some works construct cost volume in different ways. Kendall et al. [28] and Chang et al. [9] construct cost volume using concatenation. Guo et al. [29] simultaneously construct a group-wise correlation volume and a concatenation volume. Xu et al. [30] introduce a sparse points-based intra-scale and cross-scale cost aggregation method to replace 3D convolution, which significantly improves the speed. Most of these methods need to set a maximum disparity according to the datasets. Li et al. [31] exploit cross-attention for matching pixels and does not require a pre-defined disparity range. But it requires additional occlusion masks for training, which is not available for most dataset. Shen et al. [32] introduce a volume fusion module to directly combine multi-scale 4D volumes and calculate a multi-level loss to accelerate the convergence of the model. In this paper, we adopt pyramid combination and warping network (PCWNet) [32] as a baseline to combine with depth from LiDAR.

B. Depth Completion

LiDAR can provide accurate but sparse depth. To generate dense depth map from a sparse depth map, depth completion is often conducted. Many methods [33]–[37] have been proposed for LiDAR depth completion with the guidance of a reference color image to integrate structural information of objects. Typically, early fusion methods concatenate features from RGB images and sparse depth for depth completion. Yang et al. [38] use a late fusion strategy which infers the posterior distribution of a dense depth map associated with an image. Forkel et al. [39] introduce a real-time fusion method that integrates LiDAR point clouds into stereo processing using the semi-global matching algorithm. Hu et al. [40] apply a dual-branch backbone to generate color- and depth-dominant depth maps, then fuse them to get the final output. Zhao et al. [41] introduce a mechanism of neighboring attention. Rho et al. [42] introduce transformer architecture for depth completion and a guided-attention block to fuse depth and color features. Zhang et al. [43] integrate convolutional layers and transformers into a single block for deep completion, enabling the encoding of both local and global information simultaneously. To avoid large amount of parameters in transformer based approaches, we choose to use PENet [40] in the depth completion branch. From the models, we further introduce uncertainty prediction on top of PENet. Our method, as compared to existing depth completion approaches, incorporates uncertainty estimation for the completed depth map. The resulting depth map, along with its associated uncertainty, can be fused with the depth map from the stereo modality.

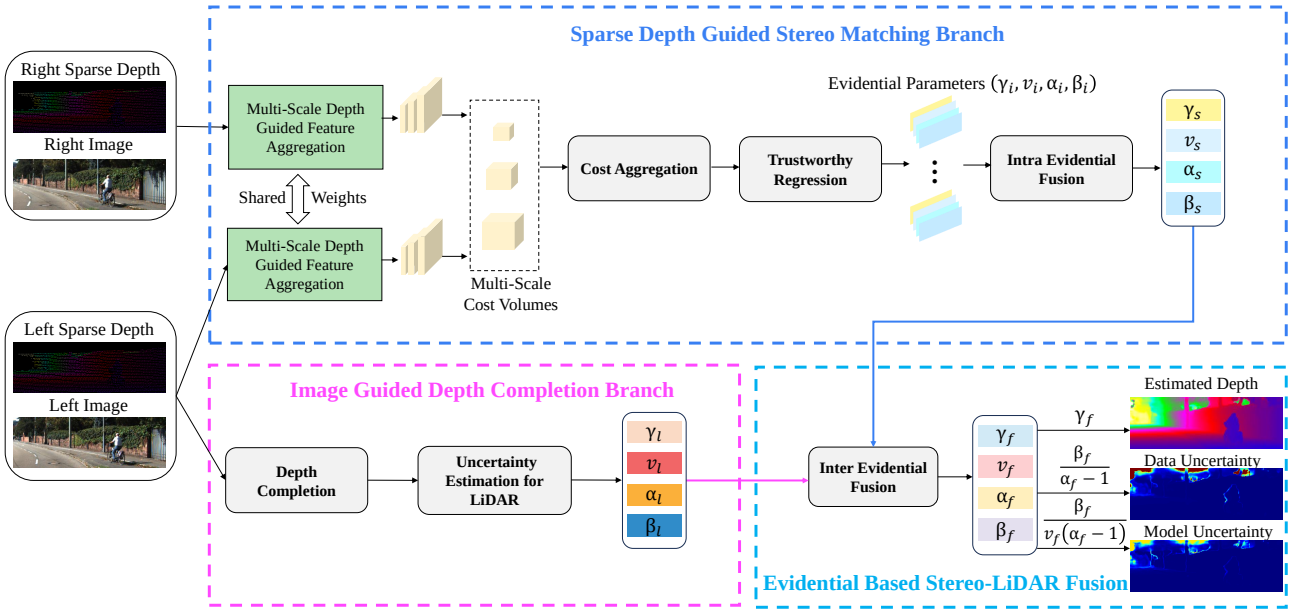


Fig. 2. The overview of our proposed method. It mainly consists of sparse depth guided stereo matching branch and image guided depth completion branch, followed by the evidential based stereo-LiDAR fusion for final dense depth generation. In the sparse depth guided stereo matching branch, two pairs of LiDAR and stereo images are fed into multi-scale depth guided feature aggregation for estimated dense depth and stereo uncertainty estimation. In the image guided depth completion branch, a pair of LiDAR and stereo images are used for estimated dense depth and LiDAR uncertainty estimation.

C. Uncertainty Estimation

The data uncertainty can be caused by the noise of training data, the model uncertainty represents the shortcomings of the network. Ensemble [44] methods predict model uncertainty by training multiple models with different initializations and modeling the distribution over parameters. Monte Carlo (MC) dropout [45] obtains multiple different prediction distributions by randomly turning off neurons in the neural network. Both ensemble and MC dropout require multiple forward passes, depth values and uncertainties can be obtained by computing the mean and variance of multiple predicted values, respectively.

Recently, trustworthy machine learning has been used to compute uncertainty for evidential based multi-view classification [46], which shows significant improvement. Many works use uncertainty for depth prediction. Gansbeke et al. [47] integrate global and local information to generate the depth map via confidence weights. Cheng et al. [11] calculate adaptive thin volume with an uncertainty-aware cascaded design. Shen et al. [48] adaptively adjust the disparity search range. In contrast to the previous methods that rely on variance-based uncertainty, deep evidential learning [18] is employed to estimate the uncertainties of the disparity map [49]. Lou et al. [50] utilize evidential based approach to fuse the outputs from different stereo matching methods.

Based on evidential learning [18], every depth estimation d is drawn from a normal distribution with unknown mean and variance (μ, σ^2), which are assumed to be drawn from normal and inverse-gamma distributions, respectively:

$$d \sim \mathcal{N}(\mu, \sigma^2), \ \mu \sim \mathcal{N}(\gamma, \sigma^2 v^{-1}), \ \sigma^2 \sim \Gamma^{-1}(\alpha, \beta), \quad (1)$$

where $\Gamma(\cdot)$ denotes the gamma function, $\gamma \in \mathbb{R}$, $v > 0$, $\alpha > 1$, $\beta > 0$.

With the assumption that the mean and the variance are independent, the posterior distribution is formulated as a normal-inverse gamma distribution NIG$(\gamma, v, \alpha, \beta)$. In this work, we follow the evidential learning approach and predict uncertainties for both stereo matching branch and depth completion branch for fusion. Different from [49], we integrate the uncertainty computation with multi-scale cost volume for an intra-modality fusion. Moreover, we also incorporate inter-modality fusion between stereo matching and LiDAR based depth completion.

D. Stereo-LiDAR Fusion

Multimodal fusion combines different kinds of data, such as images with images, images with text [51], [52], and images with events [53], [54]. For example, Bao et al. [51] employ a cross-modal transformer structure designed for deeply integrating image and text features. Additionally, Liu et al. [53] utilize an event and image fusion transform module for merging image and event features, which is based on a mutual attention mechanism. There have been some research endeavors focusing on the fusion of stereo and LiDAR modalities, aiming to exploit their complementary properties. In the method introduced by Zhang et al. [55], a straightforward concatenation of features derived from stereo images and point clouds is utilized for the purpose of depth estimation. Maddern et al. [14] use a probabilistic approach of real-time LiDAR stereo fusion for low-cost 3D perception. However, this probabilistic fusion approach performs poorly in scenarios with limited depth information. To address this issue, Park et al. [15] employ a CNN model to fuse the input LiDAR disparity and Stereo disparity. Afterwards, LiDAR and stereo disparities are utilized to generate calibration parameters [56],

which are updated every frame for the purpose of registering LiDAR and stereo disparities. Wang et al. [57] fuse LiDAR information into a stereo matching network at an early stage. Choe et al. [58] extract image and point cloud features and fuse two modalities in a depth cost volume to encode semantic and geometric information. Cheng et al. [59] tackle the noise and misalignment in the data by a feedback loop. Xu et al. [60] introduce a novel lightweight scheme that combines RGB information and sparse LiDAR points to generate a semi-dense depth map and employ a varying-weight Gaussian guiding method for effective cost volume aggregation. He et al. [61] introduce a novel framework that integrates semantic information from stereo images and spatial information from raw point clouds for enhanced 3D object detection by utilizing a residual attention learning mechanism. Different from stereo cameras, gated cameras emit light pulses and capture photons from a specific distance, filtering out backscatter from weather conditions like fog, rain, and snow [62]. In order to overcome the limitations of existing LiDAR and RGB stereo depth estimation methods, Walz et al. [63] present a method that leverages gated stereo observations and exploits both multi-view and time-of-flight cues to estimate high-resolution depth maps. Our proposed method stands out from existing stereo-LiDAR fusion techniques by not only integrating different modalities at the feature level but also leveraging uncertainty to fuse multiple depth maps.

III. METHODOLOGY

The overview of our proposed method is shown in Fig. 2. It consists of two main branches. Specifically, one branch is designed for sparse depth guided stereo matching and the other one is designed for image guided depth completion, where each branch predicts the depth map as well as modality-specific uncertainties. Finally, the outputs from the two branches are combined via evidential based fusion.

In the sparse depth guided stereo matching branch, we first apply a multi-scale depth guided feature aggregation on the stereo images and their corresponding LiDAR depth maps. Subsequently, a stereo matching network is employed, whereby the adoption of the architecture proposed by the recent PCWNet [32] is predicated upon its commendable efficacy in facilitating cross-domain generalization as well as enhancing stereo matching accuracy. By performing multi-scale cost volume and disparity regression, the estimated dense depth can be obtained by converting the disparity into depth. We integrate the uncertainty computation into the architecture to compute the uncertainty of the stereo matching results. In the depth completion branch, the left image and corresponding left sparse depth are used for deep completion network to generate the estimated dense depth. We adopt the network architecture from [40]. Similar to that in stereo matching, uncertainty of depth completion is also computed. After the depth map and uncertainty map of the two branches are obtained, we can take the pixel-wise uncertainty into account and fuse the evidential distribution of pixels of two different modalities for final depth. In this way, we can also get the model uncertainty and data uncertainty of the fused depth map.

A. Sparse Depth Guided Stereo Matching

In stereo matching, the process typically begins with feature extraction and cost volume construction. The subsequent steps in stereo matching, namely 3D cost aggregation and disparity regression, utilize the information in the cost volume to produce the final disparity map. Cost volume encodes important features for subsequent disparity calculations. Different from only using stereo images to construct cost volumes [32], we here enrich the cost volumes by combining the information from both sparse depth and stereo images using a new proposed multi-scale depth guided feature aggregation module.

1) Multi-scale Depth Guided Feature Aggregation: Inspired by multi-modality features fusion method [64], we employ a multi-scale depth guided feature aggregation strategy to incorporate depth information into RGB image, as shown in Fig. 3. Given an image I and corresponding sparse depth D, initially, we employ separate convolutional layers to generate initial features for both I and D, which serve as inputs to the initial multimodal feature aggregation (MFA) module. Subsequently, the output of each MFA layer is propagated as input to the subsequent MFA layer. A five-stage MFA extraction module is proposed to compute multi-scale feature maps, which is used for cost volume generation later.

As shown in Fig. 3, in MFA block, we use a shared convolutional layer to process RGB features f_I and sparse depth features f_D, yielding M_I and M_D. Then we concatenate M_I and M_D, thereby blending their distinctive attributes at a specific spatial location. Afterward, we use different convolutional layers to generate spatial-wise gate for M_I and M_D, respectively. After obtaining these two gates $G_I \in \mathbb{R}^{1\times H\times W}$, and $G_D \in \mathbb{R}^{1\times H\times W}$, a softmax function is used to generate weights for the fusion between RGB features and sparse depth features as follows:

$$A_I = \frac{e^{G_I}}{e^{G_I} + e^{G_D}}, \tag{2}$$

$$A_D = \frac{e^{G_D}}{e^{G_I} + e^{G_D}}. \tag{3}$$

An intermediate feature f_w is then computed by weighting the RGB and depth features as follows:

$$f_w = M_I \cdot A_I + M_D \cdot A_D. \tag{4}$$

Then we compute f'_I and f'_D as mean of f_w and RGB features or depth features, respectively, which are used for next stage of MFA:

$$f'_I = \frac{f_w + M_I}{2}, \tag{5}$$

$$f'_D = \frac{f_w + M_D}{2}. \tag{6}$$

We concatenate f_w and M_I to get f_A:

$$f_A = [\, f_w,\ M_I \,]. \tag{7}$$

After the multi-scale depth-guided feature aggregation, three feature maps with sizes of $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, and $\frac{H}{16} \times \frac{W}{16}$ are obtained for constructing the multi-scale cost volume. H and W are the height and width of the initial image.

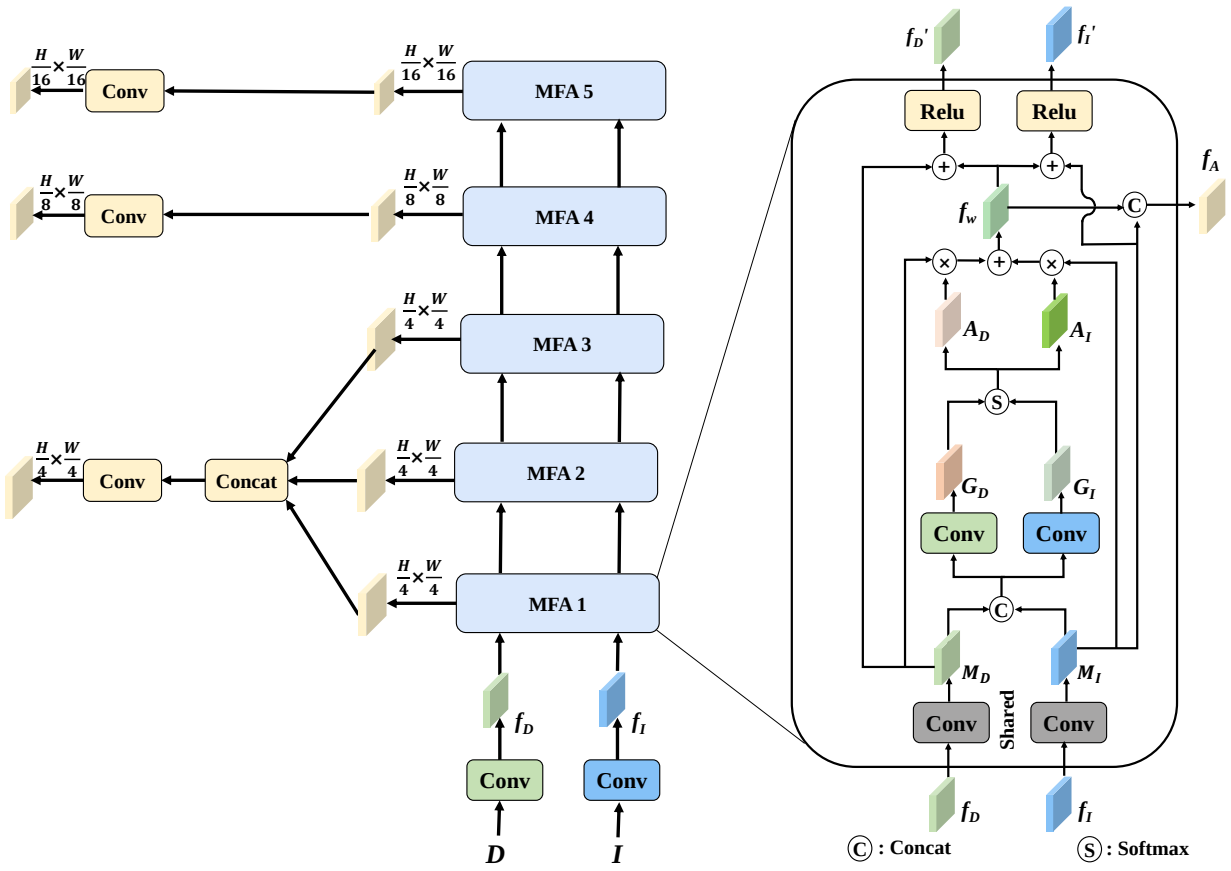


Fig. 3. Schematic diagram of the five-stage multi-scale depth guided feature aggregation and diagram of the multimodal feature aggregation (MFA) module. In the first MFA module, the inputs are RGB features f_I and sparse depth features f_D. The outputs are f_I' and f_D', which serve as the inputs for the next stage of the MFA.

2) Uncertainty Estimation in Stereo Matching: We adopt the recent PCWNet [32] because of its good performance on both cross-domain generalization and stereo matching accuracy. PCWNet constructs combination volumes on the upper levels of the pyramid and develops a cost volume fusion module to integrate them for initial disparity estimation. Then, a warping volume is constructed on the last level of the pyramid and employed to refine the initial disparity. Slightly different from the original PCWNet, we use three scales instead of four to reduce computational cost for later fusion with LiDAR, and we do not employ depth refinement. By performing multi-scale cost volume and disparity regression, the estimated dense depth is obtained. In this work, we compute the uncertainties for the multi-scale cost volume to achieve intra-scale fusion.

To estimate the parameters of NIG distribution rather than merely predicting the disparity map, we modify the disparity regression module into a trustworthy regression with multi-channel output while keeping the remaining modules unchanged. As shown in Fig. 4, the proposed trustworthy regression module uses a single branch of 3D convolution and the up-sampling block to compute a 4-channel output $O \in \mathbb{R}^{D\times H\times W\times 4}$, where D is the maximum of disparity, H and W are the height and width of the stereo images. We set $D = 192$ in our implementation. The distribution parameters can be computed as follows:

$$O_\gamma, O_v, O_\alpha, O_\beta = \text{Split}(O, dim = -1), \tag{8}$$

$$p = \text{Softmax}(O_\gamma), \tag{9}$$

$$\gamma = \sum_{k=0}^{D} k \cdot p_k, \;\; \ell_i = \sum_{k=0}^{D} O_i \cdot p_k, \tag{10}$$

where k denotes disparity level, p denotes the probability, and $i \in \{v, \alpha, \beta\}$. Finally, a Softplus activation is applied on the ℓ_i to obtain v, α, β.

After obtaining the evidence distribution parameters $(\gamma, v, \alpha, \beta)$ of each pixel, following deep evidential regression [18], predicted depth value, data and model uncertainties can be calculated as follow:

$$\underbrace{\mathbb{E}[\mu] = \gamma}_{\text{prediction}}, \tag{11}$$

$$\underbrace{\mathbb{E}\left[\sigma^2\right] = \frac{\beta}{\alpha - 1}}_{\text{data uncertainty}}, \tag{12}$$

$$\underbrace{\text{Var}[\mu] = \frac{\beta}{v(\alpha - 1)}}_{\text{model uncertainty}}. \tag{13}$$

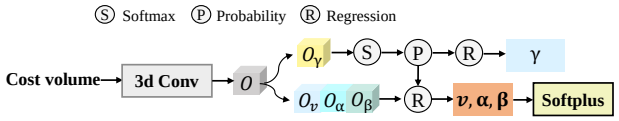


Fig. 4. The trustworthy regression module designed for stereo matching.

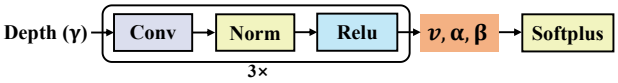


Fig. 5. The network structure for uncertainty estimation in depth completion.

Parameter γ is depth value of prediction, parameters (v, α, β) are used to calculate data and model uncertainties. In Eq. (12) and Eq. (13), parameters β and α affect both data uncertainty and model uncertainty. In Eq. (13), parameter v affects model uncertainty. With the decrease of parameter v, the model uncertainty increases.

B. Uncertainty Aware Image Guided Depth Completion

We follow the main network structure in the precise and efficient image guided depth completion network (PENet) [40] and modify it for uncertain aware image guided depth completion. PENet uses a two-branch backbone to exploit color-dominant and LiDAR-dominant information for two indepent depth maps. The two dense depth maps are then fused using the same strategy as in FusionNet [47]. We utilize the PENet while excluding the utilization of its depth refinement module.

Fig. 5 illustrates network structure of the uncertainty computation. On top of the structure from PENet, we denote its estimated depth value as γ. We further use three convolutional layers with the kernel size 3×3 followed by BatchNorm and Relu to generate per-pixel uncertainty parameters v, α and β, which enable us to compute the data and model uncertainties in depth completion branch.

C. Fusion Strategy based on Evidence

Simply taking a weighted average of the depths from different scales or modalities may result in mutual interference among the different depth maps. Therefore, it is essential to incorporate the uncertainty of each depth map for fusion. We adopt the fusion strategy with the mixture of NIG distribution (MoNIG) [65] for its excellent mathematical properties to perform both intra evidential fusion and inter evidential fusion. Given k sets of parameters of NIG distributions, the MoNIG distribution can be computed with the following operations:

$$\begin{aligned} MoNIG(\gamma, v, \alpha, \beta) = NIG(\gamma_1, v_1, \alpha_1, \beta_1) \oplus \\ NIG(\gamma_2, v_2, \alpha_2, \beta_2) \oplus \cdots \oplus NIG(\gamma_k, v_k, \alpha_k, \beta_k), \end{aligned} \tag{14}$$

where $\oplus$ represents the summation operation of two NIG distributions:

$$\begin{aligned} NIG(\gamma, v, \alpha, \beta) \triangleq NIG(\gamma_m, v_m, \alpha_m, \beta_m) \\ \oplus NIG(\gamma_n, v_n, \alpha_n, \beta_n), \end{aligned} \tag{15}$$

where $\gamma = \frac{v_m\gamma_m + v_n\gamma_n}{v_m + v_n}$, $\alpha = \alpha_m + \alpha_n + \frac{1}{2}$, $v = v_m + v_n$, $\beta = \beta_m + \beta_n + \frac{1}{2}v_m(\gamma_m - \gamma)^2 + \frac{1}{2}v_n(\gamma_n - \gamma)^2$.

1) Intra Evidential Fusion of Multi-scale Cost Volume Stereo Matching: Multi-scale cost volume is often used in stereo matching to exploit the features from different layers of the feature extractor. We construct three levels of the cost volumes with $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$ scaled features and employ the cost volume fusion module [32] to early combine the knowledge from multi-scale receptive field. These feature maps contain coarse-to-fine semantic information such as textures, boundaries, and regions. Then we apply three branches of trustworthy regression modules to generate parameters of NIG distributions $(\gamma_i, v_i, \alpha_i, \beta_i)$, where $i \in \{1, 2, 3\}$. Intra evidential fusion module integrates three NIG distributions into one NIG distribution $NIG(\gamma_s, v_s, \alpha_s, \beta_s)$:

$$\begin{aligned} NIG(\gamma_s, v_s, \alpha_s, \beta_s) = \\ NIG(\gamma_1, v_1, \alpha_1, \beta_1) \oplus \cdots \oplus NIG(\gamma_3, v_3, \alpha_3, \beta_3). \end{aligned} \tag{16}$$

The intra evidential fusion strategy integrates multi-scale features for trustworthy stereo matching.

2) Inter Evidential Fusion of LiDAR Depth Completion and Stereo Matching: Besides the intra evidential fusion of stereo matching, we further apply inter evidential fusion from LiDAR and stereo matching. We denote the NIG distribution of depth from LiDAR as $NIG(\gamma_l, v_l, \alpha_l, \beta_l)$. We fuse it with the depth from stereo matching as follows to get:

$$\begin{aligned} NIG(\gamma_f, v_f, \alpha_f, \beta_f) = \\ NIG(\gamma_s, v_s, \alpha_s, \beta_s) \oplus NIG(\gamma_l, v_l, \alpha_l, \beta_l), \end{aligned} \tag{17}$$

where $\gamma_f = (v_s + v_l)^{-1}(v_s\gamma_s + v_l\gamma_l)$, $\alpha_f = \alpha_s + \alpha_l + \frac{1}{2}$, $v_f = v_s + v_l$, $\beta_f = \beta_s + \beta_l + \frac{1}{2}v_s(\gamma_s - \gamma_f)^2 + \frac{1}{2}v_l(\gamma_l - \gamma_f)^2$.

The parameter v_s and v_l represent the confidence of the depth predicted by the stereo and LiDAR modalities, respectively. In this way, pixels with high uncertainty have lower weights in the fusion process, and the depth map after fusion is trustworthy and interpretable. The parameters after fusion $(\gamma_f, v_f, \alpha_f, \beta_f)$ also obey the NIG distribution. Therefore data uncertainty and model uncertainty can be calculated directly. The uncertainty map of the fused depth map is obtained by mixing the NIG distribution of the corresponding pixels of the two branches. It does not need parameters for learning and does not require a convolutional layer.

D. Loss Function

1) Depth Estimation Loss: We compute the RMSE losses $\mathcal{L}_D$ for depth obtained through stereo matching, depth completion, and the final fused depth, denoted as $\mathcal{L}_s$, $\mathcal{L}_l$ and $\mathcal{L}_f$ respectively. The depth map of stereo branch is obtained by converting the disparity maps using the camera focal length and baseline. The RMSE loss $\mathcal{L}_D$ is calculated based on the ground truth depth map:

$$\mathcal{L}_D = \sqrt{\frac{1}{|V|}\sum_{v \in V}\left(\hat{D}_v - D_v^{gt}\right)^2}, \tag{18}$$

where V represents the set of pixels with valid depth in the ground truth depth map, and $|V|$ represent the size of the set V. $\hat{D}_v$ refers to the predicted depth, while D_v^{gt} represents the ground truth depth.

For stereo matching branch, we also compute smooth L_1 loss on three disparity maps, which can be expressed as follows:

$$\mathcal{L}_{disp} = \sum_{j=0}^{2} w_j \cdot \mathcal{L}_{\text{smooth}L_1}\left(\hat{d}_j, d_{gt}\right), \tag{19}$$

where w_j is the weight of the j-th prediction of disparity map, we set w_0 = 0.7, w_1 = 0.7, w_2 = 1, empirically. $\hat{d}_j$ and d_{gt} are the j-th prediction and ground truth disparity, respectively. We only consider pixels with a disparity value greater than zero.

2) Uncertainty Estimation Loss: Following [18], we compute a negative logarithm of model evidence $\mathcal{L}^N(\gamma, v, \alpha, \beta)$ and an evidence regularizer $\mathcal{L}^{\mathrm{R}}(\gamma, v, \alpha)$. The model evidence $\mathcal{L}^N(\gamma, v, \alpha, \beta)$ is computed as:

$$\begin{aligned}\mathcal{L}^N(\gamma, v, \alpha, \beta) = &\frac{1}{2}\log\left(\frac{\pi}{v}\right) - \alpha\log(\Omega) \\ &+ \left(\alpha + \frac{1}{2}\right)\log\left((D^{gt} - \gamma)^2 v + \Omega\right) \\ &+ \log\left(\frac{\Gamma(\alpha)}{\Gamma\left(\alpha + \frac{1}{2}\right)}\right),\end{aligned} \tag{20}$$

where $\Omega = 2\beta(1 + v)$.

The evidence regularizer $\mathcal{L}^{\mathrm{R}}(\gamma, v, \alpha)$ is computed as:

$$\mathcal{L}^{\mathrm{R}}(\gamma, v, \alpha) = |D^{gt} - \gamma| \cdot (2v + \alpha). \tag{21}$$

The evidence regularizer encourages large uncertainty when the difference between the predicted value γ and the ground truth D^{gt} is large. This can be easily seen as the minimization of the term in Eq. (21) would lead to smaller values of v and α and therefore large uncertainties in Eq. (12) and Eq. (13) subsequently.

The total loss $\mathcal{L}^{\text{unc}}$ comprises two loss terms that serve the purpose of maximizing and regularizing evidence:

$$\mathcal{L}^{\text{unc}}(\gamma, v, \alpha, \beta) = \mathcal{L}^{\mathrm{N}}(\gamma, v, \alpha, \beta) + \lambda\mathcal{L}^{\mathrm{R}}(\gamma, v, \alpha), \tag{22}$$

where $\lambda > 0$ controls the balance between the two items. We set $\lambda = 0.01$ in our experiments.

We apply the equations above on the depth maps generated by stereo matching branch, the LiDAR depth completion branch, and the fused depth map to get their corresponding uncertainty losses $\mathcal{L}^{\text{unc}}(\gamma_s, v_s, \alpha_s, \beta_s)$, $\mathcal{L}^{\text{unc}}(\gamma_l, v_l, \alpha_l, \beta_l)$ and $\mathcal{L}^{\text{unc}}(\gamma_f, v_f, \alpha_f, \beta_f)$, respectively.

To optimize the disparity value, predicted depth value and the corresponding uncertainty value, the overall loss function is computed as follows:

$$\begin{aligned}\mathcal{L}_{Total} = &\mathcal{L}_l + \mathcal{L}_s + \mathcal{L}_f + \mathcal{L}_{disp} + \mathcal{L}^{\text{unc}}(\gamma_s, v_s, \alpha_s, \beta_s) \\ &+ \mathcal{L}^{\text{unc}}(\gamma_l, v_l, \alpha_l, \beta_l) + \mathcal{L}^{\text{unc}}(\gamma_f, v_f, \alpha_f, \beta_f).\end{aligned} \tag{23}$$

IV. Experiments

A. Datasets and Evaluation Metrics

We have implemented the proposed network and tested its performance with the KITTI depth completion dataset [5] and the Virtual KITTI2 dataset [66].

The KITTI depth completion dataset is a large-scale dataset capturing real-world driving scenarios in outdoor environments. It comprises 138 scenes, encompassing a training set of 42,949 images and a validation set of 3,426 images. Because there is no ground truth in the test set, we follow CCVN [57] and use the same 1000 pictures in the validation set for comparison with other methods. The initial resolution is 1242×352. Due to the absence of ground truth in the top portion of the initial images, a cropping approach is employed during the training process, resulting in images with dimensions of 1216×256. The depth maps provided by Velodyne HDL-64e are sparse, with ground truth depth generated by combining consecutive 11 frames to form a single depth map, resulting in approximately 30% of pixels containing depth information [57].

Virtual KITTI2 dataset is a synthetic one with ground truth depth for all pixels and is a more realistic version of the original Virtual KITTI 1.3.1 [67]. Following [68], we use "Scene01" and "Scene02" for training, "Scene06", "Scene18" and "Scene20" for evaluation. For each scene, we take the only scenario (15-deg-left) for training and evaluation. The training and test sets have 680 and 1446 images, respectively.

Following [57], we compute root mean square error (RMSE), mean absolute error (MAE), inverse root mean square error (IRMSE), and inverse mean absolute error (IMAE) to evaluate the performance of our method.

TABLE I
COMPARISON WITH OTHER METHODS ON THE KITTI DEPTH COMPLETION VALIDATION SET. S REFERS TO STEREO IMAGES, L REFERS TO LIDAR AND M REFERS TO MONOCULAR IMAGE. ↓ INDICATES THAT THE SMALLER VALUE THE BETTER PERFORMANCE.

Methods	Modality	RMSE (mm)↓	MAE (mm)↓	IRMSE (1/km)↓	IMAE (1/km)↓
ACMNet [41]	M + L	1037.3	236.4	2.38	0.91
NLSPN [69]	M + L	868.8	236.7	2.61	1.02
GuideNet [70]	M + L	857.8	234.0	2.36	0.99
PENet [40]	M + L	761.3	210.9	2.19	0.90
LiStereo [55]	S + L	832.1	283.9	2.19	1.10
CCVN [57]	S + L	749.3	252.5	**1.39**	0.80
VPN [58]	S + L	636.2	205.1	1.87	0.98
SLFNet [68]	S + L	641.1	197.0	1.77	0.87
Ours	S + L	**599.3**	**190.0**	1.43	**0.78**

B. Implementation Details

Our method is implemented with PyTorch. The training phase uses the Adam ($\beta_1 = 0.9, \beta_2 = 0.999$) optimizer. We use four NVIDIA GTX 3090 GPUs for experiments, and set the batch size to 4. The initial learning rate is set to 1e-3. After training for 10 epochs, it decreases to 5e-4. The training epoch is set to 15 for KITTI depth completion dataset. We use the weights pre-trained on the KITTI depth completion dataset to fine-tune on Vitual KITTI2 dataset for 14 epochs. The maximum depth is set to 100 and 80 on KITTI depth completion dataset and Vitual KITTI2 dataset, respectively.

C. Comparison with The State-of-the-art Methods

1) Performance Comparison in KITTI: We compare our method with four stereo-LiDAR fusion methods including LiStereo [55], CCVN [57], VPN [58], SLFNet [68] in KITTI depth completion dataset. We obtain the results from [68].

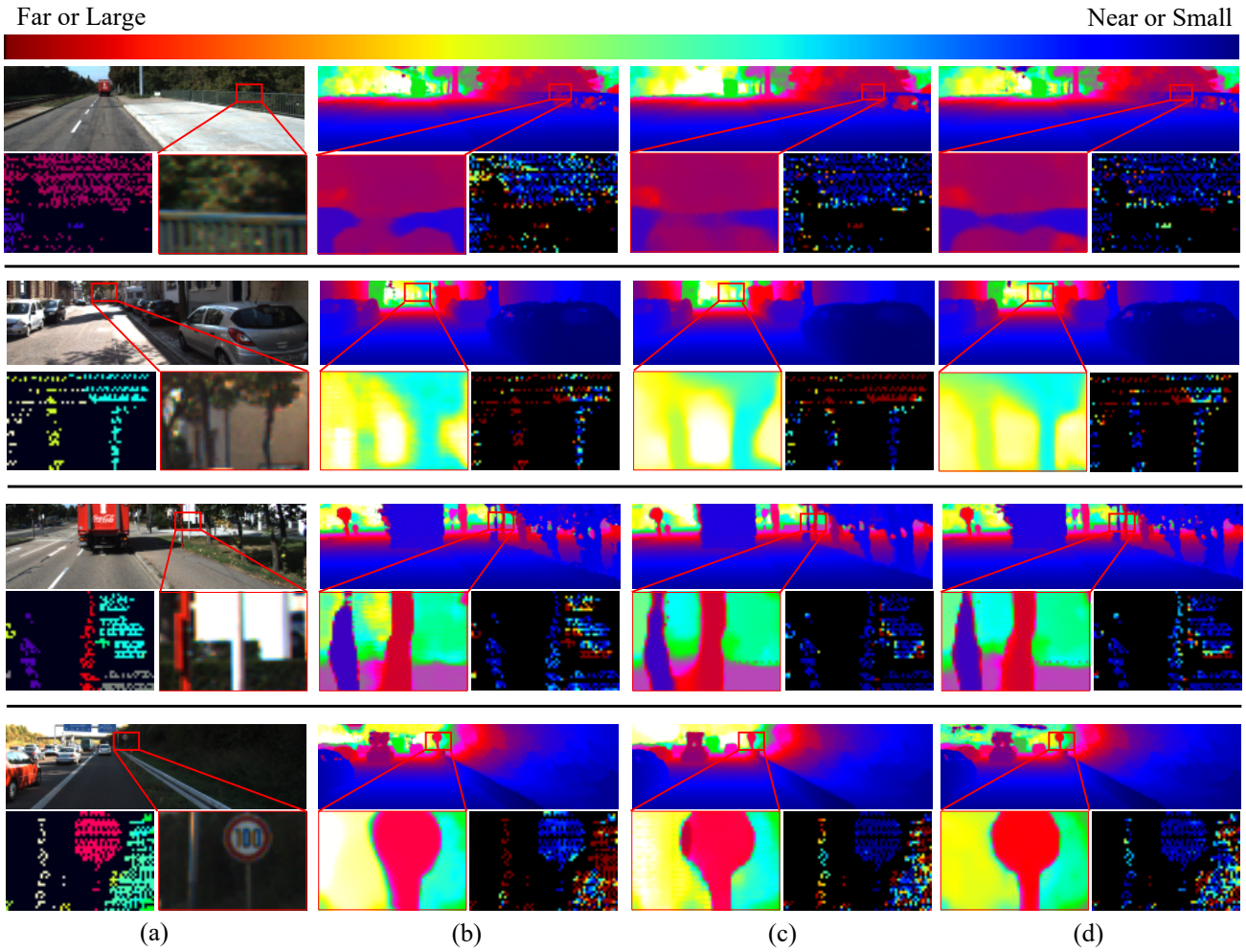


Fig. 6. Visual comparison of results. (a) The left image (top), enlarged area indicated by the rectangular in red (bottom right), the ground truth of depth values for enlarged area (bottom left). (b) The results by CCVN (top), enlarged area (bottom left), error map (bottom right). (c) The results by SLFNet (top), enlarged area (bottom left), error map (bottom right). (d) The results by our proposed HCENet (top), enlarged area (bottom left), error map (bottom right). The black color in error maps indicates an absence of ground truth. The colors from red to blue in the color bar indicate depth or error from large to small.

TABLE II
COMPARISON WITH STEREO-LIDAR FUSION METHODS ON THE VITUAL KITTI2 DATASET.

Methods	Modality	RMSE(mm)↓	MAE(mm)↓
CCVN [57]	Stereo + LiDAR	3726.8	915.6
VPN [58]	Stereo + LiDAR	3217.1	712.0
SLFNet [68]	Stereo + LiDAR	2843.1	696.2
Ours	Stereo + LiDAR	**2253.1**	**670.1**

Table I summaries the results. Our method is compared with other methods in four metrics. Specifically, it outperforms VPN and SLFNet by 5.8% and 6.5% in RMSE, 7.3% and 3.5% in MAE, 23.5% and 19.2% in IRMSE, 20.4% and 10.3% in IMAE, respectively.

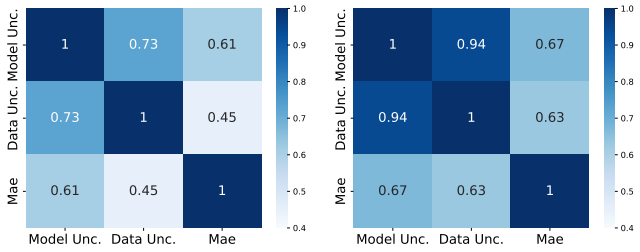
Fig. 6 shows a visual comparison of the results of four samples by the proposed method with those by CCVN and SLFNet. We zoom in on the area in the red box. The ground truth of depth maps provided by LiDAR are sparse, less than 30% of pixels in the ground truth contain depth information. Pixels without information are shown as black points in visualizations with a depth value of 0, making it hard to see the shapes of many objects. Differences in color intensity between the predicted depth map and the actual one are mostly due to errors in estimation. We have added a color bar for better visualization. In the depth map, blue stands for small distances, while red represents large distances. As for the error maps in (b), (c), and (d), blue indicates small error, whereas red signifies large error. As we can see from the comparison, our method predicts the depth more accurately in regions with objects such as railing, traffic sign board, etc. In contrast, CCVN and SLFNet produce inferior predictions where the shapes of the objects are not well preserved from the backgrounds in depth map.

Besides the stereo-LiDAR fusion methods, we also show the results from monocular depth completion fusion (M+L) methods (ACMNet [41], NLSPN [69], GuideNet [70], PENet [40]) in Table I for comparison. Compared to monocular depth completion methods, our approach also surpasses all the methods by large margins.

2) Performances in other dataset: Although KITTI depth completion dataset is a real-world dataset, it has some limita-

TABLE III
PERFORMANCE OF DIFFERENT FEATURE FUSION STRATEGIES ON KITTI DEPTH COMPLETION VALIDATION SET.

Methods	RMSE↓	MAE↓	IRMSE↓	IMAE↓
Concat	614.4	193.4	1.49	0.81
Ours	**599.3**	**190.0**	**1.43**	**0.78**



(a) (b)

Fig. 7. The Pearson correlation coefficient matrix among the data uncertainty, the model uncertainty and the MAE of estimated depth on (a) The KITTI depth completion dataset and (b) The Virtual KITTI2 dataset.

tions as the ground truth of the depth is derived by aggregating LiDAR points from surrounding frames. Therefore, it suffers from errors for regions such as the glass on cars, occluded areas, etc. Therefore, we also conduct the experiments on Virtual KITTI2 dataset. Since the Virtual KITTI2 dataset only provides stereo image pairs and dense depth pairs, we use randomly generated masks and integrate them with ground truth to obtain sparse depth maps, each with about 5k pixels. Table II shows the comparison on the Virtual KITTI2 dataset. Our method achieves the best performance as well. Following SLFNet [68], we compare our method with CCVN, VPN, and SLFNet. Our method outperforms CCVN, VPN, and SLFNet in RMSE by 39.5%, 30.0%, and 20.7%, respectively.

D. Ablation Study

In this section, we perform ablation studies to assess the impact of different design choices in our network. We first evaluate the benefits of the MFA module. We compare MFA with simple concatenation. Table III shows the performance of different feature fusion strategies on KITTI depth completion dataset. Our MFA module outperforms simple concatenation and proves that our fusion method is suitable for feature fusion of sparse depth and RGB image.

Next, we evaluate how different values of λ affect the performance, which is used to control the balance between model evidence and evidence regularizer in Eq. (22). Table IV shows the results for different λ. As we can see, $\lambda = 0.01$ leads to best results. The results show that the model performance will be worse without penalty ($\lambda = 0$), which justifies the needs of the evidence regularization. On the other side, large values would not help much.

We continue the ablation study to justify the effectiveness of intra and inter evidential fusion, and the MFA. Table V shows the results. As can be seen, if the MFA module is not used, the RMSE will increase, which shows that the introduction of sparse depth features in the feature fusion

TABLE IV
COMPARISON OF THE STRENGTH OF EVIDENCE REGULARIZATION ON KITTI DEPTH COMPLETION DATASET.

λ	RMSE↓	MAE↓	IRMSE↓	IMAE↓
0	602.6	208.9	1.43	0.82
0.5	644.0	198.3	1.49	0.82
0.02	600.6	192.6	1.44	0.79
0.01	**599.3**	**190.0**	**1.43**	**0.78**

stage can help our stereo matching branch to generate more accurate depths. This is because there are many textureless and occluded regions in the dataset, and the introduction of LiDAR information can reduce false predictions. Intra and inter evidential fusion both contribute to the improvement of the model, resulting in a respective decrease of 0.8% and 7.9% in the RMSE. Furthermore, when these three modules are used in conjunction, the model exhibits its best performance across all four metrics.

E. Comparison of Depth Map Fusion Strategies

We also conduct experiments to evaluate the performance enhancements introduced by our depth fusion strategies. In Table VI, we compare different ways to fuse the depth values predicted by the two branches of stereo matching and depth completion. Using the average depth of two branches lacks uncertainty incorporation. It simply combines the depth maps without considering uncertainty and divides the result by two, leading to poor performance. We also compute the relative model uncertainties obtained from two branches by concatenating their model uncertainties and applying the softmax function. This process allows us to obtain the relative uncertainties, which is then used for depth map fusion. Our fusion method is better than these two compared strategies, which shows that our predictive level fusion can fully exploit the uncertainty of two modalities.

F. Uncertainty Analysis

The deep evidential learning of uncertainty computes both data and model uncertainties. However, data uncertainty cannot be reduced through model improvement, whereas model uncertainty reflects the predictive capability of the model. We compute the Pearson correlation coefficient matrix among the data uncertainty, the model uncertainty and the MAE metric tested on the KITTI depth completion dataset and the Virtual KITTI2 dataset in Fig. 7. The correlation coefficients between model uncertainty and MAE are high on both datasets, with values of 0.61 and 0.67, respectively. The positive correlation coefficient indicates that our predicted uncertainty is able to effectively reflect the errors. Therefore, during inference, the accuracy of predictions can be inferred based on their associated uncertainty. For the two datasets, the correlation coefficients between model uncertainty and data uncertainty are 0.73 and 0.94, respectively. The model uncertainty and data uncertainty are highly correlated, indicating that data uncertainty plays a major role to affect the model uncertainty. We also provide a visualization of the data uncertainty and model uncertainty on KITTI depth completion dataset in Fig.

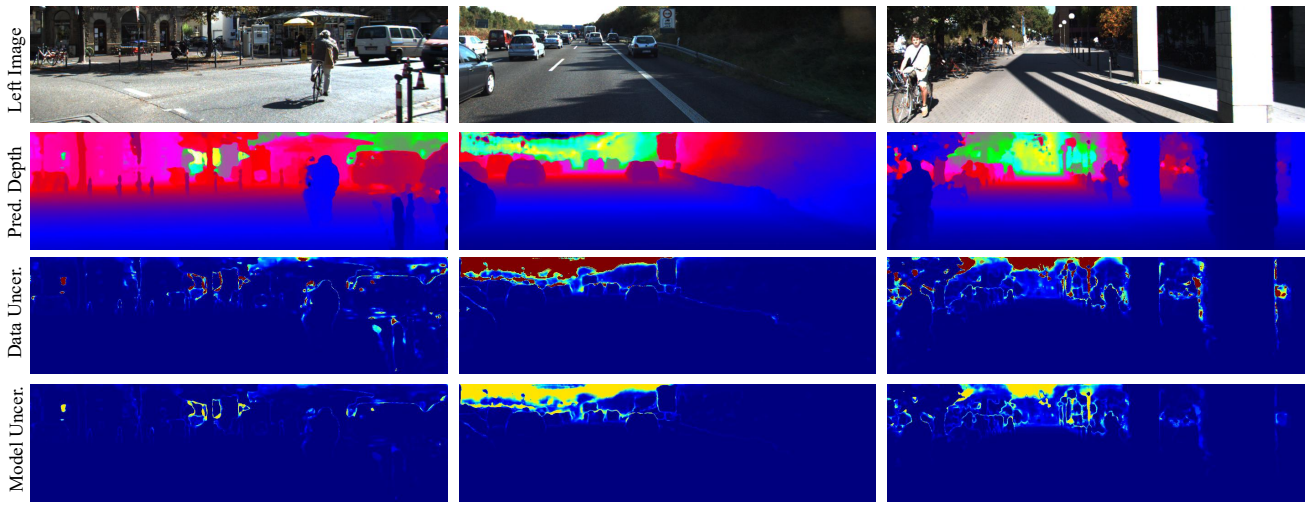


Fig. 8. The top row in the illustration represents the left image, while the second row depicts the fused depth map. The third and fourth rows correspond to the data uncertainty and model uncertainty, respectively. Both data uncertainty and model uncertainty exhibit higher values in the boundary regions of objects and at farther distances.

TABLE V
ABLATION EXPERIMENTS WITH DIFFERENT MODULE COMBINATIONS ON KITTI DEPTH COMPLETION VALIDATION DATASET.

Intra Evidential Fusion	MFA	Inter Evidential Fusion	RMSE↓	MAE↓	IRMSE↓	IMAE↓
✓	✓		651.1	224.1	1.78	0.97
✓		✓	621.6	219.4	**1.39**	0.82
	✓	✓	604.4	203.8	1.49	0.84
✓	✓	✓	**599.3**	**190.0**	1.43	**0.78**

TABLE VI
RESULTS OF DIFFERENT DEPTH FUSION STRATEGIES ON KITTI DEPTH COMPLETION VALIDATION DATASET.

Methods	RMSE↓	MAE↓	IRMSE↓	IMAE↓
Average	651.1	224.1	1.78	0.97
Relative	625.6	223.0	1.43	0.84
Ours	**599.3**	**190.0**	**1.43**	**0.78**

8 from three samples. As observed, the data and model uncertainties tend to exhibit higher values at the boundaries of objects, such as pedestrians and vehicles, which is consistent with the intuition that the decisions alone the boundaries are more challenging. Predictions at long distances also exhibit a high degree of uncertainty, such as the sky.

V. CONCLUSIONS

In this paper, we propose a novel evidential based stereo-LiDAR fusion for depth estimation. To better exploit the complementary relationship of LiDAR and stereo modalities, we introduce the MFA module at the feature level for early multi-modal fusion. Our network computes pixel-wise uncertainties and uses them to combine the depth from stereo matching and that from depth completion to get the final depth map. Compared to previous stereo-LiDAR fusion techniques, our HCENet is able to produce uncertainties for each modality and achieve trustworthy fusion. The fusion requires higher computational cost compared with stereo only or depth completion only approaches. For practical deployment at edge side, more efforts such as model compression shall be adopted to reduce the parameters for improved inference speed. Experimental results show that the proposed method outperforms state-of-the-art methods.

REFERENCES

[1] P. Nguyen, K. G. Quach, C. N. Duong, N. Le, X.-B. Nguyen, and K. Luu, "Multi-camera multiple 3d object tracking on the move for autonomous vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 2569–2578.

[2] M. Franzius, M. Dunn, N. Einecke, and R. Dirnberger, "Embedded robust visual obstacle detection on autonomous lawn mowers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 44–52.

[3] N. Garnett, S. Silberstein, S. Oron, E. Fetaya, U. Verner, A. Ayash, V. Goldner, R. Cohen, K. Horn, and D. Levi, "Real-time category-based and general obstacle detection for autonomous driving," in *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2017, pp. 198–205.

[4] X. Dong, M. A. Garratt, S. G. Anavatti, and H. A. Abbass, "Towards real-time monocular depth estimation for robotics: A survey," *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 23, no. 10, pp. 16 940–16 961, Oct. 2022.

[5] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *Proceedings of the International Conference on 3D Vision (3DV)*, 2017, pp. 11–20.

[6] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proceedings of the International Conference on 3D Vision (3DV)*, 2016, pp. 239–248.

[7] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4796–4803.

[8] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 21 919–21 928.

[9] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5410–5418.

[10] Z. Liang, Y. Guo, Y. Feng, W. Chen, L. Qiao, L. Zhou, J. Zhang, and H. Liu, "Stereo matching using multi-level cost volume and multi-scale feature constancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 43, no. 1, pp. 300–315, Jan. 2021.

[11] S. Cheng, Z. Xu, S. Zhu, Z. Li, L. E. Li, R. Ramamoorthi, and H. Su, "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2524–2534.

[12] Y. Mao, Z. Liu, W. Li, Y. Dai, Q. Wang, Y.-T. Kim, and H.-S. Lee, "Uasnet: Uncertainty adaptive sampling network for deep stereo matching," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 6311–6319.

[13] S. M. N. Uddin, S. H. Ahmed, and Y. J. Jung, "Unsupervised deep event stereo for depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 32, no. 11, pp. 7489–7504, Nov. 2022.

[14] W. Maddern and P. Newman, "Real-time probabilistic fusion of sparse 3d lidar and dense stereo," in *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 2181–2188.

[15] K. Park, S. Kim, and K. Sohn, "High-precision depth estimation with the 3d lidar and stereo fusion," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 2156–2163.

[16] X. Tang, K. Yang, H. Wang, J. Wu, Y. Qin, W. Yu, and D. Cao, "Prediction-uncertainty-aware decision-making for autonomous vehicles," *IEEE Transactions on Intelligent Vehicles (TIV)*, vol. 7, no. 4, pp. 849–862, Dec. 2022.

[17] F. Henze, D. Faßbender, and C. Stiller, "Identifying admissible uncertainty bounds for the input of planning algorithms," *IEEE Transactions on Intelligent Vehicles (TIV)*, vol. 8, no. 4, pp. 3129–3143, 2023.

[18] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 14 927–14 937.

[19] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 22 158–22 169.

[20] Y. Zhang and T. Funkhouser, "Deep depth completion of a single rgb-d image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 175–185.

[21] K. Zeng, Y. Wang, J. Mao, C. Liu, W. Peng, and Y. Yang, "Deep stereo matching with hysteresis attention and supervised cost volume construction," *IEEE Transactions on Image Processing (TIP)*, vol. 31, pp. 812–822, 2021.

[22] Z. Li, W. Zuo, Z. Wang, and L. Zhang, "Confidence-based large-scale dense multi-view stereo," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 7176–7191, 2020.

[23] C. Zhou, Y. Liu, Q. Sun, and P. Lasang, "Vehicle detection and disparity estimation using blended stereo images," *IEEE Transactions on Intelligent Vehicles (TIV)*, vol. 6, no. 4, pp. 690–698, 2021.

[24] H.-C. Yang, P.-H. Chen, K.-W. Chen, C.-Y. Lee, and Y.-S. Chen, "Fade: Feature aggregation for depth estimation with multi-view stereo," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 6590–6600, 2020.

[25] Z. Lu, J. Wang, Z. Li, S. Chen, and F. Wu, "A resource-efficient pipelined architecture for real-time semi-global stereo matching," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 32, no. 2, pp. 660–673, Feb. 2022.

[26] Y. Lee and H. Kim, "A high-throughput depth estimation processor for accurate semiglobal stereo matching using pipelined inter-pixel aggregation," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 32, no. 1, pp. 411–422, Jan. 2022.

[27] K. W. Tong, P. Z. Sun, E. Q. Wu, C. Wu, and Z. Jiang, "Adaptive cost volume representation for unsupervised high-resolution stereo matching," *IEEE Transactions on Intelligent Vehicles (TIV)*, vol. 8, no. 1, pp. 912–922, Jan. 2023.

[28] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 66–75.

[29] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3273–3282.

[30] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1959–1968.

[31] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 6197–6206.

[32] Z. Shen, Y. Dai, X. Song, Z. Rao, D. Zhou, and L. Zhang, "Pcw-net: Pyramid combination and warping cost volume for stereo matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 280–297.

[33] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys, "Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3313–3322.

[34] Y. Zhong, C.-Y. Wu, S. You, and U. Neumann, "Deep rgb-d canonical correlation analysis for sparse depth completion," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 5332–5342.

[35] Z. Yan, K. Wang, X. Li, Z. Zhang, J. Li, and J. Yang, "Rignet: Repetitive image guided network for depth completion," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 214–230.

[36] X. Cheng, P. Wang, and R. Yang, "Depth estimation via affinity learned with convolutional spatial propagation network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 103–119.

[37] G. Chen, J. Lin, and H. Qin, "Uamd-net: A unified adaptive multimodal neural network for dense depth completion," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 33, no. 10, pp. 5406–5419, Oct. 2023.

[38] Y. Yang, A. Wong, and S. Soatto, "Dense depth posterior (ddp) from single image and sparse range," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 3353–3362.

[39] B. Forkel and H.-J. Wuensche, "Lidar-sgm: Semi-global matching on lidar point clouds and their cost-based fusion into stereo matching," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2841–2847.

[40] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, "Penet: Towards precise and efficient image guided depth completion," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 656–13 662.

[41] S. Zhao, M. Gong, H. Fu, and D. Tao, "Adaptive context-aware multi-modal network for depth completion," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 5264–5276, 2021.

[42] K. Rho, J. Ha, and Y. Kim, "Guideformer: Transformers for image guided depth completion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6250–6259.

[43] Y. Zhang, X. Guo, M. Poggi, Z. Zhu, G. Huang, and S. Mattoccia, "Completionformer: Depth completion with convolutions and vision transformers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18 527–18 536.

[44] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[45] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2016, pp. 1050–1059.

[46] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 45, no. 2, pp. 2551–2566, Feb. 2023.

[47] W. Van Gansbeke, D. Neven, B. De Brabandere, and L. Van Gool, "Sparse and noisy lidar completion with rgb guidance and uncertainty," in *Proceedings of the International Conference on Machine Vision Applications (MVA)*, 2019, pp. 1–6.

[48] Z. Shen, Y. Dai, and Z. Rao, "Cfnet: Cascade and fused cost volume for robust stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 13 906–13 915.

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

[50] J. Lou, W. Liu, Z. Chen, F. Liu, and J. Cheng, "Elfnet: Evidential local-global fusion for stereo matching," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023, pp. 17 784–17 793.

[51] L. Bao, L. Wei, W. Zhou, L. Liu, L. Xie, H. Li, and Q. Tian, "Multi-granularity matching transformer for text-based person search," *IEEE Transactions on Multimedia(TMM)*, 2023.

[52] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 5814–5824.

[53] L. Liu, J. An, J. Liu, S. Yuan, X. Chen, W. Zhou, H. Li, Y. F. Wang, and Q. Tian, "Low-light video enhancement with synthetic event guidance," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 37, no. 2, 2023, pp. 1692–1700.

[54] G. Paikin, Y. Ater, R. Shaul, and E. Soloveichik, "Efi-net: Video frame interpolation from fusion of events and frames," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1291–1301.

[55] J. Zhang, M. S. Ramanagopal, R. Vasudevan, and M. Johnson-Roberson, "Listereo: Generate dense depth maps from lidar and stereo imagery," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 7829–7836.

[56] K. Park, S. Kim, and K. Sohn, "High-precision depth estimation using uncalibrated lidar and stereo fusion," *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 21, no. 1, pp. 321–335, Jan. 2020.

[57] T.-H. Wang, H.-N. Hu, C. H. Lin, Y.-H. Tsai, W.-C. Chiu, and M. Sun, "3d lidar and stereo fusion using stereo matching network with conditional cost volume normalization," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 5895–5902.

[58] J. Choe, K. Joo, T. Imtiaz, and I. S. Kweon, "Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation," *IEEE Robotics and Automation Letters (RA-L)*, vol. 6, no. 3, pp. 4672–4679, July. 2021.

[59] X. Cheng, Y. Zhong, Y. Dai, P. Ji, and H. Li, "Noise-aware unsupervised deep lidar-stereo fusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6332–6341.

[60] Z. Xu, Y. Li, S. Zhu, and Y. Sun, "Expanding sparse lidar depth and guiding stereo matching for robust dense depth estimation," *IEEE Robotics and Automation Letters (RA-L)*, vol. 8, no. 3, pp. 1479–1486, March. 2023.

[61] Q. He, Z. Wang, H. Zeng, Y. Zeng, Y. Liu, S. Liu, and B. Zeng, "Stereo rgb and deeper lidar-based network for 3d object detection in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 24, no. 1, pp. 152–162, Jan. 2023.

[62] T. Gruber, F. Julca-Aguilar, M. Bijelic, and F. Heide, "Gated2depth: Real-time dense lidar from gated images," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 1506–1516.

[63] S. Walz, M. Bijelic, A. Ramazzina, A. Walia, F. Mannan, and F. Heide, "Gated stereo: Joint depth estimation from gated and wide-baseline active stereo cues," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 13 252–13 262.

[64] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 561–577.

[65] H. Ma, Z. Han, C. Zhang, H. Fu, J. T. Zhou, and Q. Hu, "Trustworthy multimodal regression with mixture of normal-inverse gamma distributions," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 6881–6893.

[66] Y. Cabon, N. Murray, and M. Humenberger, "Virtual kitti 2," *arXiv preprint arXiv:2001.10773*, 2020.

[67] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4340–4349.

[68] Y. Zhang, L. Wang, K. Li, Z. Fu, and Y. Guo, "Slfnet: A stereo and lidar fusion network for depth completion," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 4, pp. 10 605–10 612, Oct. 2022.

[69] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. So Kweon, "Non-local spatial propagation network for depth completion," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 120–136.

[70] J. Tang, F.-P. Tian, W. Feng, J. Li, and P. Tan, "Learning guided convolutional network for depth completion," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 1116–1129, 2020.

Jiayuan Fan received the Ph.D. degree in information engineering from Nanyang Technological University, Singapore, in 2015. After her graduation, she worked as a Research Scientist at the Institute for Infocomm Research, A*STAR, Singapore. She is currently an associate professor with Academy for Engineering and Technology in Fudan University, Shanghai, China. Her main research interests include computer vision, and image forensic analysis and application.

Haixiang Chen received his Master's degree in 2024 from the Academy for Engineering and Technology in Fudan University, Shanghai, China. His primary research interests include stereo matching, depth completion, and uncertainty estimation.

Weide Liu is currently a Research Scientist at Agency for Science, Technology and Research (A*STAR). Before that, he was a Research Scientist at ByteDance AI Lab in Singapore. Weide received his Ph.D. from Nanyang Technological Unversity, where he was advised by Prof. Guosheng Lin, and he also earned a Bachelor's degree from the university's School of Electrical and Electronic Engineering (EEE). His research interests include computer vision, vision language, machine learning, federated learning, and medical image analysis.

Xun Xu received the B.E. degree from Sichuan University, in 2010 and the PhD degree from Queen Mary University of London in 2016. He was a research fellow with National University of Singapore between 2016 and 2019. He is now with I2R, A*STAR. His research interests include semi-supervised learning, domain adaptation, zero-shot learning with applications to 3D point cloud data.

Jun Cheng (SM'20) received the B. E. degree from the University of Science and Technology of China, and the PhD degree from Nanyang Technological University, Singapore. He is now a principal research scientist in the Institute for Infocomm Research, Agency for Science, technology and Research, working on AI for medical imaging, robust machine vision and perception. He is an Associate Editor for IEEE Transactions on Image Processing and IEEE Transactions on Medical Imaging.