ELFNet: Evidential Local-global Fusion for Stereo Matching

Jieming Lou[1,2], Weide Liu[1], Zhuo Chen[1], Fayao Liu[1], and Jun Cheng*[1]

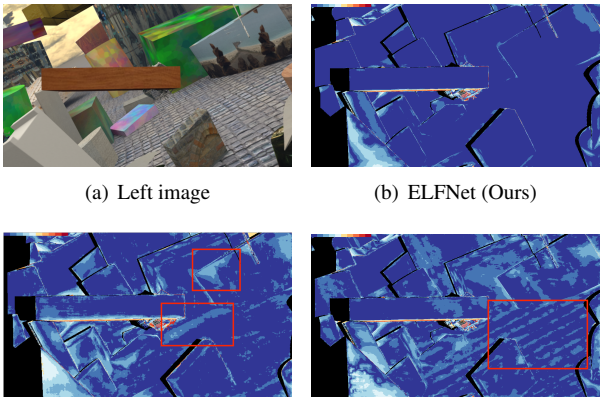
[1]Institute for Infocomm Research, A*STAR
[2]National Univeristy of Singapore

Abstract

*Although existing stereo matching models have achieved continuous improvement, they often face issues related to trustworthiness due to the absence of uncertainty estimation. Additionally, effectively leveraging multi-scale and multi-view knowledge of stereo pairs remains unexplored. In this paper, we introduce the **E**vidential **L**ocal-global **F**usion (ELF) framework for stereo matching, which endows both uncertainty estimation and confidence-aware fusion with trustworthy heads. Instead of predicting the disparity map alone, our model estimates an evidential-based disparity considering both aleatoric and epistemic uncertainties. With the normal inverse-Gamma distribution as a bridge, the proposed framework realizes intra evidential fusion of multi-level predictions and inter evidential fusion between cost-volume-based and transformer-based stereo matching. Extensive experimental results show that the proposed framework exploits multi-view information effectively and achieves state-of-the-art overall performance both on accuracy and cross-domain generalization. The codes are available at* `https://github.com/jimmy19991222/ELFNet`.

1. Introduction

Stereo matching, which aims at estimating the dense disparity map given a pair of rectified images, is one of the most fundamental problems in various applications, such as 3D reconstruction, autonomous driving, and robotics navigation [14]. Benefiting from the rapid development of convolutional neural networks, many stereo matching models have achieved promising performance by constructing cost volume and using 3D convolutions [13, 36–38]. Recently, with the support of transformer, approaches have been proposed to utilize global information using self- and cross-attention mechanisms, bringing an alternative way for the stereo matching [12, 18].



(a) Left image (b) ELFNet (Ours) (c) PCWNet (Cost-volume based) (d) STTR (Transformer based)

Figure 1. Error map visualization and comparison of the proposed ELFNet, PCWNet (a cost-volume-based model) and STTR (a transformer-based model) on a case of the Scene Flow [23]. The regions colored in blue indicate low error, while the regions in white and red indicate relatively higher error. By fusing cost-volume-based and transformer-based models reliably, our ELFNet significantly reduces the error.

Despite the improved performance, quantifying uncertainty of the stereo matching results has been overlooked. The frequently occurring overconfident predictions in the existing stereo matching limit the deployment of the algorithms, especially in safety-critical applications. The deep learning models are prone to be unreliable due to the lack of interpretability, especially when facing out-of-domain, low-quality or perturbed samples. Things are even worse in the field of stereo matching where the model is first pretrained in a large scaled synthetic dataset [23] and fine-tuned in a much smaller dataset from real-world scenes. This makes uncertainty estimation an essential part of preventing potentially disastrous decisions based on stereo matching results.

In the meantime, multi-view complementary information widely exists in stereo matching, but it remains a challenge to harness them to improve accuracy effectively and

*cheng_jun@i2r.a-star.edu.sg

efficiently. For instance, multi-scale pyramidal cost volumes are used to offer the coarse-to-fine knowledge obtained from the feature extractor [3, 4, 13, 31, 32, 36, 40], but the current fusion method fails to consider the uncertainties in different scales, which results in an untrustworthy and incomplete integration. In addition, cost-volume-based [32, 37] and transformer-based approaches [18] provide entirely different strategies for dealing with stereo pairs: the former aggregates local features with convolutions, and the latter captures global information with transformer for dense matching. We find that these two types of methods complementary to each other. For instance, as shown in the red blocks of Figure 1(c) and 1(d), cost-volume-based model is not robust in regions with large illumination changes while transformer-based model does not make full use of complex local textures. In such a scenario, uncertainty estimation is a potential module for endowing a trustworthy fusion strategy among multi-view information to alleviate the error risks without bringing much additional computational load.

With these motivations in mind, we propose an *Evidential Local-global Fusion* (ELF) framework for stereo matching to kill two birds with one stone (see Figure 2). The framework enables both uncertainty estimation and reliable fusion by taking advantage of deep evidential learning [1, 30, 42]. Specifically, we employ trustworthy heads in each branch of the model to compute the aleatoric and epistemic uncertainties [8, 15] along with the disparity. To integrate the multi-scale cost volume information and the complementary information between convolution based and transformer based approaches simultaneously, we propose an intra evidential fusion module and an inter evidential fusion module with a mixture of normal-inverse Gamma distributions (MoNIG) [7, 21]. As shown in Figure 1(b), the proposed ELFNet attains a low disparity error in most regions by leveraging respective strengths of the cost-volume-based model [32] and the transformer-based model [18] according to the evidence dynamically.

Our contributions can be summarized as follows:

1. We introduce deep evidential learning to both cost-volume-based and transformer-based stereo matching to estimate both aleatoric and epistemic uncertainties;

2. We propose a novel evidential local-global fusion (ELF) framework, which enables both uncertainty estimation and two-stage information fusion based on evidence;

3. We conduct comprehensive experiments, which demonstrate that the designed ELFNet consistently boost the performance in terms of accuracy and cross-domain generalization.

2. Related Works

2.1. Deep Stereo Matching

Cost-volume-based Deep Stereo Matching Methods. Cost-volume-based deep stereo matching methods are widely used and have achieved promising results. DispNetC [24] is the first end-to-end trainable stereo matching framework using the dot product of left and right feature maps to form a 3D cost volume. Despite of a computational-friendly way, the correlation operation cannot capture enough knowledge to obtain a satisfactory result. Following GC-Net [16], many works [3, 5, 41] employ 3D hourglass convolutions to aggregate 4D cost volume with the size of [height×width×disparity range×feature] constructed by concatenating the features of stereo pairs, which demands large memory and computational complexity. Group-wise correlation is then proposed in Gwc-Net [13] to construct a compact cost volume and facilitates a better trade-off. ACVNet [37] proposes to build attention-aware cost volume to suppress redundant information and further alleviate the burden of cost aggregation.

To reduce the high computational cost and leverage more semantic and robust information, multi-scale cost volume is introduced to stereo matching. HSMNet [39] constructs a pyramid of volumes to process high-res stereo images. AANet [38] is a lightweight framework with a feature pyramid and multi-scale correlation volume interaction. CFNet [31] constructs cascade pyramid cost volume to narrow down the disparity search range and refine the disparity map in a coarse-to-fine manner. Most recently, PCWNet [32] proposes a volume fusion module to directly combine multi-scale 4D volumes and calculate a multi-level loss to accelerate the convergence of the model.

Transformer-based Deep Stereo Matching Methods. With the support of attention mechanisms, transformer-based deep stereo matching methods have become another line of stereo matching research. STTR [18] matches pixels from a sequence-to-sequence matching perspective to impose a uniqueness constraint and avoid the construction of fixed disparity cost volume. CSTR [12] employs a plug-in module, Context Enhanced Path, to help better integrate global information and deal with hazardous regions, such as texturelessness, specularity, or transparency.

Generally, transformer-based methods excel at modeling long-range global knowledge, but do not perform well in regions with local texture details. This encourages us to improve the overall performance by fusing transformer-based and cost-volume-based methods to capture complementary information. In this paper, we use PCWNet [32] and STTR [18] as examples for fusion, but other cost-volume-based and transformer-based approaches can be used as well.

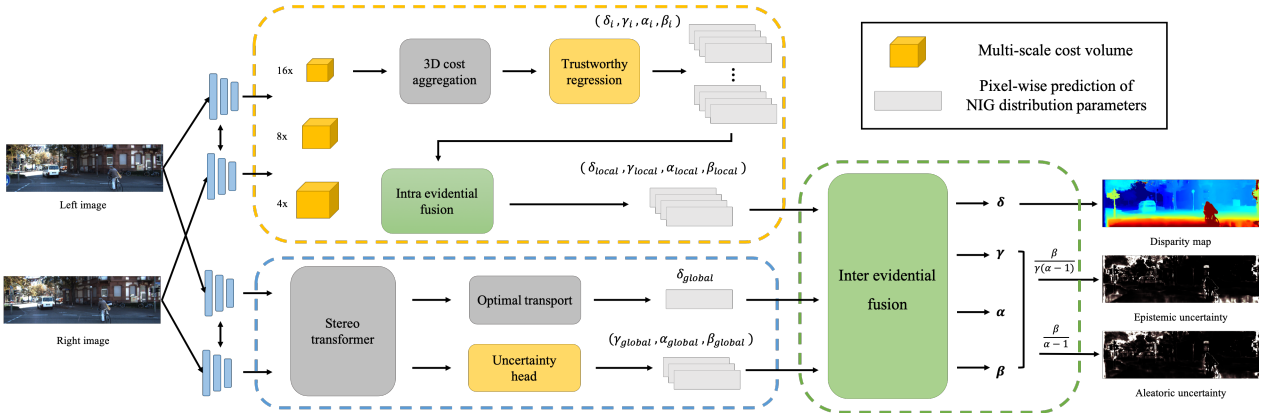


Figure 2. Illustration of the proposed evidential local-global fusion (ELF) for stereo matching. The model architecture comprises three parts: the cost-volume-based module with intra evidential fusion(yellow block), the transformer-based module(blue block), and the inter evidential fusion module(green block). The framework leverages evidential estimation to accomplish two-stage fusion and generate both aleatoric and epistemic uncertainty maps along with disparity.

2.2. Uncertainty Estimation

As deep learning techniques are increasingly applied in safety-sensitive real-world scenarios, the measurement of confidence thus becomes vital for distinguishing the doubtful predictions and aiding in decision making [10]. Bayesian neural networks [15] enable deep models with uncertainty by substituting the deterministic weight parameters with distributions. However, the computational expense of optimization is unaffordable with such a huge amount of parameters. Monte Carlo dropout (MC Dropout) [9] is the most well-known method to tackle the problem, which formulates dropout as Bernoulli distributed random variables to approximate the training process as variational inference. Deep ensemble methods [6, 17, 35] predict the overall result based on multiple models with different architectures and have gained popularity in modeling uncertainty in the past several years. Several methods are introduced to help ensemble methods become more practical, such as pruning [2] and distillation [19]. Deterministic neural network [20, 27, 33] is a more efficient estimation approach, which directly computes the uncertainty of prediction distributions based on one single forward pass. For instance, evidential deep learning [30] and prior networks [22] place Dirichlet priors over discrete classification predictions. Deep evidential regression [1] proposes to extend [30] to regression tasks to estimate the parameters of the normal inverse gamma distribution over an underlying Normal distribution, which ensures explicit representation of epistemic and aleatoric uncertainties. Considering a multi modalities setting, Ma et al. [21] uses a mixture of the normal inverse gamma distribution (MoNIG) to allow a trustworthy regression which characterizes both modality-specific and integrated uncertainties.

Several works seek to incorporate uncertainty estimation into stereo tasks. UCSNet [4] proposes calculating adaptive thin volume with an uncertainty-aware cascaded design in a multi-view stereo setting. Inspired by UCSNet, CFNet [31] employs uncertainty estimation to adjust the disparity search range adaptively. In contrast to the previous methods using variance-based uncertainty, Wang et al. [34] first applies deep evidential learning to predict uncertainties of disparity map in stereo matching. In this paper, we further extend deep evidential learning to fully utilize the multi-level knowledge in stereo matching task with both intra and inter evidential fusion strategies.

3. Method

This section explains the proposed Evidential Local-global Fusion (ELF) framework based on uncertainty estimation for stereo matching. As illustrated in Figure 2, our network architecture is divided into three parts: the cost-volume-based module with intra evidential fusion, the transformer-based module, and the inter evidential fusion module. Given a stereo pair, pyramid combination network with intra evidential fusion and trustworthy stereo transformer respectively predict the distribution parameters $\{\delta_{local}, \gamma_{local}, \alpha_{local}, \beta_{local}\}$ and $\{\delta_{global}, \gamma_{global}, \alpha_{global}, \beta_{global}\}$. Then, disparity, aleatoric uncertainty and epistemic uncertainty are deduced from the syncretic $\{\delta, \gamma, \alpha, \beta\}$, which are obtained by inter evidential fusion module based on the multi-view mixture of the normal-inverse gamma distribution.

3.1. Evidential Deep Learning for Stereo Matching

3.1.1 Background and Uncertainty Loss

Stereo matching strives to estimate the disparity between the given stereo pair for each pixel. From the viewpoint of evidential learning, every disparity d is drawn from a

normal distribution, but with unknown mean and variance (μ, σ^2). To model the distribution, μ and σ^2 are assumed to be drawn from normal and inverse-gamma distributions respectively:

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

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

With the assumption that the mean and the variance are independent, the posterior distribution $q(\mu, \sigma^2) = p(\mu, \sigma^2 | d_1, ..., d_N)$ can be formulated as a normal-inverse gamma distribution NIG$(\delta, \gamma, \alpha, \beta)$. Amini et al. [1] then define the total evidence $\Phi = 2\gamma + \alpha$ intuitively to measure the confidence of the predictions. The disparity d, aleatoric uncertainty al and epistemic uncertainty ep can be then derived as

$$\begin{aligned} d = \mathbb{E}(\mu) = \sigma,\ al = \mathbb{E}(\sigma^2) = \frac{\beta}{\alpha - 1}, \\ ep = \mathrm{Var}(\mu) = \frac{\beta}{\gamma(\alpha - 1)}, \end{aligned} \quad (2)$$

During training, the loss function $\mathcal{L}^N$ can be defined as the negative logarithm of model evidence,

$$\begin{aligned} \mathcal{L}^N(w) = \frac{1}{2}\log(\frac{\pi}{\gamma}) - \alpha\log(\Omega) + \\ (\alpha + \frac{1}{2})\log((y - \delta)^2\gamma + \Omega) + \log\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha + \frac{1}{2})}\right), \end{aligned} \quad (3)$$

where $\Omega = 2\beta(1 + \gamma)$, w denotes the set of the estimated distribution parameters.

To reduce the evidence where prediction is incorrect, a regularization term is introduced,

$$\mathcal{L}^R(w) = |d^{gt} - \mathbb{E}(\mu_i)| \cdot \Phi = |d^{gt} - \delta| \cdot (2\gamma + \alpha), \quad (4)$$

where d^{gt} is the ground truth disparity map.

To extend deep evidential learning to the dense stereo matching task, the total uncertainty loss $\mathcal{L}^U$ can be defined as the expectation over all the pixels,

$$\mathcal{L}^U(w) = \frac{1}{N}\sum_{0}^{N-1}\left(\mathcal{L}_i^N(w) + \tau\mathcal{L}_i^R(w)\right), \quad (5)$$

where $\tau > 0$ controls the degree of regularization, N denotes the total number of pixels.

3.1.2 Uncertainty Estimation in Stereo Matching

Uncertainty estimation in cost-volume-based stereo matching. Cost-volume-based stereo matching networks have five typical main modules in detail: shared-weights feature extraction, cost volume construction, 3D cost aggregation, disparity regression module, and disparity refinement. To estimate the parameters of NIG distribution rather than merely predicting the disparity map, we substitute the disparity regression module into a trustworthy regression with multi-channel output and keep the remaining modules unchanged. Shown in Figure 3 (top), considering the computational complexity of 3D convolution operation, the proposed trustworthy regression module employs a single branch of two 3D convolutions with Mish activation [26] and the up-sample method to output a 4-channel volume $V_{output} \in \mathbb{R}^{D_{max} \times H \times W \times 4}$, where H and W are the height and width of the input stereo pairs, and D_{max} is the max value of disparity candidates. The distribution parameters can be computed as follows:

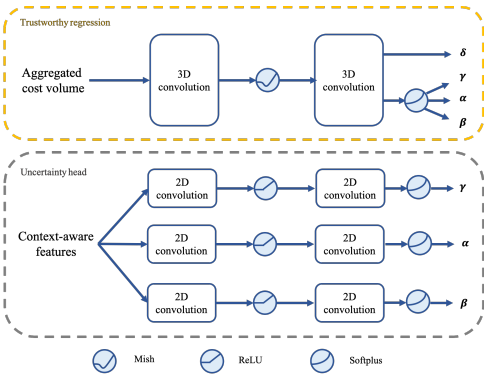


Figure 3. Overview of the trustworthy regression in cost-volume-based stereo matching(top) and uncertainty head in transformer-based stereo matching(bottom). Uncertainty head only predicts γ, α, and β, whereas δ is generated by optimal transport module.

$$V_\delta, V_\gamma, V_\alpha, V_\beta = \mathrm{Split}(V_{out}, dim = -1), \quad (6)$$

$$p = \mathrm{Softmax}(V_\delta), \quad (7)$$

$$\delta = \sum_{k=0}^{D} k \cdot p_k,\ \mathrm{logit}_i = \sum_{k=0}^{D} V_i \cdot p_k, \quad (8)$$

where k denotes disparity level, p denotes the probability, and $i \in \{\gamma, \alpha, \beta\}$. Then a Softplus activation is applied on the logits_i to generate γ, α, β (and additional plus 1 to ensure $\alpha > 1$).

Uncertainty estimation in transformer-based stereo matching. Transformer-based stereo matching networks take advantage of cross- and self-attention mechanisms to predict the expectation of disparity δ and occlusion probability p_{occ} without fixed-disparity cost volume and 3D convolutions. The Stereo Transformer outputs the attention weights w_{attn}, then the optimal transport module regresses the disparity and calculates occlusion probability based on the cost matrix, which is set as $-w_{attn}$. On top of this, we add an uncertainty head consisting of two 2D convolution blocks and Softplus activation(shown in Figure 3 (bottom))

to generate the parameters γ, α, β from the concatenation of left and right context-aware features converted by the Stereo Transformer. To obtain better-calibrated results, we utilize the context adjustment layer in accordance with STTR [18] to refine the parameter δ and the occlusion probability p_{occ}.

3.2. Fusion Strategy based on Evidence

We adopt the fusion strategy with the mixture of normal-inverse gamma distribution (MoNIG) [21] for its excellent mathematical properties to perform both intra evidential fusion and inter evidential fusion. Specifically, given M sets of parameters of NIG distributions, the MoNIG distribution can be computed with the following operations:

$$\begin{aligned}&\text{MoNIG}(\delta,\gamma,\alpha,\beta) = \text{NIG}(\delta_1,\gamma_1,\alpha_1,\beta_1)\oplus \\ &\text{NIG}(\delta_2,\gamma_2,\alpha_2,\beta_2)\oplus\cdots\oplus\text{NIG}(\delta_M,\gamma_M,\alpha_M,\beta_M),\end{aligned} \tag{9}$$

where $\oplus$ represents the summation operation of two NIG distributions, which is defined as

$$\text{NIG}(\delta,\gamma,\alpha,\beta) \triangleq \text{NIG}(\delta_1,\gamma_1,\alpha_1,\beta_1)\oplus\text{NIG}(\delta_2,\gamma_2,\alpha_2,\beta_2),$$

where

$$\begin{aligned}&\delta = (\gamma_1+\gamma_2)^{-1}(\gamma_1\delta_1+\gamma_2\delta_2),\\ &\gamma = \gamma_1+\gamma_2,\ \alpha = \alpha_1+\alpha_2+\frac{1}{2},\\ &\beta = \beta_1+\beta_2+\frac{1}{2}\gamma_1(\delta_1-\delta)^2+\frac{1}{2}\gamma_2(\delta_2-\delta)^2.\end{aligned} \tag{10}$$

The parameter δ of the combined distribution is the summation of δ_1 and δ_2 weighted by γ, which measures the confidence level of the expectation. The final β is defined as not only the sum of β_1 and β_2, but also the variance between the combined distribution and each individual distribution, as it provides insight into aleatoric and epistemic uncertainties simultaneously.

3.2.1 Intra Evidential Fusion of Cost-volume-based Stereo Matching

Multi-scale cost volume has been frequently used in stereo matching tasks to exploit the features from different layers of the extractor. We construct three levels of the group-wise correlation 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 3D cost aggregation and trustworthy regression module to generate parameters of NIG distributions $(\delta_i,\gamma_i,\alpha_i,\beta_i)$, where $i\in\{1,2,3\}$. Intra evidential fusion module integrates three NIG distributions into one distribution as the final pyramid combined result

$$\begin{aligned}&\text{MoNIG}(\delta_{local},\gamma_{local},\alpha_{local},\beta_{local}) = \\ &\text{NIG}(\delta_1,\gamma_1,\alpha_1,\beta_1)\oplus\cdots\oplus\text{NIG}(\delta_3,\gamma_3,\alpha_3,\beta_3).\end{aligned} \tag{11}$$

The uncertainty-aware fusion strategy equips our framework with the ability to integrate reliable outputs from multi-scale features.

3.2.2 Inter Evidential Fusion between Cost-volume-based and Transformer-based Stereo Matching

The intrinsic inductive bias of locality of convolutions makes cost-volume-based stereo matching models easy to model local features, whereas the transformer-based models capitalize on the long-range dependencies of attention mechanism to capture global information. The different foci of the two methods lead to the difference in strengths and weaknesses for predicting disparities and are likely to complement one another in some instances. The inter evidential fusion with MoNIG distribution provides an elegant and computationally efficient mechanism to merge two predictions into one. We apply the fusion strategy based on uncertainty to obtain the final distribution

$$\begin{aligned}\text{MoNIG}(\delta,\gamma,\alpha,\beta) = &\text{MoNIG}(\delta_{local},\gamma_{local},\alpha_{local},\beta_{local})\\ &\oplus\text{NIG}(\delta_{global},\gamma_{global},\alpha_{global},\beta_{global}).\end{aligned} \tag{12}$$

3.3. Loss

We compute the uncertainty losses on local outputs, global outputs and final combined outputs, denoted as $\mathcal{L}^U(w_{local})$, $\mathcal{L}^U(w_{global})$ and $\mathcal{L}^U(w)$, respectively. In the transformer-based stereo matching module, we obtain the attention weights and occlusion probability p_{occ} as well. Besides the uncertainty loss, we adopt the same loss functions as STTR [18], relative response loss $\mathcal{L}^{RR}(w_{attn})$ to maximize the attention on the true target location and binary-entropy loss $\mathcal{L}^{BE}(p_{occ})$ to supervise occlusion map. The overall loss function is described as

$$\begin{aligned}&\mathcal{L} = \mathcal{L}^U(w_{local})+\lambda_1\mathcal{L}^U(w_{global})\\ &+\lambda_2\mathcal{L}^U(w)+\lambda_3\mathcal{L}^{RR}(w_{attn})+\lambda_4\mathcal{L}^{BE}(p_{occ}),\end{aligned} \tag{13}$$

where λ_i, $i=1,2,3,4$ control the loss weights.

4. Experiments

In this section, we evaluate the proposed ELFNet on various datasets including Scene Flow [23], KITTI 2012 & KITTI 2015 [11, 25] and Middlebury 2014 [29]. Additionally, we conduct uncertainty analyses to explore the relationship between model performance and uncertainties.

4.1. Datasets and Evaluation Metrics

Scene Flow FlyingThings3D subset [23] is a large-scale synthetic dataset of random objects. The subset provides about 25,000 stereo frames with the resolution of

Table 1. Ablation study on Scene Flow [23].

Components				Scene Flow	
STTR [18]	Uncertainty	Inter Evidential Fusion	Intra Evidential Fusion	EPE(px)↓	D1-1px(%)↓
✓				0.42	1.37
✓	✓			0.41	1.31
✓	✓	✓		0.38	1.32
✓	✓	✓	✓	**0.33**	**1.28**

960×540 and corresponding sub-pixel ground truth disparity maps and occlusion regions.

KITTI 2012 & KITTI 2015 [11, 25] are collected from the real-life driving scenario. KITTI 2012 provides 194 training and 195 testing images pairs, and KITTI 2015 provides 200 training and 200 testing image pairs. The resolution of KITTI 2012 and KITTI 2015 stereo pairs are 1226×370 and 1242×375. Both datasets provide sparse disparity maps.

Middlebury 2014 [28] is an indoor dataset with 15 training image pairs and 15 testing image pairs. It contains a set of 23 high resolution stereo pairs in three different resolutions, and we select the quarter-resolution ones.

Evaluation metrics. We use end point error (EPE), percentage of disparity outliers by 1px or 3px (D1-1px / D1-3px) and percentage of errors larger than 3px (3px Err) as evaluation metrics for disparity prediction.

4.2. Implementation Details

Our ELF framework is inherently compatible with all transformer-based and multi-scale cost-volume-based models. For our experiments, we set STTR [18] as the transformer-based part and PCWNet [32] as the cost-volume-based part for its good balance between accuracy and generalization. We train the model in an end-to-end manner using AdamW optimizer with a weight decay of 1e-4. We pre-train on Scene Flow FlyingThings3D subset [23] for 16 epochs with the initial learning rate of 2e-4 for the transformer-based part and 2e-3 for the cost-volume-based part. After epochs 4, 8, 10, and 12, we decrease the learning rate by a factor of 2. We use 6 self- and cross-attention layers with 4 heads for the transformer-based part and set the maximum disparity $D = 192$ for the cost-volume-based part. The regularization weight in the uncertainty loss $\tau = 0.5$. The weights of four outputs are set as $\lambda_1 = 1.0$, $\lambda_2 = 2.0$, $\lambda_3 = 1.0$, $\lambda_4 = 1.0$ empirically. We apply several data augmentations to stimulate real-world scenarios during training, including Gaussian noise, brightness/contrast shift, and random size cropping. All the experiments are conducted on a NVIDIA RTX 3090 GPU and implemented with PyTorch. The codes are available at `https://github.com/jimmy19991222/ELFNet`.

Table 2. Comparison with state-of-the-art on Scene Flow [23].

	Disparity <192		All Pixels	
	EPE(px)↓	D1-1px(%)↓	EPE(px)↓	D1-1px(%)↓
PSMNet [3]	0.95	2.71	1.25	3.25
GwcNet [13]	0.76	3.75	3.44	4.65
CFNet [31]	0.70	3.69	1.18	4.26
PCWNet [32]	0.85	1.94	0.97	2.48
GANet [41]	0.48	4.02	0.97	4.89
STTR [18]	0.42	1.37	0.45	**1.38**
CSTR [12]	0.41	1.41	0.45	1.39
ELFNet(Ours)	**0.33**	**1.28**	**0.40**	1.39

4.3. Comparison with State-of-the-art

To evaluate the effectiveness of the proposed approach, we compare it to several state-of-the-art methods, including PSMNet [3], GwcNet [13], CFNet [31], PCWNet [32], GANet [41], STTR [18] and CSTR [12]. Table 2 presents a comparison of our method with the previous state-of-the-art methods on Scene Flow [23]. Our proposed method outperforms all other methods in both EPE and D1-1px metrics with different settings. Notably, our approach demonstrates an improvement of 19.5% (0.33 v.s. 0.41) in EPE and 9.2% (1.28 v.s. 1.41) in D1-1px compared to the best performing method, CSTR [12], in the Disparity < 192 setting.

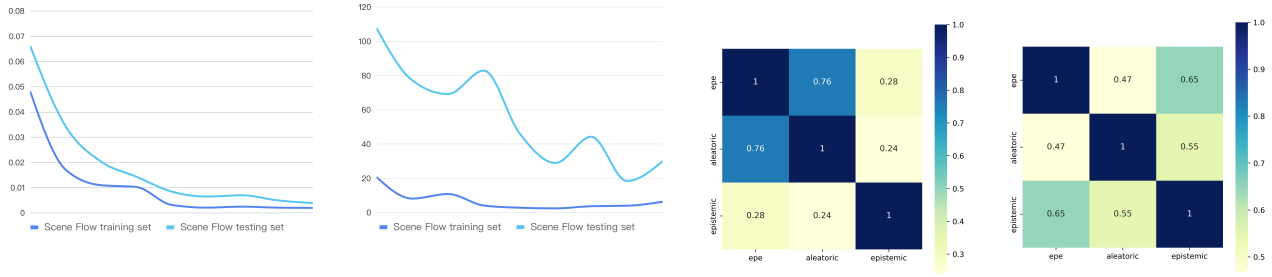
In the All Pixels setting, our approach demonstrates a reduced EPE over the current state-of-the-art method by 11.2% from 0.45 to 0.40. Overall, our ELFNet outperforms the cost-volume-based and transformer-based models in disparity estimation accuracy by leveraging their advantages in a trustworthy manner. Meanwhile, it also inherits and maintains occlusion estimation from transformer. It achieves comparable occlusion estimation with occlusion intersection over union score 0.98 compared with 0.97 by STTR [18].

4.4. Ablation Studies

To verify the effects of modules in our framework, we provide quantitative results in Table 1. The proposed ELF framework has three main designs, including uncertainty estimation, inter evidential fusion, and intra evidential fusion. Uncertainty estimation enables the model to predict the epistemic and aleatoric uncertainties respectively; the inter evidential fusion combines the information of the transformer-based and the cost-volume based models; and the intra evidential fusion leverages different levels of local

Table 3. Cross-domain evaluation without *fine-tuning* on Middleburry 2014 [29], KITTI 2012 [11] and KITTI 2015 [25].

	Middlebury 2014		KITTI 2012		KITTI 2015	
	EPE(px)↓	3px Err(%)↓	EPE(px)↓	D1-3px(%)↓	EPE(px)↓	D1-3px(%)↓
PSMNet [3]	3.05	13.0	3.46	25.9	6.59	16.3
GwcNet [13]	1.89	8.95	1.68	11.7	2.21	12.2
CFNet [31]	**1.69**	7.73	**0.96**	**4.74**	2.27	**5.76**
PCWNet [32]	2.17	9.09	1.32	5.37	1.88	6.03
STTR [18]	2.33	6.19	1.82	6.96	**1.50**	6.40
ELFNet (Ours)	1.79	**5.72**	1.18	**4.74**	1.57	5.82



(a) Aleatoric uncertainty curve (b) Epistemic uncertainty curve (c) Scene Flow [23] (d) KITTI 2015 [25]

Figure 4. Uncertainty analysis. (a) and (b) are the training curve of two uncertainties on the Scene Flow [23]; (c) and (d) are the heat maps of the Pearson correlation analysis between the EPE and the uncertainties.

knowledge inside the cost-volume-based model.

The ablation study indicates that all three designs are indispensable, and the evidential fusion parts play a crucial role in boosting performance. To be specific, on Scene Flow [23], we observe that STTR [18] with uncertainty estimation module shows a comparable EPE score (0.42 v.s. 0.41) compared with the baseline. When employing the inter evidential fusion module, the model achieves a 0.03 improvement in terms of EPE score. With the additional intra evidential fusion module of the cost-volume-based part, the EPE score of the model boosts by a remarkable 0.05. Overall, the EPE is reduced by 21.4% (0.42 v.s. 0.33) with our ELF framework on Scene Flow [23]. ELFNet also outperforms the baseline by 6.6% (1.28 v.s. 1.37) for the D1-1px metric.

4.5. Fusion Strategies Comparison

Table 4. Fusion strategy comparison on Scene Flow [23]. Avg.: take the average of the separate estimation as the final disparity map. Conv.: add a 2D convolution layer as the late fusion module, and fine-tune for 5 epochs with the former parts frozen.

	EPE(px)↓	D1-1px(%)↓
PCWNet [32]	0.85	1.94
STTR [18]	0.42	1.37
Avg.	0.65	1.75
Conv.	0.42	1.38
ELFNet(Ours)	**0.33**	**1.28**

The proposed ELF framework can be viewed as a powerful late fusion strategy. To further verify the fusion performance, we compare other late fusion strategies with our ELFNet. In Table 4, we observe that simply computing the average of the output disparity maps or employing the convolution layer as a late fusion cannot achieve satisfactory results and may perform worse than STTR [18]. In contrast, ELFNet combines the cost-volume-based model and transformer-based model effectively with improved results.

4.6. Cross-domain Generalization

We conduct experiments to prove that our pre-trained model on the synthetic Scene Flow dataset [23] can achieve strong cross-domain generalization in the zero-shot setting. As shown in Table 3, our proposed ELFNet presents comparable generalization ability with the existing state-of-art models on the real-world datasets. Specially, compared with PCWNet [32] baseline, ELFNet brings a 17.5% EPE score (1.79 v.s. 2.17) and a 37.1% 3px Err score (5.72 v.s. 9.09) improvement on Middlebury 2014 [29], and a 10.6% EPE score (1.18 v.s. 1.32) and 11.7% D1-3px score (4.74 v.s. 5.37) improvement on KITTI 2012 [11]. On the KITTI 2015 [25], our ELFNet also achieves overall competitive generalization results. ELFNet improves 9.1% on the D1-3px metric (5.82 v.s. 6.40) compared with STTR [18].

4.7. Uncertainty Analysis

Uncertainty estimation with deep evidential learning provides both aleatoric and epistemic uncertainties [15] of the prediction. The aleatoric uncertainty reflects the extent to which the disparity value differs from the ground truth, which is related to the data noise and cannot be reduced by optimization, while epistemic uncertainty represents the degree of dispersion of the disparity values, which is related

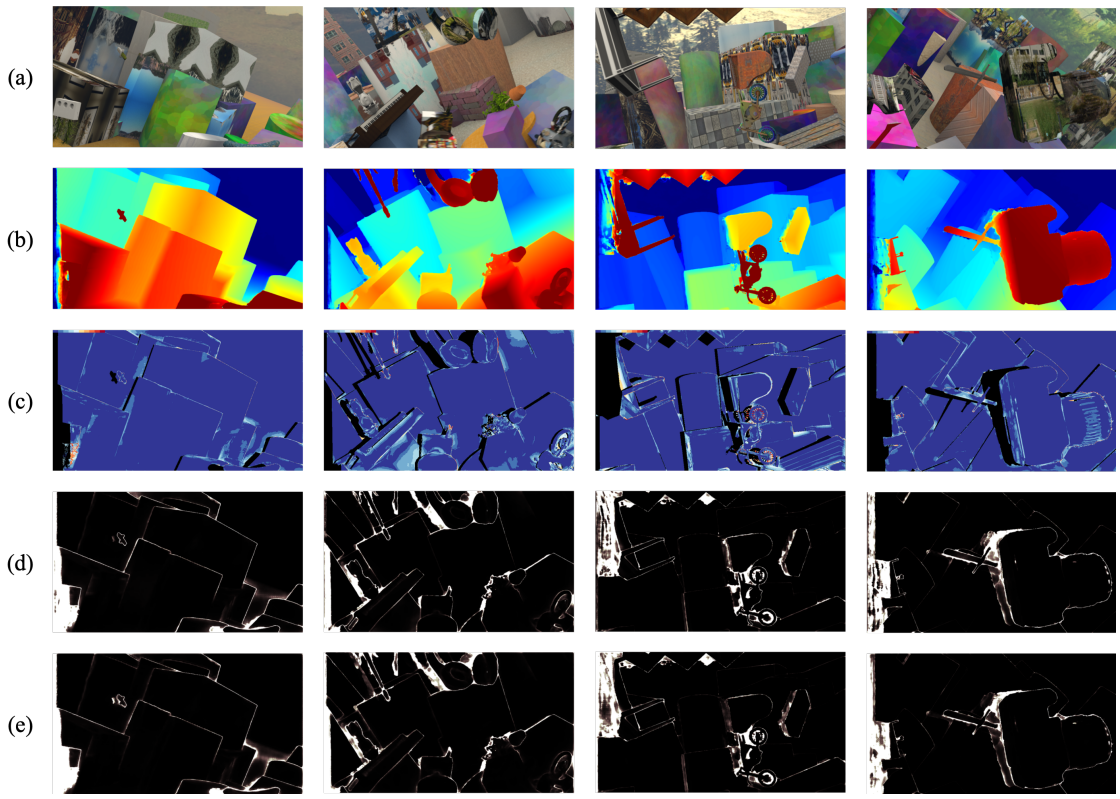


Figure 5. Visualization on Scene Flow of the proposed ELFNet. (a) The left image. (b) The estimated disparity map. (c) The error map. (d) The epistemic uncertainty map. (e) The aleatoric uncertainty map.

to the model capacity.

Figure 4 shows several results of uncertainty analysis. As presented in Figure 4(a) and 4(b), both aleatoric and epistemic uncertainties represent a general decline during training, which shows that the model assigns lower uncertainties as learning more from the data. Based on the ELFNet pretrained on the Scene Flow [23] dataset, we conduct the Pearson correlation analysis between the EPE metric and two uncertainties tested on the Scene Flow [23] and KITTI 2015 [25] separately to shed light on the uncertainty estimation mechanism. The correlation heat maps are shown in Figure 4(c) and 4(d). We have two main observations. Firstly, the correlation value is consistently positive, which indicates that uncertainties are positively correlated to the EPE and can serve as an indicator of accuracy during inference. Secondly, the correlation value between EPE and the aleatoric uncertainty on the Scene Flow [23] dataset is higher than the corresponding value (0.76 v.s. 0.47) on the KITTI 2015 [25] dataset, while the correlation value between EPE and epistemic uncertainty is lower (0.28 v.s. 0.65). It can be inferred that when dealing with out-of-domain samples, the disparity error is more closely related to epistemic uncertainty. In contrast, if the model has been trained on similar data distributions, the error is more likely to be influenced by aleatoric uncertainty.

We provide the qualitative results on Scene Flow [23] in Figure 5. Although there is no ground truth for uncertainties, we observe that high uncertainties are assigned in the occluded and boundary regions. Compared with the error maps, the uncertainty maps are also active in the areas where the error occurs, such as the wheel hub of a bicycle, which suggests that uncertainty maps provide clues for estimation offset.

5. Conclusion

In this paper, we have proposed an Evidential Local-global Fusion (ELF) framework to fuse multi-view information for stereo matching reliably. We leverage deep evidential learning to estimate multi-level aleatoric and epistemic uncertainties alongside the disparity maps, which further allows a trustworthy fusion strategy based on evidence to exploit complementary knowledge. Experimental results show that our model performs well on both accuracy and generalization across different datasets.

Acknowledgement

This work was supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

[2] George DC Cavalcanti, Luiz S Oliveira, Thiago JM Moura, and Guilherme V Carvalho. Combining diversity measures for ensemble pruning. *Pattern Recognition Letters*, 74:38–45, 2016.

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[4] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020.

[5] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Tom Drummond, Hongdong Li, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[6] Kashyap Chitta, Jose M Alvarez, and Adam Lesnikowski. Large-scale visual active learning with deep probabilistic ensembles. *arXiv preprint arXiv:1811.03575*, 2018.

[7] Daniela De Canditiis and Brani Vidakovic. Wavelet bayesian block shrinkage via mixtures of normal-inverse gamma priors. *Journal of Computational and Graphical Statistics*, 13(2):383–398, 2004.

[8] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

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

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[12] Weiyu Guo, Zhaoshuo Li, Yongkui Yang, Zheng Wang, Russell H Taylor, Mathias Unberath, Alan Yuille, and Yingwei Li. Context-enhanced stereo transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 263–279. Springer, 2022.

[13] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3268–3277, 2019.

[14] Mohd Saad Hamid, NurulFajar Abd Manap, Rostam Affendi Hamzah, and Ahmad Fauzan Kadmin. Stereo matching algorithm based on deep learning: A survey. *Journal of King Saud University-Computer and Information Sciences*, 34(5):1663–1673, 2022.

[15] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[16] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017.

[17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[18] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers, 2021.

[19] Jakob Lindqvist, Amanda Olmin, Fredrik Lindsten, and Lennart Svensson. A general framework for ensemble distribution distillation. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2020.

[20] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.

[21] Huan Ma, Zongbo Han, Changqing Zhang, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. *Advances in Neural Information Processing Systems*, 34:6881–6893, 2021.

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

[23] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

[25] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. In *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015.

[26] Diganta Misra. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*, 2019.

[27] Marcin Możejko, Mateusz Susik, and Rafał Karczewski. Inhibited softmax for uncertainty estimation in neural networks. *arXiv preprint arXiv:1810.01861*, 2018.

[28] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 31–42. Springer, 2014.

[29] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, pages 131–140, 2001.

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

[31] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13906–13915, June 2021.

[32] Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 280–297. Springer, 2022.

[33] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR, 2020.

[34] Chen Wang, Xiang Wang, Jiawei Zhang, Liang Zhang, Xiao Bai, Xin Ning, Jun Zhou, and Edwin Hancock. Uncertainty estimation for stereo matching based on evidential deep learning. *Pattern Recognition*, 124:108498, 2022.

[35] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*, 2020.

[36] Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Song Wang, and Lili Ju. Semantic stereo matching with pyramid cost volumes. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7483–7492, 2019.

[37] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12981–12990, 2022.

[38] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1956–1965, 2020.

[39] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019.

[40] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020.

[41] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019.

[42] Xujiang Zhao, Yuzhe Ou, Lance Kaplan, Feng Chen, and Jin-Hee Cho. Quantifying classification uncertainty using regularized evidential neural networks. *arXiv preprint arXiv:1910.06864*, 2019.