2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Consistency-Driven Confidence Estimation for Stereo Matching

Shuheng Lu[1], Zaiwang Gu[2*], Xudong Jiang[1] and Jun Cheng[2*]

***Abstract*— Confidence estimation for stereo matching is crucial for enhancing the reliability and accuracy of depth perception in real-world applications. Despite effectively capturing aleatoric uncertainty through probabilistic modeling and statistical aggregation, current regression-based confidence estimation methods neglect uncertainty arising from unstable training dynamics, resulting in over-confident predictions near occlusion boundaries, textureless regions, and reflective surfaces where errors are most severe. We propose a novel epoch-wise consistency accumulation algorithm that explicitly incorporates training dynamics into confidence estimation. Specifically, we design a full-image cross-epoch alignment mechanism to dynamically quantify pixel-wise training consistency between consecutive epochs, thereby significantly enhancing the estimation of confidence. We further propose a consistency-ranked evidential discrepancy loss, which aligns evidential uncertainty estimates with consistency-derived ordinal supervision, aiming to improve the correlation between confidence scores and actual prediction errors. Our approach is incorporated into MonSter, an advanced stereo baseline, achieving SOTA performance in confidence estimation across KITTI 2012, KITTI 2015 and Middlebury benchmarks.**

I. INTRODUCTION

Stereo matching estimates dense disparity maps from rectified image pairs, serving as a fundamental technique in computer vision applications such as 3D reconstruction, autonomous driving, and robotic perception [1]. While modern stereo networks achieve remarkable accuracy, they often struggle to provide reliable confidence estimates. Existing approaches primarily rely on probabilistic modeling or statistical aggregation to capture aleatoric uncertainty, which reflects noise inherent in the observations. However, they largely overlook epistemic uncertainty, stemming from parameter ambiguity and training instability [2]. As a result, these models tend to produce overconfident predictions, especially in the most error-prone regions, such as occlusion boundaries, textureless surfaces, and reflective areas. This miscalibration between predictive confidence and actual accuracy poses significant risks to depth perception in safety-critical applications. To address this, there is an urgent need for confidence estimation mechanisms that reflect both epistemic uncertainty and inconsistencies arising from training dynamics without compromising the accuracy of the underlying disparity predictions.

*Corresponding author: Zaiwang Gu and Jun Cheng (guzw, cheng_jun@a-star.edu.sg).

[1]S. Lu and X. Jiang are with the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore.

[2] Z. Gu and J. Cheng are with the Institute for Infocomm Research (I^{2}R), Agency for Science, Technology and Research (A*STAR), Singapore.

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

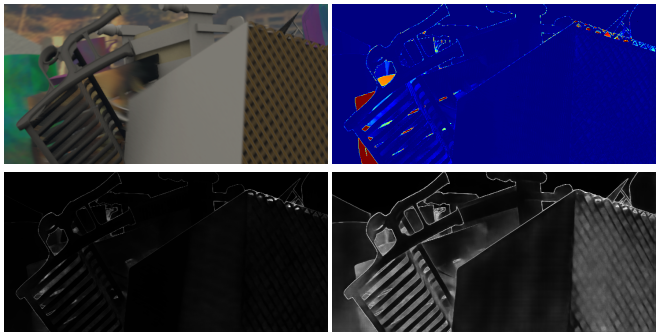

Fig. 1. From left to right, top to bottom: input left image, ours error map, MoNIG epistemic uncertainty map, and ours epistemic uncertainty map. Our method delivers more stable and trustworthy confidence estimation than MoNIG, without sacrificing disparity accuracy.

Advances in uncertainty estimation have driven the development of various methods for quantifying predictive confidence. For instance, Bayes by Backprop [3] achieves principled uncertainty estimation through approximate Bayesian posterior inference. Deep Ensembles [4] train multiple independently initialized networks and aggregate their outputs. These approaches typically entail considerable computational cost and increased inference latency due to repeated sampling procedures and ensemble evaluation. Evidential deep learning [5] places a Dirichlet prior over discrete classification predictions, while deep evidential regression [2] extends this framework to continuous regression tasks. Current frameworks predominantly focus on data-centric uncertainty quantification, while paying insufficient attention to the information embedded in training dynamics. To address this limitation, Moon *et al.* [6] and Li *et al.* [7] estimate prediction confidence by exploiting training dynamics, either through consistency across epochs or correctness frequency.

Nevertheless, these methods remain confined to classification tasks, and naively extending them to dense regression problems such as stereo matching entails several fundamental challenges. First, dense regression produces continuously fluctuating outputs that hinder consistency enforcement through exact matching. Meanwhile, the predicted class is determined by the highest probability in classification, and this probabilistic output itself serves as a measure of epistemic uncertainty [8]. Dense regression outputs lack an explicit uncertainty representation that can be directly aligned with consistency-based supervision. Furthermore, computing consistency signals across training epochs becomes computationally demanding, as dense prediction requires maintaining pixel-wise histories over large-scale inputs. The use of data augmentations such as random cropping further complicates

this process by disrupting spatial alignment and obfuscating the correspondence between predicted confidence scores and their associated consistency-based supervision.

Motivated by these limitations, we introduce an epoch-wise consistency accumulation framework that explicitly embeds training dynamics into the confidence estimation process for stereo matching. Specifically, we design a full-resolution cross-epoch prediction mechanism that captures pixel-wise training dynamics by aligning disparity predictions of the same sample across successive epochs, and we compute inter-epoch deviation at each pixel to derive training consistency signals that reflect the model's evolving confidence during optimization. We adopt deep evidential regression to model disparity uncertainty, wherein each pixel-wise prediction is represented by a Normal Inverse-Gamma (NIG) distribution. To improve robustness and generality, we aggregate evidential signals extracted from heterogeneous geometric pathways via intra-distribution fusion. The resulting epistemic component constitutes a principled uncertainty basis for quantifying model confidence. To supervise uncertainty estimation, we introduce a consistency-ranked evidential discrepancy loss that encourages alignment between evidential confidence outputs and training-derived consistency signals, enforcing ordinal agreement across pixel-wise predictions. Extensive experiments demonstrate that the proposed framework integrates naturally with MonSter [9] , a high-performing stereo architecture, and consistently achieves stable and reliable confidence estimation across standard benchmarks, as shown in Fig. 1.

Our main contributions can be summarized as follows:

- We propose an epoch-wise consistency accumulation framework that injects training dynamics into stereo confidence estimation via cross-epoch alignment.
- The proposed method addresses the difficulty of constructing reliable supervision in dense regression by capturing inter-epoch prediction coherence without relying on ground-truth uncertainty labels.
- We design a consistency-ranked evidential discrepancy loss to align model-predicted uncertainty with training-derived consistency supervision.
- Extensive experiments validate the method's strong generalizability and its uncertainty estimates demonstrate a strong alignment with disparity prediction errors across mainstream stereo benchmarks.

II. Related Works

A. Stereo Matching

Over the past decades, stereo matching has witnessed remarkable advancements, propelled by the integration of deep learning techniques [1], [10], [11]. Early CNN-based frameworks such as GC-Net [12] pioneered 3D cost volume regularization, followed by multi-scale and context-aware designs including PSMNet [13] and AANet [14]. Subsequent improvements focused on enhancing matching robustness and efficiency through group-wise correlation and cascaded cost volumes, as in GwcNet [15] and CFNet [16], while SOMNet [17] further combined 2D and 3D cues to refine depth estimation, particularly around object boundaries. Complementary efforts addressed stereo failure modes and alignment strategies, including the orthogonal stereo setups in [18] and Stereoscopic Cross Attention (SCA) [19]. Later developments explored feature aggregation and geometric alignment via pyramid composition and cost warping [20], convolution-attention hybrid modeling [21], and recurrent refinement mechanisms such as RAFT-Stereo [22] and CREStereo [23]. Implicit geometric encoding and texture-aware enhancement were introduced by IGEV [24] and SelectStereo [25], respectively. Transformer-based architectures further advanced global reasoning without explicit cost volumes [26], extended to multi-modal fusion [27] and temporal modeling [28], while GOAT [29] strengthened occlusion reasoning through global attention.

Recently, foundational models have started to be employed in stereo matching. MonSter [9] introduces a dual-branch architecture to incorporate foundation priors into stereo estimation. DEFOM-Stereo [30] embeds foundation depth cues within a recurrent refinement pipeline with scale-aware updates. FoundationStereo [31] further improves cross-domain generalization by mitigating the sim-to-real gap. Stereo Anywhere [32] focuses on geometry-aware fusion mechanisms. In this paper, we adopt MonSter as the baseline architecture.

B. Uncertainty Estimation

As stereo matching moves toward real-world deployment, especially in high-stakes domains, quantifying prediction uncertainty has become indispensable [33]. Kendall *et al.* [34] considered both epistemic and aleatoric uncertainty, employing a unified Bayesian deep learning framework to learn distributions over weights. To address the prohibitive computational overhead of exact inference in large parameter spaces, Monte Carlo Dropout [35] uses dropout to approximate variational inference. By leveraging independently trained networks with diverse initializations, Deep Ensembles [4] achieve reliable uncertainty estimates through the variance among predictions. Evidential approaches [5], in contrast, place a Dirichlet distribution over classification outputs to jointly capture aleatoric and epistemic uncertainty, and have later been extended to regression [2] by inferring the parameters of a NIG distribution. Building upon unimodal evidential regression, the Mixture of NIG framework [36] further extends uncertainty modeling to multimodal scenarios by capturing both modality-specific and global uncertainty.

Beyond general tasks, uncertainty estimation has also been explored in the context of stereo matching. LAF-Net [37] introduces attention maps and scale-aware sub-networks to adaptively fuse multi-modal information. SEDNet [38] employs a differentiable soft histogram technique and incorporates KL divergence to align the predicted uncertainty distribution with the actual error distribution. ELFNet [39] adopted evidence theory to design a Transformer-based local-global fusion network. Disparity Plane Sweep [40] compares shifted-image disparity profiles with ideal ones to estimate

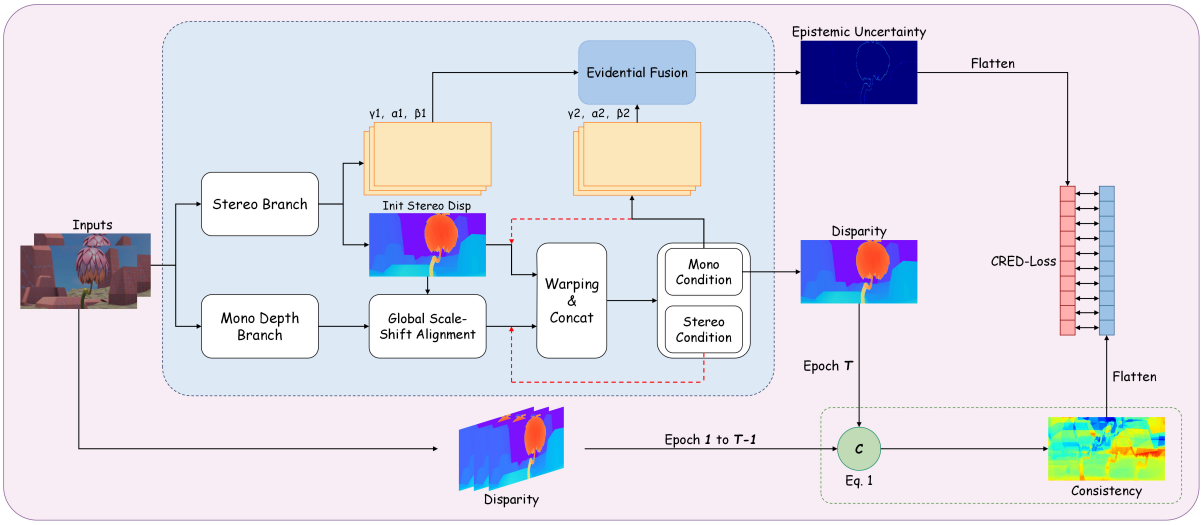

Fig. 2. Network architecture of the proposed method. The light blue box represents the MonSter backbone, with the addition of the Evidential module. The pink box illustrates the cross-epoch accumulation of training consistency. The consistency computation branch, enclosed by the green dashed line, is used during training only and discarded during testing.

uncertainty externally, without relying on internal network features. Although trained end-to-end, these methods neglect to extract supervisory signals from the training process itself. To address this, CAL [6] introduces a correctness-aware objective that supervises confidence using pseudo labels indicating prediction validity, while TC [7] introduces training consistency between predictions and targets during training. However, these approaches remain limited to classification, whereas our method extends training consistency to the dense regression setting of stereo matching.

III. Method

We propose an epoch-wise accumulation framework to inject training dynamics into confidence estimation in stereo matching. Our method consists of three main components: an epoch-wise consistency accumulation module, an intra-evidence fusion module based on the cost volume, and a consistency ranking loss module, as shown in Fig. 2. The epoch-wise consistency accumulation module accumulates per-pixel consistency supervision labels throughout training. The intra-evidential fusion module predicts the hyperparameters $\{\delta, \gamma, \alpha, \beta\}$ of an uncertainty-aware distribution [5], [2], [36], and leverages the derived epistemic uncertainty as a confidence measure to fit the consistency supervision labels through a ranking-based loss.

A. Training Consistency for Stereo Matching

To enable consistency computation under the continuous and non-repetitive nature of regression outputs, we formulate a relaxed consistency metric tailored for stereo matching. As shown in Eq. (1), the overall consistency score c across the entire image is computed over T epochs as follows:

$$c = \text{Norm} \sum_{t=1}^{T-1} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \max \left\{ 0,\ 1 - \frac{\left| \hat{d}_{i,j}^{t} - \hat{d}_{i,j}^{t+1} \right|}{\hat{d}_{i,j}^{t}} \right\}, \tag{1}$$

where $\hat{d}_{i,j}^{t}$ denotes the predicted disparity at pixel location (i, j) in the t-th training epoch.

By formulating consistency based on the normalized difference between successive predictions, our method captures how close the model's output remains across epochs relative to its own magnitude. This proportion-aware design not only accommodates the scale variability of disparity values across different regions but also avoids imposing hard thresholds, enabling a smoother and more adaptive consistency signal tailored for dense regression.

B. Consistency-Aware Confidence Estimation

Although the above consistency measures how the model converges during the training, it cannot be computed at test time. To make use of consistency during testing, we include a branch for confidence estimation while using consistency to regularize the training of confidence estimation. More specifically, we use deep evidential learning as our basic confidence estimation and minimize the difference between model uncertainty and consistency.

1) Evidential Uncertainty: We adopt deep evidential regression for its probabilistic framework that jointly models predictions and uncertainty quantification [5], [2], [36], [39]. Unlike conventional approaches that treat confidence estimation as an unconstrained auxiliary task, evidential deep learning formulates uncertainty modeling as a process of accumulating statistical evidence over likelihood functions. The framework is implemented by assuming that the observed disparity values d are drawn from a Gaussian distribution with unknown mean and variance (μ, σ^2), where the mean μ follows a Gaussian prior and the variance σ^2 follows an Inverse Gamma prior [41]:

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

where $\Gamma(\cdot)$ denotes the Gamma function, and $\mathbf{m} = (\delta, \gamma, \alpha, \beta)$ represents the set of hyperparameters.

To jointly characterize the disparity prediction and its associated uncertainty, the task is formulated as inferring a posterior distribution $q(\mu, \sigma^2) = p(\mu, \sigma^2 \mid d)$. To simplify the inference, the posterior is approximated using a NIG distribution. As derived in Eq. (3), the total evidence $\Phi = 2\gamma + \alpha$ quantifies the model's prediction confidence, while the hyperparameters naturally disentangle aleatoric and epistemic uncertainty[5], [2], [36]:

$$\underbrace{\mathbb{E}[\mu]}_{\text{prediction}} = \gamma, \quad \underbrace{\mathbb{E}[\sigma^2]}_{\text{aleatoric}} = \frac{\beta}{\alpha - 1}, \quad \underbrace{\mathrm{Var}[\mu]}_{\text{epistemic}} = \frac{\beta}{\gamma(\alpha - 1)}. \tag{3}$$

This framework ensures mathematically grounded confidence estimates that align with actual prediction reliability during training.

2) Uncertainty-Aware NIG Fusion: In the initial disparity estimation stage, the MonSter model constructs a group-wise correlation cost volume at $\frac{1}{4}$ resolution to effectively capture geometric relationships from binocular inputs, encoding the structural and semantic information necessary for disparity estimation. This 3D volume is then processed by a cost aggregation module to generate a compact geometric representation, serving as the foundation for subsequent evidential modeling. On top of this, we introduce four lightweight 3D convolutional regression heads [12], [42], [43] to predict the hyperparameters $(\delta_b, \gamma_b, \alpha_b, \beta_b)$ of a binocular-guided NIG distribution, enabling joint modeling of disparity and its associated uncertainty.

During the iterative refinement stage, the model dynamically adjusts disparity predictions and their uncertainties through interactions between monocular and stereo features. Specifically, MonSter leverages deformable alignment to fuse monocular depth priors with stereo residual cues, enabling fine-grained modeling of local geometric deviations. A gating mechanism is further introduced to suppress unreliable features. The fused representations are then fed into the residual evidential modeling module [9], where we extend the module's architecture by adding additional convolutional branches to predict the monocular-guided hyperparameters of a NIG distribution $(\delta_m, \gamma_m, \alpha_m, \beta_m)$.

We introduce the NIG summation operator to achieve a mixture of NIG distributions, which fuses two NIG distributions into a unified representation while preserving their evidential structure. The summation operation is defined as follows [36], [39]:

$$\mathrm{NIG}(\delta, \gamma, \alpha, \beta) \triangleq \mathrm{NIG}(\delta_b, \gamma_b, \alpha_b, \beta_b) \oplus \mathrm{NIG}(\delta_m, \gamma_m, \alpha_m, \beta_m), \tag{4}$$

where

$$\begin{aligned} \delta &= (\gamma_b + \gamma_m)^{-1}(\gamma_b \delta_b + \gamma_m \delta_m), \\ \alpha &= \alpha_b + \alpha_m + \frac{1}{2}, \quad \gamma = \gamma_b + \gamma_m, \\ \beta &= \beta_b + \beta_m + \frac{1}{2}\gamma_b(\delta_b - \delta)^2 + \frac{1}{2}\gamma_m(\delta_m - \delta)^2. \end{aligned} \tag{5}$$

In this fusion mechanism, the parameter γ_m represents the confidence of the m-th modality in its predicted mean δ_m. The parameter β consists of two parts: the sum of β_b and β_m from multiple modalities, and the variance between the final prediction and each individual modality's prediction. Overall, the fusion strategy effectively integrates modality-specific uncertainty and the prediction deviation among different modalities, enabling a comprehensive representation of the final uncertainty.

We train our model by minimizing the difference between uncertainty from the above mixture of NIG distribution and the consistency such that the consistency knowledge can be captured in the trained models. At test time, the consistency part enclosed in the dash line is discarded.

C. Loss

1) Consistency-Ranked Evidential Discrepancy Loss: A fundamental assumption is that pixels exhibiting higher training consistency should be assigned correspondingly higher confidence scores. However, directly regressing raw consistency values may lead to optimization instability and cause the model to overfit to the absolute scale of the supervision signal, rather than capturing the relative structure that reflects true confidence ordering. Instead, we design the consistency-ranked evidential discrepancy loss, which encourages the relative ordering between training consistency and evidential uncertainty to align. This objective is formally expressed in Eq. (6) as follows:

$$\mathcal{L}_{\mathrm{C}} = \sum_{(p,q)\in\mathcal{P}} \max\{0,\ -\mathrm{sgn}(s_q - s_p)\cdot(c_q - c_p) + |s_q - s_p|\}, \tag{6}$$

where $\mathcal{P}$ denotes the set of sampled pixel pairs (p, q) within the image. For each pixel p, s_p represents the predicted evidential uncertainty score, and c_p denotes the corresponding consistency supervision label. The function $\mathrm{sgn}(\cdot)$ refers to the sign function.

The loss enforces alignment between the ordering of consistency labels and predicted confidence scores. Specifically, when $c_p < c_q$, the model is encouraged to produce confidence estimates that satisfy $s_p < s_q$. A positive penalty is incurred whenever the confidence ranking violates the expected consistency, and it is further amplified in proportion to the magnitude of the confidence gap.

2) Evidential Uncertainty Loss: We define the loss function $\mathcal{L}_i^{\mathrm{N}}(\mathbf{w})$ at the i-th pixel as the negative logarithm of model evidence, and optimize the model parameters $\mathbf{w}$ by minimizing this negative log-likelihood loss during training:

$$\begin{aligned} \mathcal{L}_i^{\mathrm{N}}(\mathbf{w}) &= \frac{1}{2}\log\left(\frac{\pi}{\gamma}\right) - \alpha\log(\Omega) + \\ &(\alpha + \tfrac{1}{2})\log\left((d_i - \delta)^2\gamma + \Omega\right) + \log\left(\frac{\Gamma(\alpha)}{\Gamma\left(\alpha + \frac{1}{2}\right)}\right), \end{aligned} \tag{7}$$

where $\Omega = 2\beta(1+\gamma)$, and d_i denotes the predicted disparity at the i-th pixel.

Meanwhile, to impose stronger penalties on incorrect evidence in stereo matching, we also adopt an evidence

regularizer $\mathcal{L}_i^{\mathrm{R}}(\mathbf{w})$, scaled proportionally to the disparity prediction error,

$$\mathcal{L}_i^{\mathrm{R}}(\mathbf{w}) = \left|d_i^{gt} - \mathbb{E}(\mu_i)\right| \cdot \Phi = \left|d_i^{gt} - \delta\right| \cdot (2\gamma + \alpha), \quad (8)$$

where d_i^{gt} is the ground truth disparity at pixel i.

Lou *et al.* [39] define the total uncertainty loss for dense stereo matching as the expected loss over all pixels, comprising a likelihood maximization term and a regularization term for evidential learning,

$$\mathcal{L}_{\mathrm{U}}(\mathbf{w}) = \frac{1}{N} \sum_{i=0}^{N-1} \left(\mathcal{L}_i^{\mathrm{N}}(\mathbf{w}) + \tau \mathcal{L}_i^{\mathrm{R}}(\mathbf{w})\right), \quad (9)$$

where N denotes the total number of pixels, and the regularization term uses the coefficient τ to balance uncertainty inflation and model fit.

3) Total Loss: The total loss $\mathcal{L}$ is formulated as a weighted combination of disparity regression loss $\mathcal{L}_D$, evidential uncertainty loss $\mathcal{L}_U$, and training consistency-guided ranking loss $\mathcal{L}_C$. Here, $\mathcal{L}_D$ corresponds to the total loss formulation adopted in MonSter,

$$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_U + \lambda \cdot \mathcal{L}_C \quad (10)$$

Considering that the consistency-guided ranking loss has a smaller magnitude compared to other loss terms, we introduce a weighting factor λ to incorporate it as a secondary regulatory component in the model optimization.

IV. Experiments

A. Experimental Settings

In this section, we describe the experimental setup and evaluation protocol in detail. We comprehensively evaluate our approach on several widely-used stereo benchmarks. The model is pre-trained on SceneFlow [44] and directly evaluated in a zero-shot manner on the remaining datasets, demonstrating its generalizability and robustness across diverse real-world scenarios.

1) Datasets: The SceneFlow dataset, a synthetic benchmark for optical flow and disparity, contains 35,454 training and 4,370 testing 960×540-pixel stereo image pairs, and serves as the primary dataset for our model's pre-training and performance evaluation. The KITTI 2012 & 2015 datasets [45] are pivotal real-world benchmarks from autonomous driving scenarios. Our experiments utilize the 194 training stereo image pairs with ground truth from KITTI 2012 and 200 from KITTI 2015 for real-world zero-shot generalization. Additionally, for evaluating zero-shot generalization in indoor scenes, our experiments employ 15 images at $\frac{1}{4}$ resolution from the Middlebury 2014 [46], a well-known indoor benchmark offering high-resolution stereo pairs.

2) Implementation Details: Our implementation leverages a PyTorch-based MonSter architecture. For optimization, we employ the AdamW optimizer coupled with a OneCycleLR learning rate schedule [47], initialized at 2e-4. Training is conducted on the SceneFlow dataset, where input images for disparity estimation are cropped to 736×320 following MonSter's preprocessing protocol. Consistency is computed at the full 960×540 resolution, as random cropping would otherwise invalidate consistency comparison between adjacent epochs. Initially, we train the model from scratch for 10 epochs with a batch size of 8 to accumulate training consistency. Subsequently, we load the pretrained MonSter backbone and resume training for 100k iterations, integrating the accumulated consistency to obtain the final pretrained weights.

TABLE I

Comparison with state-of-the-art. Bold shows the best results and underline shows the second best. The ↑ indicates the higher the better and the ↓ indicates the smaller the better. EPE is a tracking indicator.

Method	SceneFlow			
	EPE	Pearson↑	AUC_e ↓	n-AUSE↓
SEDNet	0.387	0.179	0.114	3.260
Ensemble	0.383	0.531	1.098	41.116
Bayesian	0.378	0.117	0.088	2.374
Evidential	0.379	<u>0.641</u>	<u>0.038</u>	<u>0.505</u>
Ours	0.379	**0.691**	**0.037**	**0.403**

3) Evaluation Protocol: Our approach is primarily focused on estimating predictive uncertainty, with the goal of capturing the model's confidence in each disparity output. We employ the Pearson correlation coefficient [48] to assess the consistency between predicted disparities and the associated confidence scores. Specifically, it measures the linear correlation between the absolute disparity error and the predicted confidence across all valid pixels, indicating how well the confidence estimates align with actual prediction errors. In addition to correlation-based measures, we further adopt AUC_e [49], [50] and its normalized variant n-AUSE [51], [52] to evaluate the quality of confidence estimation from a ranking perspective:

$$\mathrm{AUC_e} = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{r_t N} \sum_{i=1}^{r_t N} e_{(i)}^{\mathrm{conf}} \right) \quad (11)$$

AUC_e reflects the model's ability to prioritize reliable predictions by progressively removing pixels with the highest predicted confidence, while n-AUSE normalizes the gap between AUC_e and the AUC_{opt} derived from ground-truth error rankings to mitigate the influence of disparity accuracy. n-AUSE is computed as follows:

$$\text{n-AUSE} = \frac{\mathrm{AUC}_e - \mathrm{AUC}_{opt}}{\mathrm{AUC}_{opt}} \quad (12)$$

In addition, we adopt the EPE as a tracking indicator, serving as a standardized reference to align with the backbone performance to ensure that our uncertainty-aware framework does not compromise disparity estimation quality.

B. Comparison with State-of-the-art

We adopt the MonSter architecture as our backbone network, following its original training configuration. To comprehensively evaluate our approach, we compare it against

TABLE II

ZERO-SHOT GENERALIZATION PERFORMANCE OF SCENEFLOW TRAINED MODEL IN KITTI AND MIDDLEBURY DATASETS. METHODS ABOVE THE HORIZONTAL LINE USE DIFFERENT BACKBONE ARCHITECTURES.

Method	KITTI 2012			KITTI 2015			Middlebury		
	Pearson↑	AUC_e ↓	n-AUSE↓	Pearson↑	AUC_e ↓	n-AUSE↓	Pearson↑	AUC_e ↓	n-AUSE↓
ELFNet	0.291	1.071	1.019	0.277	0.916	1.189	0.183	0.777	1.092
DPSNet	/	/	/	0.336	1.344	0.955	0.370	1.754	0.938
SEDNet	0.182	0.514	1.835	0.144	0.817	1.877	0.228	0.267	1.865
Ensemble	0.402	1.438	8.147	0.333	1.488	4.816	0.440	1.182	11.997
Bayesian	0.079	0.497	2.170	0.007	0.809	2.168	0.044	0.375	3.115
Evidential	0.474	0.342	0.920	0.456	0.497	**0.771**	0.567	**0.160**	**0.642**
Ours	**0.512**	**0.335**	**0.886**	**0.531**	**0.494**	0.788	**0.635**	0.163	0.677

TABLE III

ABLATION STUDIES TO SHOW EFFECTIVENESS OF THE COMPONENTS IN THE PROPOSED METHOD.

Method	EPE	Pearson↑	AUC_e ↓	n-AUSE↓
Baseline	0.378	-	-	-
+Uncertainty	0.384	0.028	0.604	21.170
+Evidential	0.379	0.641	0.038	0.505
+Consistency	0.379	**0.691**	**0.037**	**0.404**

TABLE IV

ABLATION STUDIES ON THE WEIGHTS OF THE PROPOSED CRED LOSS, TRAINED FOR 40,000 STEPS ON THE SCENE FLOW DATASET.

	EPE	Pearson↑	AUC_e ↓	n-AUSE↓
$\lambda = 0.05$	0.426	0.675	0.051	0.450
$\lambda = 0.08$	0.415	0.670	0.043	**0.413**
$\lambda = 0.10$	**0.398**	**0.680**	**0.041**	0.417
$\lambda = 0.12$	0.425	0.675	0.050	0.442

several strong and representative baselines. For methods that can be adapted, such as Ensemble, Bayesian [53], [54], [55], Evidential, and SEDNet, we re-implement them using the MonSter backbone to ensure fair and consistent comparisons.

After completing the training process described in the implementation details, we evaluate our model on the Scene Flow test set to assess its generalization ability. Through empirical analysis, we set the weighting coefficient of the proposed consistency loss $\mathcal{L}_C$ to $\lambda = 0.1$, which strikes a favorable balance between disparity prediction and uncertainty estimation. Details of the ablation studies are given in Section IV-D. To validate the effectiveness of our method, we conduct comprehensive comparisons with several strong and representative baselines under consistent evaluation settings, including SEDNet [38], Ensemble [4], Bayesian [54], and Evidential [5], [2], [36]. To achieve a fair comparison, we implement the confidence estimation from these methods on MonSter baseline.

As shown in Table I, our method surpasses the previously best-performing approach in confidence estimation on the Scene Flow test set while maintaining the similar performance in EPE. Specifically, compared to the previous Evidential approach, our approach improves Pearson correlation by 7.8% from 0.641 to 0.691 and reduces n-AUSE by 20.2% from 0.505 to 0.403, demonstrating that our predicted confidence scores align more precisely with the spatial distribution of disparity errors. Unlike the over-smoothed patterns from Ensemble methods and the unstable confidence of Bayesian models, our approach guided by consistency accumulation captures error structures more faithfully by emphasizing regions of persistent prediction disagreement. We also conducted a statistical t-test and the results show that the improvement is significant with $p < 0.001$.

C. Zero-shot Generalization

While training and testing on datasets from same domain shows the performance in one aspect, the zero-shot performance in datasets from different domain is also important. We further show the performance of the model trained from the synthetic Scene Flow dataset in confident estimation without any additional fine-tuning across multiple standard benchmarks, including KITTI 2012, KITTI 2015 and Middleburry. Besides the methods in Section IV-B, we also compare with ELFNet and DPSNet, following their original training protocols to preserve the integrity of their reported performance. Table II summarizes the results.

Our method improves Pearson correlation coefficient compared with the state-of-the-art evidential approach by 8.0%, 16.4% and 12.0% relatively in KITTI 2012, KITTI 2015, and Middleburry respectively. Meanwhile, it reduces AUC_e and n-AUSE in most scenarios. Fig. 3 shows a few examples to compare the confidence estimation between our method and state-of-the-art methods. As shown, our proposed method shows more accurate confidence of the predictions.

D. Ablation Studies

To systematically evaluate the contribution and necessity of each proposed component, we conduct an ablation study by incrementally integrating evidential uncertainty (denoted as +Uncertainty), evidential fusion (denoted as +Evidential), and consistency (denoted as +Consistency) into the baseline model. Results presented in Table III clearly indicate that each module is indispensable, jointly contributing to enhanced confidence estimation performance. We observe a low correlation in adapting evidential uncertainty in a single branch, due to the lack of uncertainty estimation in the other branch. By including uncertainty estimation in both

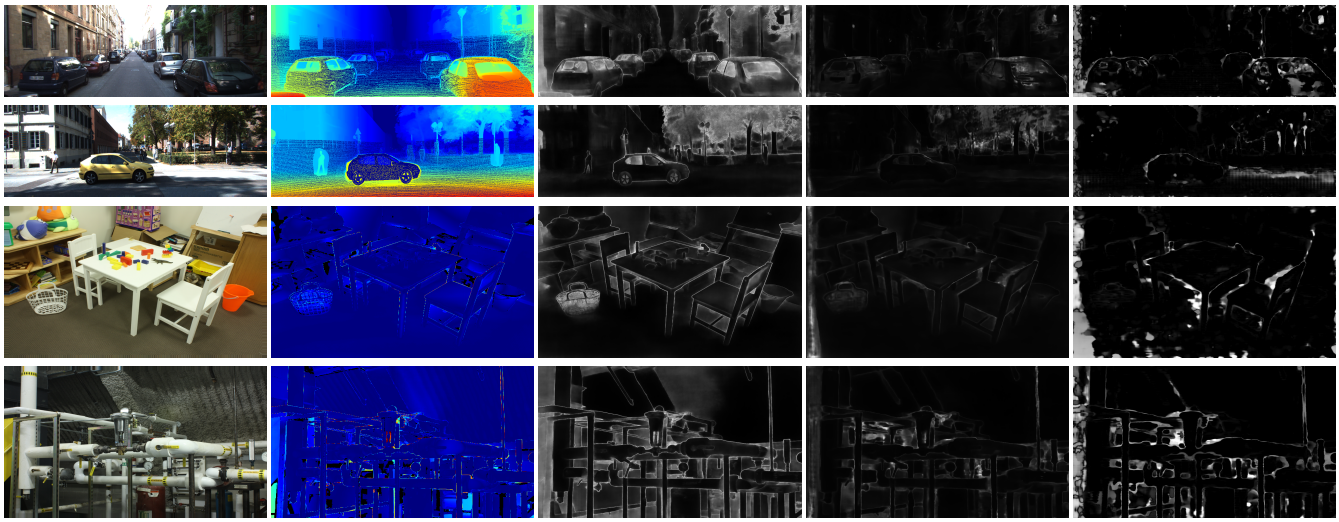

Fig. 3. Visualization of confidence estimation by our methods and its comparison with Evidential and SEDNet. Rows 1–2 present results on KITTI 2012 and KITTI 2015, and Rows 3–4 correspond to Middlebury.

branches via evidential fusion, we observe an improvement in Pearson correlation. The further introduction of the proposed consistency leads to an improvement of 7.8% relatively while the n-AUSE is reduced by 20.0%. It shall be noted that we have also tested our model with different λ from 0.06 to 0.12 with a step of 0.02. The results after 40K iteration is shown in Table IV. The results showed best performance with $\lambda = 0.1$. We set $\lambda = 0.1$ in the rest of our experiments.

E. Limitations

Despite these promising results, several limitations remain. The Gaussian assumption in the evidential formulation may not fully capture complex disparity distributions. The optimization effects of early-epoch consistency supervision require further study. The computational and memory overhead during model execution, as well as adaptation beyond MonSter to diverse mono-stereo architectures, remain to be explored.

V. Conclusion

Accurate and reliable confidence estimation is a cornerstone of stereo matching, particularly in safety-critical applications where erroneous depth perception can have severe consequences. Despite recent advances, most existing methods still fail to effectively exploit training dynamics, leaving a gap in their ability to capture and represent uncertainty in a principled way. To bridge this gap, this paper introduces an epoch-wise consistency accumulation framework that explicitly incorporates training dynamics into the uncertainty estimation process. The key idea is to leverage training consistency signals and combine them with a deep evidential intra-distribution fusion strategy, which together enable a more reliable modeling of predictive uncertainty. Extensive experiments conducted on multiple benchmark datasets demonstrate that our framework exhibits strong cross-dataset generalization. Most importantly, it consistently achieves stable and trustworthy confidence estimation without compromising disparity accuracy, highlighting its potential for deployment in real-world, safety-critical systems.

References

[1] F. Tosi, L. Bartolomei, and M. Poggi, "A survey on deep stereo matching in the twenties," *International Journal of Computer Vision*, vol. 133, no. 7, pp. 4245–4276, 2025.

[2] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 927–14 937, 2020.

[3] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1613–1622.

[4] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[5] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[6] J. Moon, J. Kim, Y. Shin, and S. Hwang, "Confidence-aware learning for deep neural networks," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7034–7044.

[7] C. Li, X. Hu, and C. Chen, "Confidence estimation using unlabeled data," in *International Conference on Learning Representations*, 2023.

[8] J. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," *Advances in Neural Information Processing Systems*, vol. 2, 1989.

[9] J. Cheng, L. Liu, G. Xu, X. Wang, Z. Zhang, Y. Deng, J. Zang, Y. Chen, Z. Cai, and X. Yang, "Monster: Marry monodepth to stereo unleashes power," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[10] R. Gong, K.-H. Yap, W. Liu, X. Yang, and J. Cheng, "Rectification-specific supervision and constrained estimator for online stereo rectification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 22 348–22 358.

[11] J. Cheng, Z. Gu, W. Liu, J. Fan, Z. Li, and C.-S. Foo, "Srnet: Self-supervised structure regularization for stereo matching," *Neurocomputing*, vol. 661, p. 131907, 2026.

[12] 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 *IEEE/CVF International Conference on Computer Vision*, 2017, pp. 66–75.

[13] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[14] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1959–1968.

[15] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3273–3282.

[16] Z. Shen, Y. Dai, and Z. Rao, "Cfnet: Cascade and fused cost volume for robust stereo matching," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 906–13 915.

[17] J. Choe, K. Joo, F. Rameau, and I. So Kweon, "Stereo object matching network," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 12 918–12 924.

[18] L. Meier, D. Honegger, V. Vilhjalmsson, and M. Pollefeys, "Real-time stereo matching failure prediction and resolution using orthogonal stereo setups," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 5638–5643.

[19] H. Sakuma and Y. Konishi, "Geometry-aware unsupervised domain adaptation for stereo matching," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 6116–6123.

[20] 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 *European Conference on Computer Vision*, 2022, pp. 280–297.

[21] J. Cheng, G. Xu, P. Guo, and X. Yang, "Coatrsnet: Fully exploiting convolution and attention for stereo matching by region separation," *International Journal of Computer Vision*, vol. 132, no. 1, pp. 56–73, 2024.

[22] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 218–227.

[23] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 263–16 272.

[24] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 919–21 928.

[25] X. Wang, G. Xu, H. Jia, and X. Yang, "Selective-stereo: Adaptive frequency information selection for stereo matching," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[26] 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 *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6197–6206.

[27] Q. Su and S. Ji, "Chitransformer: Towards reliable stereo from cues," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1939–1949.

[28] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht, "Dynamicstereo: Consistent dynamic depth from stereo videos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 229–13 239.

[29] Z. Liu, Y. Li, and M. Okutomi, "Global occlusion-aware transformer for robust stereo matching," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 3535–3544.

[30] H. Jiang, Z. Lou, L. Ding, R. Xu, M. Tan, W. Jiang, and R. Huang, "Defom-stereo: Depth foundation model based stereo matching," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 21 857–21 867.

[31] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 5249–5260.

[32] L. Bartolomei, F. Tosi, M. Poggi, and S. Mattoccia, "Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 1013–1027.

[33] M. Poggi, S. Kim, F. Tosi, S. Kim, F. Aleotti, D. Min, K. Sohn, and S. Mattoccia, "On the confidence of stereo matching in a deep-learning era: a quantitative evaluation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5293–5313, 2021.

[34] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in Neural Information Processing Systems*, 2017.

[35] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1050–1059.

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

[37] S. Kim, S. Kim, D. Min, and K. Sohn, "Laf-net: Locally adaptive fusion networks for stereo confidence estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 205–214.

[38] L. Chen, W. Wang, and P. Mordohai, "Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 235–17 244.

[39] J. Lou, W. Liu, Z. Chen, F. Liu, and J. Cheng, "Elfnet: Evidential local-global fusion for stereo matching," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 784–17 793.

[40] J. Y. Lee, W. Ka, J. Choi, and J. Kim, "Modeling stereo-confidence out of the end-to-end stereo-matching network via disparity plane sweep," in *AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 2901–2910.

[41] J. E. Griffin and P. J. Brown, "Inference with normal-gamma prior distributions in regression problems," *Bayesian Analysis*, vol. 5, no. 1, pp. 171–188, 2010.

[42] L. Yu, Y. Wang, Y. Wu, and Y. Jia, "Deep stereo matching with explicit cost aggregation sub-architecture," in *AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[43] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2495–2504.

[44] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040–4048.

[45] M. Menze, C. Heipke, and A. Geiger, "Joint 3d estimation of vehicles and scene flow," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 2, pp. 427–434, 2015.

[46] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German Conference on Pattern Recognition*. Springer, 2014, pp. 31–42.

[47] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, 2019, pp. 369–386.

[48] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*. Springer, 2009, pp. 1–4.

[49] L. E. Dodd and M. S. Pepe, "Partial auc estimation and regression," *Biometrics*, vol. 59, no. 3, pp. 614–623, 2003.

[50] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, 2005.

[51] F. K. Gustafsson, M. Danelljan, and T. B. Schon, "Evaluating scalable bayesian deep learning methods for robust computer vision," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 318–319.

[52] S. K. Lind, Z. Xiong, P.-E. Forssen, and V. Krüger, "Uncertainty quantification metrics for deep regression," *Pattern Recognition Letters*, vol. 186, pp. 91–97, 2024.

[53] R. Van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, *et al.*, "Bayesian statistics and modelling," *Nature Reviews Methods Primers*, vol. 1, no. 1, p. 1, 2021.

[54] U. Upadhyay, S. Karthik, Y. Chen, M. Mancini, and Z. Akata, "Bayescap: Bayesian identity cap for calibrated uncertainty in frozen neural networks," in *European Conference on Computer Vision*, 2022.

[55] U. Upadhyay, Y. Chen, and Z. Akata, "Robustness via uncertainty-aware cycle consistency," in *Advances in Neural Information Processing Systems*, 2021.