Uncertainty-Guided Iterative Architecture for Stereo Matching

1st Weiqing Xiao
Electronics and Information Engineering
Beihang University
Beijing, China
xiaowqtx@buaa.edu.cn

2nd Fengjun Zhong
Electronics and Information Engineering
Beihang University
Beijing, China
by2402235@buaa.edu.cn

3rd Hao Zhao
Institute for AI Industry Research
Tsinghua University
Beijing, China
zhaohao@air.tsinghua.edu.cn

***Abstract*—Iteration-based methods have achieved impressive achievements in stereo matching tasks. However, due to the lack of disparity reliability judgment, existing iteration-based methods have difficulties obtaining reliable iterative results in a stable manner. In this paper, we propose a novel uncertainty-guided iterative architecture. Specifically, we first design a lightweight Cost volume-based disparity Uncertainty Estimation method (CUE), which efficiently accomplishes multiple inference during the iteration process. Subsequently, we propose Uncertainty-based Disparity update Control (UDC), which provides update magnitudes for different pixels that match the current disparity reliability. Additionally, we propose Uncertainty-based Disparity Rectification (UDR) for burden-free optimization of the initial disparity. Experiments on SceneFlow, KITTI, Middlebury 2014, and ETH3D demonstrate that the proposed method effectively improves the accuracy and efficiency of existing iterative methods. Particularly, when running only 15 updates, our method outperforms the baseline IGEV-Stereo that runs 32 updates, which saves around 40% of the inference time.**

***Index Terms*—Stereo Matching, Iteration-based Method, Disparity Uncertainty**

I. Introduction

Depth perception is the basis for computer vision and graphics research in 3D scenes. Stereo matching, as an efficient and low-cost depth estimation method, aims to estimate the pixel horizontal displacement map (also known as disparity) between calibrated left and right images. By combining disparity and camera c alibration parameters, depth can be calculated. In recent years, many learning-based stereo networks [1], [2] have achieved exciting successes in the quality and efficiency of disparity estimation.

Depending on how the cost volume is utilized, learning-based stereo networks can be mainly divided into two types: cost aggregation-based methods and iteration-based methods. Cost aggregation-based methods [1], [3] use 3D convolutions to aggregate and regularize the 4D cost volume, and then regress the disparity from the regularized cost volume. This approach effectively combines contextual information and stereo geometric information, showing outstanding performance. However, the cost aggregation and regularization requires a large amount of 3D convolutions, thus limiting practical applications. In contrast, iteration-based methods [4]–[8] use convolutional GRU [9] or LSTM [10] that retrieve features from the cost volume to update the disparity, thus avoiding the expensive cost aggregation operations. Iteration-based methods have achieved comprehensive leadership in both performance and efficiency, becoming the mainstream of recent research.

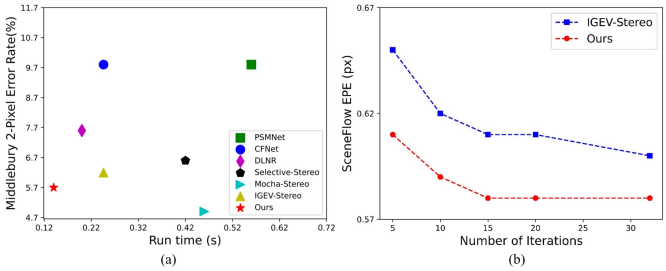

Fig. 1. (a) Comparison with state-of-the-art stereo methods [1], [7], [8], [11], [12] on Middlebury 2014 (Table III). (b) Performance comparison with the baseline model IGEV-Stereo [6] on SceneFlow (Table A1 in Appendix).

However, existing iteration-based methods [4], [6], [7] have difficulties obtaining reliable iterative results in a stable manner. Specifically, accurate disparity may fluctuate during the iterative process. Furthermore, the disparity iteration in ill-posed regions may be unable to escape local optima, i.e., fluctuate around the inaccurate disparity. In this paper, we attempt to address this issue by introducing disparity reliability knowledge during the iterative process.

Existing research [11], [13], [14] uses uncertainty to quantitatively assess the reliability of disparity. Currently, mainstream uncertainty estimation methods [13], [14] take a single image and its corresponding disparity map as inputs and directly predict uncertainty based on the consistency of local geometric features. However, their feature extraction requires complex networks and high computational cost. Some other methods [5], [15] use feature variance in the cost volume as the uncertainty, which are limited to rough assessments. In summary, we need to design a new method that achieves both low cost and high accuracy to meet the demands of multiple uncertainty estimations in iteration-based methods.

In this paper, we propose a novel iteration-based architecture to address the aforementioned issues. This architecture integrates uncertainty with stereo matching, using uncertainty to guide disparity for stable and reliable iterations. To achieve

low-cost and high-accuracy uncertainty estimation, we propose a lightweight plug-and-play uncertainty estimation module (Cost volume-based disparity Uncertainty Estimation, CUE). It consists of only 2 residual blocks and a 1×1 convolution. Leveraging the abundant contextual and local similarity information in the cost volume, CUE accurately estimates disparity uncertainty with minimal computational cost, eliminating the need for additional feature extraction steps. To encourage disparities in ill-posed regions to escape local optima while stabilizing accurate disparities, we propose Uncertainty-based Disparity update Control (UDC), which provides update magnitudes for different pixels that match the current disparity uncertainty. Additionally, we propose Uncertainty-based Disparity Rectification (UDR), which updates the initial disparity based on the fine-tuned changes in uncertainty, achieving effortless optimization of the iterative starting point. All these methods ensure more stable and improved iterative results.

We conducted extensive experiments on SceneFlow, KITTI, Middlebury 2014, and ETH3D. The experimental results indicate that our method effectively improves the performance of existing iteration-based methods for disparity estimation with very low computational cost. As shown in Figure 1, compared to the baseline IGEV-Stereo [6], our method achieves 8.06% performance improvement and 40.7% speedup. Furthermore, plugging the proposed methods into the latest iteration-based method Selective-Stereo [7], we achieve a new state-of-the-art performance on the KITTI online leaderboard.

II. Related Works

1) Iteration-based Methods: Many iteration-based stereo networks [4]–[6], [12], [16] have been successful in terms of quality and efficiency of disparity estimation. RAFT-Stereo [4] is the first iteration-based stereo architecture to be proposed. The overall design is based on RAFT, replacing the all-pairs of 4D correlation volume with a 3D volume. In addition, it introduces a multilevel GRU unit [9] , which remains hidden at multiple resolutions with cross connectivity, but still generates a single high-resolution disparity update. CREStereo [5] designs a hierarchical network with recurrent refinement, updating the disparity in a coarseto-fine pattern, which leads to a better restoration of fine depth details. DLNR [12] proposes an LSTM-based decoupling module to iteratively update the disparity and allows features containing fine details to be shifted iteratively, mitigating the problem that information can be lost during iteration. IGEV-Stereo [6] constructs a combined geometric encoding volume that encodes geometric and contextual information along with local matching details, and iteratively indexes it to update the disparity. Selective-Stereo [7] proposes the use of GRUs with varying kernel sizes to extract disparity information at different frequencies. However, due to the lack of knowledge about disparity reliability, the disparity may fluctuate during the iterative process, making it difficult to obtain stable and reliable iterative results.

2) Disparity Uncertainty Estimation: High-performance stereo methods are not error-free, it is vital to correlate uncertainty with disparity estimation. UCFNet [14] uses uncertainty estimation to filter out highly uncertain pixels from the predicted disparity map, generating sparse but reliable pseudo-labels. By fine-tuning the model through pseudo-labels, UCFNet is able to adapt to new domains. SEDNet [13] proposes a new loss function and an uncertainty estimation subnetwork, which are used for joint disparity and uncertainty estimation. With multi-task learning, SEDNet improves performance on all tasks. CREStereo++ [15] uses a variance-based uncertainty estimation module to adaptively adjust the sampling range during warping, which enhances the robustness of the same model in different scenarios. However, these methods either inefficiently estimate the uncertainty using disparity and the original image as inputs, or only estimate a rough disparity. In this work, we predict the disparity uncertainty by indexing features of the disparity-pair cost volume, ensuring low cost and high accuracy of uncertainty estimation. Moreover, unlike Bayesian-based methods [17], Monte Carlo dropout [18], Deep evidential learning [19], etc., our CUE module can be fully embedded into existing stereo matching frameworks (not specific to iterative methods) and estimate the uncertainty in real-time and iteratively during the inference process.

III. Method

In this section, we first detail CUE that estimates uncertainty with low cost and high accuracy. Then, we propose two uncertainty-guided disparity update methods, UDR and UDC, to achieve stable and reliable disparity iteration. Our proposed methods are plug-and-play. With IGEV-Stereo [6] as the base model, we show how to integrate our methods into existing iteration-based models. The overall architecture is shown in Figure 2.

A. Cost volume-based Uncertainty Estimation

1) Uncertainty Estimation: While estimating uncertainty from disparity and RGB images is the obvious paradigm, its feature extraction requires fine network design and non-negligible computational cost. Especially considering that uncertainty information is required for each disparity in the iteration-based architecture, designing an efficient new paradigm becomes the first problem to be addressed in this paper. It is worth noting that the iteration-based method predicts the error of the current disparity by the features in the cost volume, whose nonlinear transformation is the uncertainty. Therefore, following IGEV-Stereo [6] and Selective-Stereo [7], we retrieve the features of the cost volume through disparity and utilize them to estimate the uncertainty. The simplified formula for this process is as follows:

$$f_{cost}(d) = \sum_{i=-r}^{r} Concat\left\{C_{volume}(d+i)\right\} \tag{1}$$

in which

$$C_{volume}(d,x,y) = \langle f_l(x,y), f_r(x-d,y)\rangle \tag{2}$$

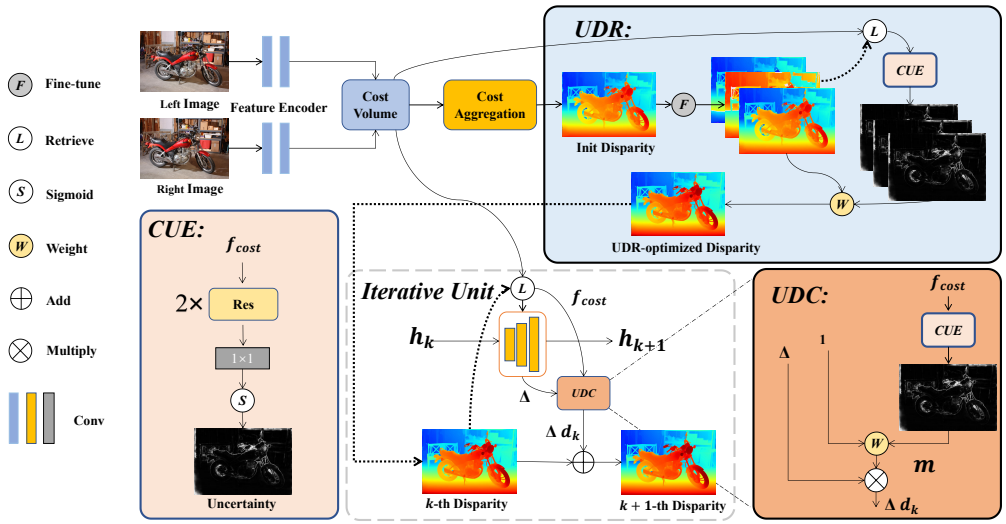

Fig. 2. Overview of our proposed methods. We estimate the disparity uncertainty by cost volume. The init disparity is coarsely optimised once in the UDR and then finely optimised several times through the iterative unit. In the iterative unit, the proposed UDC adjusts the update magnitude to achieve stable and reliable disparity iteration.

where r is the index radius, d is the disparity index, i.e., the current disparity, and $fl(r)$ is the left (right) feature map. The cost volume feature $f_{cost}(d)$, which represents the context consistency and local geometric correlation under the current disparity, can be effectively used to infer disparity uncertainty. Therefore, the uncertainty of the current disparity d can be accurately estimated by the following simple formula:

$$U(d) = \sigma(conv_{1\times1}(Res(Res(f_{cost}(d))))) \tag{3}$$

where σ represents the sigmoid function and res denotes the residual block. The proposed method accomplishes disparity uncertainty estimation using only two residual blocks and a 1×1 convolution. As a result, it can perform dozens of inferences during the iterative process at an extremely low cost, estimating the uncertainty of each disparity. In addition, we show how to compute the ground truth uncertainty in the Appendix.

B. Uncertainty-Guided disparity Update

1) Uncertainty-based Disparity update Control: In order to achieve stable and reliable disparity update, we propose to introduce update reliability knowledge (i.e., uncertainty) in the iterative process. First, we analyze the causes of the existing problems. Taking IGEV-Stereo [6] and Selective-Stereo [7] as examples, their disparity update method can be described as:

$$h_{k+1} = Unit_{GRU}(f_{cost}(d_k), h_k, d_k) \tag{4}$$

$$d_{k+1} = d_k + Decoder_d(h_{k+1}) \tag{5}$$

where d_k is the current disparity, $Unit_{GRU}$ is the disparity update unit, h_k is the current hidden state, and $Decoder_d$ is the decoder used to output the residual disparity.

As commonly understood, apart from storing memories, the hidden state also conveys summarizing feature (also known as historical information) from the iterative process. This feature offers extra guidance for the disparity update. Ideally, the accurate disparity should not change after the update. However, akin to inertia in physics, the guidance of summarizing feature may make the disparity fluctuate. In addition, ambiguous cost information in the ill-posed regions, which leads to conservative and slight update, makes it difficult for the inaccurate disparity to escape from the local optima. Figure 3 illustrates these two phenomena. To address this problem, we propose an uncertainty-guided disparity update method, called UDC. In the UDC, each pixel is provided with an update magnitude that matches the current uncertainty:

$$d_{k+1} = d_k + w_{udc}(d_k) \odot Decoder_d(h_{k+1}) \tag{6}$$

in which

$$w_{udc}(d_k) = 1 + m \times U(d_k) \times \frac{total_{itr} - k}{total_{itr}} \tag{7}$$

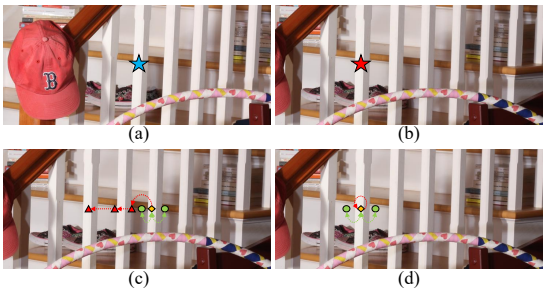

Fig. 3. (a): left image. (b) to (d): right image. In (a) and (b), the blue and red pentagons are the corresponding pixels. (c) and (d) show the cases that inaccurate disparity cannot get rid of the local optima and accurate disparity fluctuates, respectively. The orange diamond represents the pixel that the current disparity points to. The green circles and arrows represent the update results of the existing method, while the red triangles and arrows show the update results of our method. Due to the conditioning of the update magnitude by the uncertainty, our method is not only able to deal with local optima, but also to stabilize the disparity at the correct position.

where m is the modulation coefficient, $total_{itr}$ is the total number of iterations, and $\odot$ denotes the Hadamard Product. We use Eq 6 to enhance the disparity update in Eq 5, as shown in Figure 2. Intuitively, a smaller update magnitude is beneficial for stabilizing accurate disparity, while a larger update magnitude helps the disparity to get rid of the local optima. Thus, in UDC, the update magnitude is dynamic and will be adjusted according to the uncertainty. We set the base weight "1" in Eq 7 to keep the network training and inference stable. Note that the update magnitude in UDC decreases as the number of iterations increases. This strategy contributes to the stability of the final iteration disparity and brings another benefit: pixels with high uncertainty are provided with a larger update magnitude during the previous iterations, which helps the hidden state to convey summarizing feature with a larger receptive field, thereby predicting more accurate disparity results.

2) Uncertainty-based Disparity Rectification: The uncertainty indicates the reliability of the disparity. Therefore, we consider that when uncertainty is significantly reduced after an update, then the update is very likely to be beneficial. Accordingly, we propose a disparity update method called UDR, which optimizes the starting point of iteration without burden. Specifically, we perform a rough update of the disparity based on the uncertainty to correct the obvious errors in it. First, we establish the candidate disparity set $\{d \pm s\}$ centered on the current disparity d, where s denotes the search radius. Then, we weight all the candidate disparities based on the uncertainty with the formula:

$$d_{UDR} = d + s \times (U(d-s) - U(d+s)) \tag{8}$$

where d_{UDR} is the roughly updated disparity. Unlike recurrent convolution-based disparity update, UDR does not introduce an additional module at all, i.e., it updates the disparity only through uncertainty. In addition, another important difference from the recurrent convolution-based methods is that UDR does not rely on learned feature to update the disparity, and is therefore more robust.

C. Training Loss

We compute the smooth L1 loss on the initial disparity d_0 and all disparity uncertainties $U(d_i)$, as well as the L1 loss on all updated disparities $d_{i=1,\dots,total_{itr}}$, see the Appendix for more details.

IV. Experiments

A. Datasets

SceneFlow [20] is a large synthetic dataset containing 35,454 training pairs and 4,370 test pairs at a resolution of 960×540. Following IGEV-Stereo [6] and Selective-Stereo [7], we use Finalpass of Scene Flow for training and testing. **KITTI 2015** [21] is a dataset of realistic driving scenarios. It contains 200 training pairs and 200 test pairs with sparse ground truth disparity. **Middlebury 2014** [22] is an indoor dataset containing 15 training pairs and 15 test pairs, where each scene is provided with three resolutions. **ETH3D** [23] is a grayscale dataset of indoor and outdoor scenes, containing 27 training pairs and 20 test pairs.

B. Implementation Details

We implement the proposed architecture with pytorch and perform experiments on two NVIDIA RTX 3090 GPUs. For all experiments, we use the AdamW optimizer and a one-cycle learning rate schedule with a learning rate of 2e-4. In addition, we preprocess the training pairs using saturation transformation and random cropping. The cropping size is 320×736. Due to constraints on training costs, we use several settings for the batch size and the number of training steps. On SceneFlow, we train the model for 200k steps with a batch size of 8. For the KITTI benchmark, we fine-tune the pre-trained SceneFlow model for 50k steps with a batch size of 8. For the ablation and uncertainty experiments, we train the model for 100k steps with a batch size of 4.

C. Uncertainty Estimation

In this paper, the uncertainty estimation module is the core component of the proposed iterative architecture. Therefore, we first compare the efficiency and accuracy between the proposed CUE module and existing uncertainty estimation methods. We keep the network complexity consistent, i.e., all use two residual blocks to estimate the uncertainty (except for the feature variance-based methods). There are two main reasons for this setting: 1. Our goal is to design a low-cost method. 2. Existing uncertainty estimation methods lack comparisons on generalized datasets.

In addition to using the area under the ROC curve (AUC), we present the **P**er-pixel **U**ncertainty **E**rror (PUE) to evaluate the uncertainty estimation performance (See Appendix for calculation formula). Table I shows the quantitative comparison results. The proposed CUE achieves the best performance on multiple datasets. Particularly, for scenarios not seen in the training set, our method achieves the best accuracy under all metrics. Thus, our method is more robust compared to existing methods. Moreover, our method is only 0.2ms slower than feature variance-based methods, and achieves a significant accuracy advantage. The time cost of the CUE module (less than 1ms) is negligible even when tens of inference runs are performed.

D. Comparisons with State-of-the-art Methods

In this section, we compare the proposed iterative architecture with the state-of-the-art methods on multiple datasets. Note that, compared to the baseline model IGEV-Stereo, our method runs fewer updates in all the experiments. We use this setting to highlight the advantage of the uncertainty-guided architecture for efficiency.

1) In-Domain Evaluation: Table II provides the statistical comparison results with competing methods on SceneFlow and KITTI 2015 benchmarks. Our method shows competitive results on all metrics. On KITTI 2015, our method outperforms IGEV-Stereo by 7.49% on the D1-fg metric.

TABLE I

COMPARISON OF UNCERTAINTY ESTIMATION PERFORMANCE ON SCENEFLOW, MIDDLEBURY AND ETH3D. ALL MODELS ARE TRAINED ONLY ON THE SCENEFLOW TRAINING SET. WE RECORD THE PARAMETER NUMBERS AND THE INFERENCE TIME ON DIFFERENT DATASETS. **BOLD**: BEST.

Methods	SceneFlow AUC↓	SceneFlow PUE↓	Middlebury-H AUC↓	Middlebury-H PUE↓	Middlebury-Q AUC↓	Middlebury-Q PUE↓	ETH3D AUC↓	ETH3D PUE↓	Params(M)
Feature variance-based [15]	0.372	0.134	0.624	0.141	0.613	0.138	0.255	0.128	0
	$0.60ms$		$1.66ms$		$0.40ms$		$0.40ms$		
Image + Disparity [14]	0.139	0.055	0.475	0.085	0.455	0.086	0.217	**0.022**	0.07
	$0.73ms$		$1.40ms$		$0.46ms$		$0.46ms$		
CUE (Ours)	**0.122**	**0.045**	**0.327**	**0.071**	**0.311**	**0.071**	**0.199**	0.029	0.17
	$0.86ms$		$2.20ms$		$0.60ms$		$0.60ms$		

TABLE II

QUANTITATIVE EVALUATION ON SCENEFLOW AND KITTI 2015. OUR METHOD RUNS 15 UPDATES AT INFERENCE. **BOLD**: BEST. BLUE: SECOND. BOTTOM RIGHT CORNER: COMPARED TO THE BASELINE.

Methods	SceneFlow EPE	KITTI 2015 D1-bg	KITTI 2015 D1-fg	KITTI 2015 D1-all	KITTI 2012 2-noc	KITTI 2012 2-all
CREStereo [5]	-	1.45	2.86	1.69	1.72	2.18
UPFNet [14]	-	1.38	2.85	1.62	1.67	2.17
DLNR [12]	0.48	1.60	2.59	1.76	-	-
Croco-Stereo [24]	-	1.38	2.65	1.59	-	-
RAFT-Stereo [4]	0.56	1.58	3.05	1.82	1.92	2.42
NMRF [25]	0.45	**1.28**	3.13	1.59	1.59	2.07
MoCha-Stereo [8]	**0.41**	1.36	**2.43**	1.53	-	-
IGEV-Stereo [6]	0.47	1.38	2.67	1.59	1.71	2.17
Selective-Stereo [7]	0.438	1.33	2.61	1.55	1.59	2.05
Ours(IGEV version)	0.45	1.38	$2.47_{-0.20}$	$1.56_{-0.03}$	$1.67_{-0.04}$	$2.08_{-0.09}$
Ours(Selective version)	0.433	$1.30_{-0.03}$	$2.57_{-0.04}$	**1.51**$_{-0.04}$	**1.56**$_{-0.03}$	**2.02**$_{-0.03}$

TABLE III

SYNTHETIC DATA GENERALIZATION EXPERIMENTS. WE PRE-TRAIN OUR MODEL ON SCENEFLOW AND TEST IT DIRECTLY ON MIDDLEBURY 2014 AND ETH3D. THE 2-PIXEL ERROR RATE IS USED FOR MIDDLEBURY 2014, AND 1-PIXEL ERROR RATE FOR ETH3D. †: IGEV VERSION. ∗: SELECTIVE VERSION. **BOLD**: BEST. BOTTOM RIGHT CORNER: TIMES.

Methods	Middlebury Half	Middlebury Quarter	ETH3D
PSMNet [1]	15.8	9.8	10.2
DSMNet [26]	13.8	8.1	6.2
STTR [2]	15.5	9.7	17.2
CFNet [11]	15.3	9.8	5.8
RAFT-Stereo [4]	8.7	7.3	**3.2**
NMRF [25]	-	7.5	3.8
IGEV-Stereo [6]	$7.1_{0.79s}$	$6.2_{0.25s}$	$3.6_{0.31s}$
Selective-Stereo [7]	$6.8_{1.06s}$	$6.6_{0.34s}$	$5.4_{0.38s}$
Ours†(10 updates)	$7.2_{0.35s}$	$5.8_{0.11s}$	$3.3_{0.15s}$
Ours†(15 updates)	$6.6_{0.46s}$	$5.7_{0.14s}$	$3.3_{0.19s}$
Ours∗(15 updates)	**6.3**$_{0.59s}$	**5.3**$_{0.19s}$	$4.6_{0.23s}$

2) Cross-Domain Generalization: In this experiment, we compare disparity estimation methods on unseen scenes. Specifically, all methods were trained on SceneFlow and then evaluated directly on Middbury 2014 and ETH3D. Table III provides the statistical comparison results. On both datasets, our methods achieve competitive performance (See Appendix for visual results). Particularly, compared to IGEV-Stereo, which runs 32 updates, our method outperforms it by running 10 updates and achieves a speed-up of about 60%. The larger the image resolution, the more pronounced this speed advantage becomes.

E. Ablation Studies

We perform ablation studies to investigate the effectiveness of the proposed methods. All models are trained on the SceneFlow training set, and then evaluated directly on the SceneFlow test set, the Middbury 2014 training set and the ETH3D training set.

We explore the validity and configuration of the proposed modules. As before, our method runs only 15 updates at inference, while the baseline model IGEV-Stereo runs 32 updates. As shown in Table IV, the proposed UDC can significantly improve the prediction accuracy and efficiency. Meanwhile, with the introduction of UDR, our method achieves further performance improvement. In addition, since both UDC and UDR share the parameters from CUE, which are also shared in each iteration, the number of parameters added is only 0.17M. All these results effectively demonstrate the state-of-the-art of our method. More importantly, the modules are effective with almost all configurations. Considering the comprehensive performance on different scenarios, we set m to 0.5 and s to 2.0.

V. CONCLUSION

We propose a novel uncertainty-guided iterative architecture, which solves the problem that existing iteration-based methods are difficult to obtain reliable iteration results stably. Our method consists of three plug-and-play lightweight modules. CUE utilizes the abundant image pair similarity information in the cost volume to infer uncertainty. UDC can dynamically adjust the magnitude of the disparity iteration based on uncertainty, while UDR can achieve efficient optimization of the iteration starting point at almost no cost. We conducted extensive experiments on SceneFlow, KITTI, Middlebury 2014, and ETH3D to validate the effectiveness and superiority of our method. On the KITTI 2015 leaderboard, our method outperforms the baseline IGEV-Stereo by 7.49% on the D1-fg metric. Moreover, our method achieves better

TABLE IV

ABLATION STUDY AND CONFIGURATION EXPLORATION OF THE PROPOSED METHOS ON SCENEFLOW, MIDDLEBURY AND ETH3D. ALL MODELS ARE TRAINED ONLY ON THE SCENEFLOW TRAINING SET. OUR METHOD RUNS ONLY 15 UPDATES AT INFERENCE, WHILE THE BASELINE MODEL IGEV-STEREO RUNS 32 UPDATES. **BOLD**: BEST.

Model	UDC	UDR	SceneFlow $> 3px$	SceneFlow time	Middlebury-H $> 2px$	Middlebury-H time	Middlebury-Q $> 2px$	Middlebury-Q time	ETH3D $> 1px$	ETH3D time	Params. (M)
Baseline			3.16	0.343	8.21	0.787	7.64	0.244	4.36	0.291	12.60
+UDC	$m = 0.5$		2.91		6.76		6.82		3.47		
	$m = 1.0$		2.91	**0.200**	6.77	**0.468**	6.91	**0.142**	3.50	**0.190**	12.77
	$m = 2.0$		2.92		6.74		6.93		3.51		
Full model	$m = 0.5$	$s = 2$	**2.72**	0.202	**6.59**	0.476	**6.66**	0.144	**3.64**	0.192	12.77
	$m = 0.5$	$s = 3$	2.79		6.61		6.68		3.71		

generalization performance on Middlebury 2014 and ETH3D, and improves inference speed by about 60%. In the future, we plan to design a more sophisticated uncertainty-guided stereo architecture to achieve better performance.

REFERENCES

[1] Jia-Ren Chang and Yong-Sheng Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5410–5418.

[2] 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," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6197–6206.

[3] Menglong Yang, Hanyong Wang, and Yang Ren, "A self-attention network for stereo matching," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–10.

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

[5] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16263–16272.

[6] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21919–21928.

[7] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang, "Selective-stereo: Adaptive frequency information selection for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19701–19710.

[8] Ziyang Chen, Wei Long, He Yao, Yongjun Zhang, Bingshu Wang, Yongbin Qin, and Jia Wu, "Mocha-stereo: Motif channel attention network for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27768–27777.

[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[10] Alex Graves and Alex Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[11] 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*, 2021, pp. 13906–13915.

[12] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao, "High-frequency stereo matching network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1327–1336.

[13] Liyan Chen, Weihan Wang, and Philippos Mordohai, "Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17235–17244.

[14] Zhelun Shen, Xibin Song, Yuchao Dai, Dingfu Zhou, Zhibo Rao, and Liangjun Zhang, "Digging into uncertainty-based pseudo-label for robust stereo matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[15] Junpeng Jing, Jiankun Li, Pengfei Xiong, Jiangyu Liu, Shuaicheng Liu, Yichen Guo, Xin Deng, Mai Xu, Lai Jiang, and Leonid Sigal, "Uncertainty guided adaptive warping for robust and efficient stereo matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3318–3327.

[16] Zeyu Ma, Zachary Teed, and Jia Deng, "Multiview stereo with cascaded epipolar raft," in *European Conference on Computer Vision*. Springer, 2022, pp. 734–750.

[17] Abdullah A Abdullah, Masoud M Hassan, and Yaseen T Mustafa, "Leveraging bayesian deep learning and ensemble methods for uncertainty quantification in image classification: A ranking-based approach," *Heliyon*, vol. 10, no. 2, 2024.

[18] Mahesh Subedar, Ranganath Krishnan, Paulo Lopez Meyer, Omesh Tickoo, and Jonathan Huang, "Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6301–6310.

[19] Helbert Paat, Qing Lian, Weilong Yao, and Tong Zhang, "Medl-u: Uncertainty-aware 3d automatic annotation based on evidential deep learning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13976–13982.

[20] 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*, 2016, pp. 4040–4048.

[21] Moritz Menze and Andreas Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3061–3070.

[22] 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*. Springer, 2014, pp. 31–42.

[23] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3260–3269.

[24] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud, "Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17969–17980.

[25] Tongfan Guan, Chen Wang, and Yun-Hui Liu, "Neural markov random field for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5459–5469.

[26] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr, "Domain-invariant stereo matching networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 420–439.