Adaptive Error Aware Cost Volume for Stereo Matching

Anonymous ECCV 2024 Submission

Paper ID #4123

Abstract. Stereo matching, a fundamental challenge in computer vision, has seen remarkable achievements through iterative mechanisms such as RAFT-Stereo and IGEV-Stereo. However, these methods still underperform in ill-posed regions, such as those with occlusions and challenging textures, due to a lack of sufficiently effective motion information in the correlations for these areas to support iterative refinement. How to employ a more rational iterative sampling strategy when obtaining sampled correlations is a question that warrants further investigation. This paper introduces the Adaptive Error Aware Cost Volume (AEACV), which addresses these challenges by integrating the following two modules: 1) Adaptive Error Aware Sampling (AEAS) module dynamically adjusts the sampling range by estimating the error map, effectively optimizing the convergence speed during the disparity estimation process. 2) Error Aware Correlation (EAC) technique that excludes ill-posed regions from cost-volume significantly improves the accuracy of stereo matching. The effectiveness of AEACV-Stereo is validated through extensive experiments, showcasing its superior performance in various scenarios, including those with challenging occlusions. Our method (AEACV-Stereo) ranks first on KITTI 2015, Middlebury, and ETH3D, surpassing existing published methods. It achieves comparable accuracy by utilizing just one-third of the cost volume sampling iterations. Additionally, our method outperforms existing works in zero-shot generalization capabilities across various datasets.

1 Introduction

The emergence of stereo-matching algorithms has revolutionized various fields, including autonomous driving, robotics, and virtual reality, by enabling depth perception through the estimation of disparities between pairs of camera-captured images.

Recent advancements have been made in leveraging 3D Convolutional Neural Networks (CNNs) for cost volume aggregation, yielding promising results [1, 7, 33, 35]. Nevertheless, these methods are computationally intensive. To address this challenge, researchers have turned to iterative algorithms based on cost volume refinement [12, 14, 34, 40, 44], which significantly reduce computational

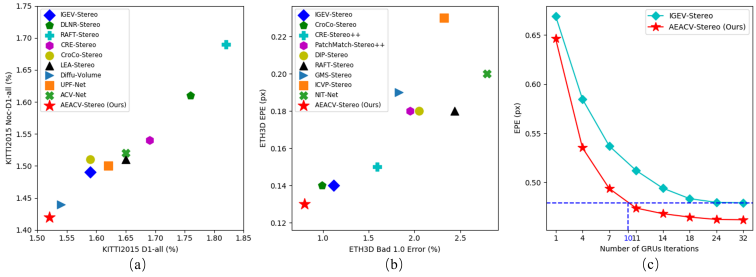


Fig. 1: Performance of our AEACV-Stereo. (a) Comparison with current stereo methods [2, 4, 12, 14, 30, 33, 34, 44, 45] on KITTI 2015 leaderboard. (b) Comparison with current stereo methods [8, 11, 14, 22, 28, 30, 34, 37, 46] on ETH3D leaderboard. (c) End Points Error (EPE) comparison with IGEV-Stereo [34] on SceneFlow test set as the number of iterations changes.

complexity and allow for higher-resolution cost volumes, enhancing algorithmic performance.

Despite these advancements, these methods ignore the noise caused by occlusion during the iteration process. The existing work still has the following problem. The existing work RAFT-Stereo, IGEV, Selective-IGEV, they use GRU module to enhance the edge details, especially, Selective IGEV uses different frequences to balance the edge region and the smooth region effect. However, the occluded error information is also repeated in the refine module.

Considering this problem, we design a new refine mechanism. Firstly, update the occlusion region and correlation volume, then update the final disparity.in iterative methods, pixels with a disparity far from the ground truth and low precision require a larger sampling range to achieve swift convergence. The occlusion information is embedded by the network sending to the correlation module, and GRU module receive a complete information to regress the final disparity. In this process, we can use less iteration numbers in the training process, because of the occlusion region aware refine module. In contrast, pixels with a disparity close to the ground truth demand smaller steps for more nuanced refinement. Second, correlations in these regions lack reliable motion information, hindering the iterative process [12, 14, 44].

In this paper, we introduce a simple and effective approach to tackle these issues. We propose a dynamic sampling strategy based on an error map derived from a lightweight, unsupervised error estimation module. This strategy dynamically adjusts the sampling range and planes, enabling faster convergence. Furthermore, we present a method for constructing a cleaner cost volume by filtering out noise from ill-posed areas, preventing interference during the iterative process for more accurate and cleaner boundaries, as shown in Fig. 2.

Our proposed method, AEACV-Stereo, demonstrates superior performance across benchmark datasets such as SceneFlow [18], KITTI 2015 [19], Middlebury [23] and ETH3D [24], surpassing existing state-of-the-art methods. In terms of iterative speed, our method achieves comparable accuracy to the 32^{nd} iteration

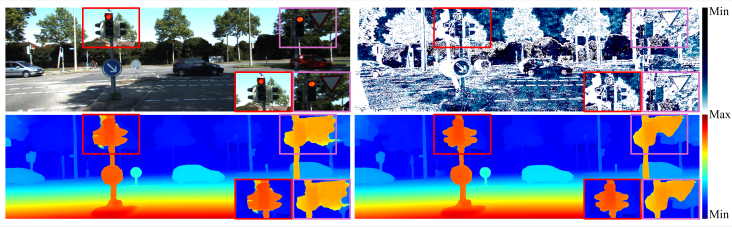


Fig. 2: Visualization of our AEACV-Stereo and IGEV-Stereo [34] on KITTI 2015. With the error aware mask distinguishing the Ill-Posed and normal regions more clearly, our method can perform cleaner boundaries from foreground and background. Up-Left: Input Left Image. Up-Right: Generated Error Aware Mask. Bottom-Left: Disparity Predicted by IGEV-Stereo [34]. Bottom-Right: Disparity Predicted by AEACV-Stereo.

of IGEV-Stereo [34] with only the 10^{th} iteration, as shown in Fig.1. Additionally, the AEACV-RAFT version significantly outperforms RAFT-Stereo [14], validating the method's transferability.

The primary contributions of our work can be summarized as follows: 1) We present a dynamic sampling strategy based on an error map, significantly accelerating iterative speed. 2) We introduce a cost volume construction method that effectively filters out noise from ill-posed regions, enhancing the accuracy of disparity prediction. 3) Our method has demonstrated its effectiveness by outperforming existing methods across various public benchmarks, reinforcing the practicality and potential impact of our approach.

2 Related Works

Learning-based & Iterative Approaches. With the development of deep learning methods, a proliferation of neural network architectures [1,5,7,9,13,14, 20,31–35,38,39] for stereo matching has emerged, marking a significant evolution in the domain over recent years. These methods, like GC-Net [9], PSM-Net [1], and ACV-Net [33], compute the 4D cost volume and deploy the 3D convolutional layers to aggregate the cost volume for predicting disparities. However, their high computational demands limit their efficiency, especially with high-resolution inputs. To address this limitation, cascade, and cost volume pyramid methods [6, 25, 36] have been introduced. For instance, CasStereo [6] and CF-Net [25] employ a coarse-to-fine strategy to improve efficiency. Yet, this approach carries the risk of propagating coarse disparity errors. In this case, iterative methods [12,14,34,40,44] have been proposed. RAFT-Stereo [14] starts to recurrently update the disparity field using local correlation information retrieved from the 3D cost volume. IGEV-Stereo [34] further uses lightweight 3D convolution to obtain the representation of global information and combines it with local information, which significantly improves the effect of challenging areas. This iterative scheme stands out for its ability to balance computational efficiency with

the need for precise disparity estimation. CREStereo [12] introduces a hierarchical network architecture that employs a coarse-to-fine strategy, complemented by a stacked cascaded approach for inference, effectively supplanting the traditional single-resolution iterative framework. Similarly, DLNR [44] innovates by employing LSTM instead of GRU, which offers the distinct benefit of decoupling the updating of hidden states from the disparity prediction process. Most methods [21,29] utilize CNN-based spatial propagation to refine the disparity in occluded regions, yet they demonstrate limited efficacy in addressing large and irregular occluded regions. Moreover, occlusion-aware techniques like GOAT [16] integrate an occlusion estimation module to produce occlusion masks, intending to tack occluded areas. However, potential inaccuracies in these masks might inadvertently affect non-occluded regions.

Sampling Strategies for Cost Volume. Initially, iterative optimization schemes have introduced fixed sampling range [14,34] for cost volume as a strategy to reduce computational load. However, adaptively retrieving information from the cost volume based on the current disparity status is more reasonable. For instance, a global sampling range can provide more information to speed up convergence when disparity worsens, and a local one can prevent fluctuations when disparity is close to ground truth. CREStereo [12] extends the sampling interval to the y direction, improving the performance when epipolar lines are not well aligned; CREStereo++ [8] and UASNet [17] employ cost volume to compute error maps and adjust the sampling space accordingly, facilitating robust stereo matching adaptable to a range of datasets. However, the computation of error maps is contingent upon the recalculation of the cost volume at each iteration. PCV-Stereo [40] introduced an adaptive sampling strategy, utilizing a multi-Gaussian distribution to fit the ground truth and dynamically adjust the sampling range. Despite its potential to significantly reduce the number of iterations, this method incorporates additional supervision, which introduces a degree of complexity.

3 Method

In this section, we introduce three main components of our **A**daptive **E**rror **A**ware **C**ost **V**olume Stereo matching network (AEACV-Stereo), which are 1) Adaptive Error Aware Sampling, 2) Error Aware Correlation and 3) Confidence Aware Refinement. The whole pipeline is shown in Fig. 3.

3.1 Adaptive Error Aware Sampling

To achieve an adaptive sampling mechanism, previous methods mainly use cost volume to estimate the error map from coarse to fine [8, 17] or employ Gaussian distribution [40] to fit ground truth. However, these approaches involve redundant computations of cost volume in each stage or introduce additional supervisory signals like KL divergence. In our design, we introduce a lightweight

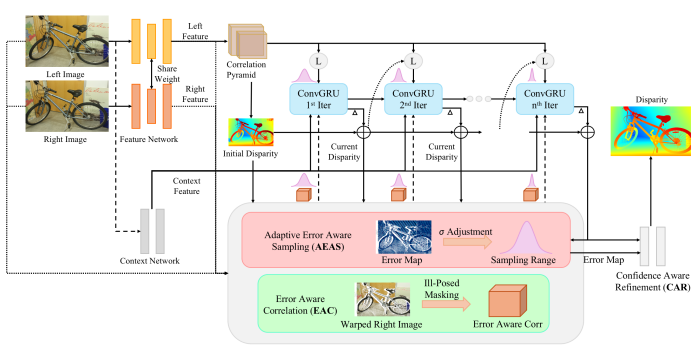


Fig. 3: Our proposed AEACV-Stereo comprises three key components: 1) Adaptive Error Aware Sampling, 2) Error Aware Correlation, and 3) Confidence Aware Refinement. Our method first predicts an Error Map to adaptively adjust the sampling range and employs the sampled planes to compute Error Aware Correlation based on the generated ill-posed mask. The Error Aware Correlation is then combined with traditional correlations to iteratively update the predicted disparity through ConvGRUs.

error estimation module to generate the Error Map, measuring the disparity status without any supervision. This decoupled design separates the computations between cost volume and error map, enabling the cost volume to be calculated once and shared for all iterations. Our sampling strategy is illustrated in Fig. 4.

Error Estimation. In our Error Estimation Module, we refine left and right features $F^r, F^l \in \mathbb{R}^{d\times H\times W}$ and warp the refined right feature F^r_{refine} by applying current disparity ($disp$) as follows:

$$\begin{aligned} F^i_{refine} &= CNN(F^i) \\ F^{rw}_{refine} &= warp(F^r_{refine}, disp) \end{aligned} \tag{1}$$

where $i \in [l, r]$ represents left and right features, F^{rw}_{refine} represents warped refined right feature. In this case, we can decouple the computations of cost volume from generating the error map ($error$) by employing cosine similarity:

$$error = 1 - \frac{F^{rw}_{refine} * F^l_{refine}}{||F^{rw}_{refine}|| * ||F^l_{refine}||} \tag{2}$$

Adaptive Sampling Range. With pre-defined σ and $error$, we can obtain adaptive sampling range r_{adp} without any supervision. Equation 3 shows how to get r_{adp}:

$$r_{adp} = error * \frac{\sigma}{2^{dr}} * r \tag{3}$$

where $\sigma = 32$ is a hyperparameter that represents the max sampling boundary when error equals 1 for the input resolution, dr on behalf of downsampling ratio, $r \in \{-R, -R+1, \ldots, 0, \ldots, R-1, R\}$, and R is a hyperparameter. Then, we directly plus the r_{adp} to the current disparity to obtain all sampling planes.

Fig. 5 illustrates how our adaptive error aware sampling mechanism worked

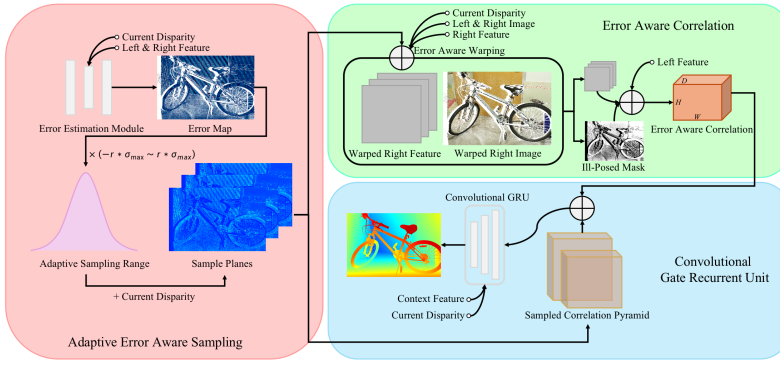


Fig. 4: The architecture of proposed modules. Left: Adaptive Error Aware Sampling. Up-Right: Error Aware Correlation. Bottom-Right: ConvGRU with Error Aware Correlation and Sampled Correlation Pyramid. Details are described in Section 3.

for each iteration. Compared with the fixed sampling method employed in [12, 14,34,44], our sampling mechanism can adjust sampling planes and range during convergence. Unlike [8,17,40], our method doesn't introduce any extra supervisions and redundant computations of correlation, which is easier to implement and transfer to other architectures.

3.2 Error Aware Correlation

For ill-posed areas, the correlations derived from conventional computation methods are less effective, given the absence of corresponding regions in the right image, hindering the provision of meaningful motion information. We have observed that both all-pair correlations [14,34] or correlations derived from warp-based methods [8,17,26] are susceptible to the impact of this issue, as shown in Fig. 6. Therefore, we introduce an Error Aware Correlation (EAC) to address this problem more clearly, which is illustrated in Fig. 4 and 6. Our EAC module effectively reduces noise in ill-posed regions within correlations and provides boundary information to distinguish these areas. This enables the CAR module to refine these regions accordingly.

Error Aware Warping. To distinguish valid and invalid areas, we propose generating a $mask \in \mathbb{R}^{1\times H\times W}$ through error aware warping to obtain clearer information in advance from left and right images $I_l, I_r \in \mathbb{R}^{3\times H\times W}$, according to equation 4.

$$\begin{aligned} I_r^W &= warp(I_r, disp) \\ mask &= \frac{\Sigma_c ||I_l - I_r^W||_1}{c} < \tau \end{aligned} \tag{4}$$

where $c = 3$ is the RGB channel and $\tau = 0.05$ is a threshold, which is a hyperparameter, to control whether the pixel inside valid or invalid areas. Based

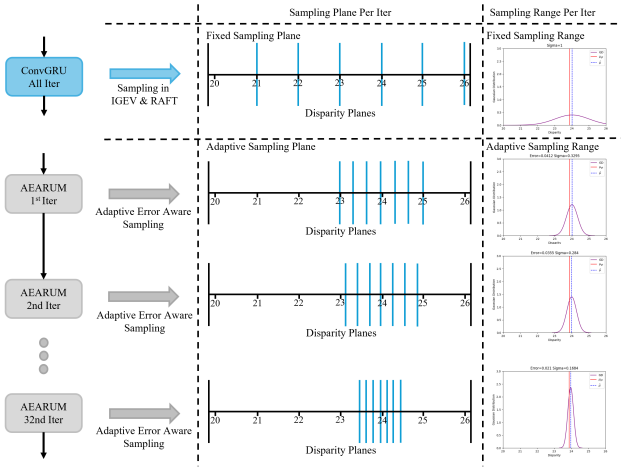


Fig. 5: The visualization of the fixed sampling method (First-Row) and our adaptive error aware sampling method (Bottom-Rows) at each iteration.

on second equation in 1, we use F^r and current disparity to yield warped right feature $F^{rw} \in \mathbb{R}^{d\times H\times W}$, like [8, 17, 26]. After that, we apply the error aware *mask* to mask out F^{rw} and explicitly separate invalid and valid regions.

Correlation Computation. Given all sampling planes, we compute Error Aware Correlation by using F^l and masked out F^{rw} individually. We pick up $p = 9$ sampling planes from r_{adp} (r_{adp} func: 3 and p represents the number of sampling planes), which consists of the current disparity plane and $p-1$ planes symmetric with it. Then we employ each plane to compute Error Aware Warping mentioned before to obtain p pairs of F^l and mask out F^{rw}. These pairs of features will be used to compute correlations individually as follows:

$$Corr(i_p) = \frac{1}{C}\Sigma_{j=1}^{C} F^l(i_p)F^{rw}(i_p) \tag{5}$$

where i_p means i^{th} planing feature, and C represents feature channel. Afterward, all Error Aware Correlations will be concatenated together and deployed in Convolutional GRU (ConvGRU) [14, 34] to generate the motion features with other correlations.

In comparison with other warping-based methods [8, 26], our Error Aware Correlation can generate clearer boundaries between ill-posed and normal regions, and explicitly eliminating the effects of ill-posed areas from the cost volume using an error aware *mask*, as demonstrated in Fig. 6. Due to the strong connection between warping quality and disparity status. Considering the significant impact of warping quality on disparity status, we integrate our Error Aware Correlation (EAC) with Geometric Encoding Volume (GEV) and All-Pairs Correlation (APC) to predict motion features. This hybrid approach employs GEV and APC to guide the initial convergence, with EAC contributing to refinement

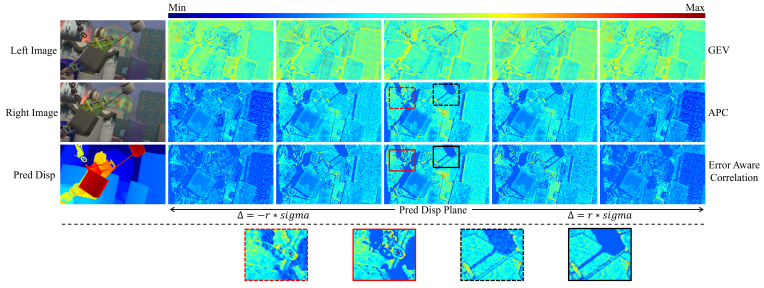


Fig. 6: Visualization of matching correlations in different sampling planes. The GEV treats all regions the same. Although APC can differentiate these two regions, it cannot provide clear boundaries. In contrast, our Error Aware Correlation can clearly distinguish ill-posed and normal regions, providing more accurate matching details to the network. Left-Column: Left and Right Images from SceneFlow and Predicted Disparity. Up-Row: Geometric Encoding Volume (GEV) in IGEV-Stereo [34]. Middle-Row: All-Pairs Correlation (APC) in RAFT-Stereo [14]. Bottom-Row: Our Error Aware Correlation.

in later stages, ensuring stable convergence and preventing the accumulation of errors caused by poor warping quality at the outset.

3.3 Confidence Aware Refinement

To further integrate the error map into the entire model, we draw inspiration from PCV-Stereo [40] and refine the last output disparity map with the final error map. We employ a confidence-aware refinement module to enhance the disparity in detailed areas. The confidence map is reconstructed as:

$$conf = 1 - error \tag{6}$$

Given the confidence map, we refined the final disparity step by step as follows:

$$\begin{aligned} F^{disp} &= CNN_1(disp) \\ F^{cat} &= Concat(F^{disp}, F^{l}, F^{l}_{refine}) \\ x &= CNN_3(ReLU(CNN_2(F^{cat}))) \\ disp_{refined} &= disp + x * conf \end{aligned} \tag{7}$$

where F^l means left feature, F^l_{refine} means refined left feature.

3.4 Loss Function

Similar to [34], we employ smooth L1 Loss [1] on initial disparity d_0 by using equation 8, if present. Otherwise, the L_{init} will be zero.

Table 1: Quantitative evaluation on SceneFlow test set. The best result is in bold.

Method	GwcNet [7]	GANet [41]	CSPN [3]	LEAStereo [4]	ACVNet [33]	IGEV-Stereo [34]	ours
AvgErr	0.76	0.84	0.78	0.78	0.48	0.47	**0.46**

$$L_{init} = Smooth_{L_1}(d_0 - d_{gt}) \tag{8}$$

where d_{gt} means the ground truth disparity. We calculate the L1 loss on all predicted disparities $\{d_i\}_{i=1}^{N}$. We follow [14] to exponentially increase weights, and the total loss is defined as:

$$L_{stereo} = L_{init} + \Sigma_{i=1}^{N} \gamma^{N-i} ||d_i - d_{gt}||_1 \tag{9}$$

where $\gamma = 0.9$, and d_{gt} represent ground truth.

4 Experiments

4.1 Implementation Details

Our model's performance is evaluated on the SceneFlow [18] dataset and three public benchmarks: ETH3D [24], Middlebury [23], and KITTI-2015 [19]. The model is implemented using PyTorch and experiments are conducted on NVIDIA A100 GPUs. For pretraining and ablation studies, the model is initially trained on the synthetic SceneFlow [18] training set (both cleanpass and finalpass) for 200k iterations with a batch size of 8, followed by evaluation on the SceneFlow test set. We employ the AdamW optimizer [10] with an initial learning rate of $2e^{-4}$ and use a OneCycle scheduler [27] with a warm-up strategy. Data augmentation is applied in accordance with the settings in IGEV-Stereo [34]. Images are randomly cropped to 320×736.

When fine-tuning on KITTI [19], the pre-trained SceneFlow model is used on mixed KITTI 2012 and KITTI 2015 training image pairs for 20k iterations, with a learning rate of $1e^{-4}$. For Middlebury [23] and ETH3D [24], given the limited size of the training sets, the pre-trained model is further trained on mixed datasets comprising synthetic SceneFlow [18], CREStereo [12], ETH3D [24], and Middlebury (2005, 2006, 2021, and V3) [23]. Consistent data augmentation including saturation change, image perturbance, and random scale is applied during both the pretraining and fine-tuning stages.

4.2 Comparisons with State-of-the-art

We evaluate AEACV-Stereo on SceneFlow, KITTI 2015, ETH3D, and Middlebury, comparing its performance with the published state-of-the-art methods.

As shown in Tab. 1, on the SceneFlow test set, when compared to existing methods, we achieve a new state-of-the-art EPE of 0.46, highlighting the superior performance of our approach.

Table 2: Quantitative evaluation on the KITTI-2015 leaderboard. "Noc" and "All" indicate the non-occluded and overall regions, respectively. The best results for each evaluation metric are bolded, and the second-best are underlined.

Method	Noc(%)			All(%)		
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all
RAFT-Stereo [14]	1.45	2.94	1.69	1.58	3.05	1.82
CREStereo [12]	1.33	2.60	1.54	1.45	2.86	1.69
DLNR [44]	1.45	<u>2.39</u>	1.61	1.60	<u>2.59</u>	1.76
CroCo v2 [30]	1.30	2.56	1.51	<u>1.38</u>	2.65	<u>1.59</u>
IGEV-Stereo [34]	<u>1.27</u>	2.62	<u>1.49</u>	<u>1.38</u>	2.67	<u>1.59</u>
AEACV-Stereo (ours)	**1.23**	**2.36**	**1.42**	**1.35**	**2.38**	**1.52**

Table 3: Quantitative evaluation on the Middlebury leaderboard. "Noc" and "All" indicate the non-occluded and overall regions, respectively. The best results for each evaluation metric are bolded, and the second-best are underlined.

Method	Noc(%)				All(%)			
	AvgErr	RMS	Bad 2.0	Bad 4.0	AvgErr	RMS	Bad 2.0	Bad 4.0
RAFT-Stereo [14]	1.27	8.40	4.74	2.75	2.71	12.6	9.37	6.42
CroCo v2 [30]	1.76	8.91	4.90	4.18	2.36	10.6	11.1	6.75
GMStereo [30]	1.31	<u>6.45</u>	7.14	2.96	<u>1.89</u>	**8.03**	11.7	6.07
CREStereo [12]	1.15	7.70	<u>3.71</u>	2.04	2.10	10.5	8.13	5.05
DLNR [44]	<u>1.06</u>	7.78	**3.20**	**1.89**	1.91	10.2	**6.98**	<u>4.77</u>
IGEV-Stereo [34]	2.89	12.8	4.83	3.33	3.64	15.1	8.16	5.79
AEACV-Stereo (ours)	**0.99**	**6.32**	4.15	<u>1.97</u>	**1.56**	<u>8.13</u>	<u>7.35</u>	**4.22**

We assess AEACV-Stereo on the KITTI-2015 test set, and the results are submitted to the online KITTI leaderboard. As presented in Tab. 2, our method secures the 1^{st} position among all published methods and ranks 2^{nd} overall out of 300+ submissions. We outperform IGEV and Croco v2 by 7% on the D1-all metric, achieving the best result across all listed metrics, with notable improvements observed in both non-occluded and overall regions.

Regarding Middlebury, our model is trained on mixed datasets with 5% augmented training data from Middlebury datasets. The training process continues for an additional 150k iterations with 768×1024 image crops, considering the larger size of Middlebury image pairs. The results are submitted to the online evaluation benchmark. As shown in Tab. 3, our method achieves the best average error and competitive Bad 2.0 and Bad 4.0 metric compared to all published methods. Given our method's focus on filtering out noises and adverse effects in challenging regions, the outstanding results in average error highlight the overall improvement of our method across the entire image. As depicted in Fig 7 and Fig. 8, our method demonstrates more accurate predictions in challenging areas, such as the boundaries of different objects, enabling better differentiation between them.

For ETH3D, we train our network on the complete training set, incorporating 2.5% augmented training data from the ETH3D low-res two-view stereo dataset.

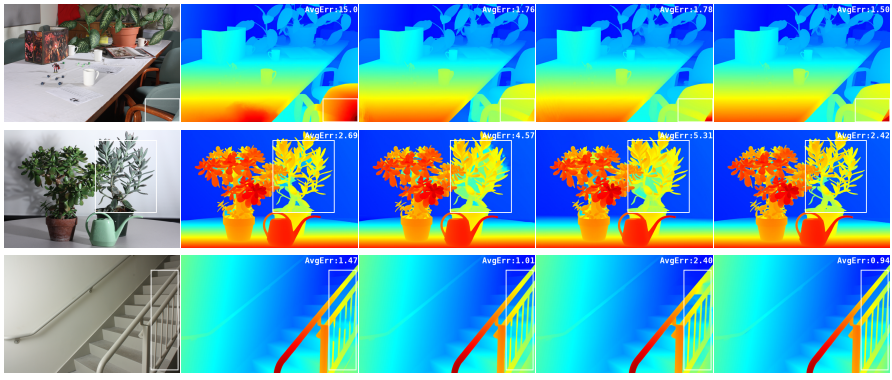


Fig. 7: The visualization of AEACV-Stereo(ours) and other state-of-the-art methods on Middlebury dataset.

Table 4: Quantitative evaluation on the ETH3D leaderboard in non-occluded (noc) regions.

Method	AvgErr	Bad 1.0	Bad 0.5
HITNet [28]	0.20	2.79	7.89
RAFT-Stereo [14]	0.18	2.44	7.04
DIP-Stereo [46]	0.18	1.97	6.74
GMStereo [37]	0.19	1.83	5.94
CREStereo [12]	**0.13**	0.98	3.58
CroCo v2 [30]	0.14	0.99	3.27
IGEV-Stereo [34]	0.14	1.12	3.52
AEACV-Stereo(ours)	**0.13**	**0.80**	**3.17**

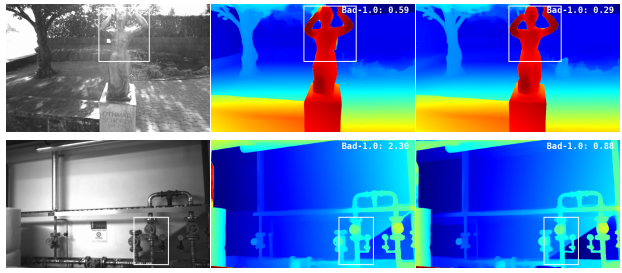
Table 5: Model generalization experiments. 2-pixel error rate for Middlebury and 1-pixel error rate for ETH3D.

Model	Middlebury half	Middlebury quater	ETH3D
GANet [41]	13.5	8.5	6.5
DSMNet [42]	13.8	8.1	6.2
FC-GANet [43]	10.2	7.8	5.8
Graft-GANet [15]	9.8	-	6.2
RAFT-Stereo [14]	8.7	7.3	**3.2**
IGEV-Stereo [34]	7.1	6.2	3.6
AEACV-Stereo(ours)	**5.9**	**5.1**	3.7

The same number of iterations are applied with 416×640 image crops. Quantitative comparisons are provided in Tab. 4, which showcase our state-of-the-art performance among published methods on the online benchmark for the majority of metrics. Remarkably, our method outperforms the published state-of-the-art by 18% on the Bad 1.0 metric, establishing itself as the current state-of-the-art on the leaderboard.

4.3 Zero-Shot Generalization

Exploring the adaptability of AEACV-Stereo, we investigate its potential to generalize from synthetic training data to previously unseen real world datasets. Given the difficulty in acquiring large real world datasets for training, the ability of stereo models to exhibit generality is of paramount importance. We train AEACV-Stereo using the SceneFlow dataset, adopting the identical settings as IGEV-Stereo [34], and then we directly evaluate its effectiveness on the Middlebury 2014 and ETH3D training sets. As depicted in Tab. 5, AEACV-Stereo demonstrates competitive performance in the same zero-shot setting.



(a) Left Image **(b)** IGEV-Stereo [34] **(c)** Ours

Fig. 8: The visualization of IGEV-Stereo and AEACV-Stereo(ours) on ETH3D dataset.

Table 6: Ablation study of proposed module on the SceneFlow test set. AEAS, CAR and EAC refer to Adaptive Error Aware Sampling, Confidence Aware Refinement and Error Aware Correlation module, respectively. Baseline is the official IGEV-Stereo [34].

Model	*AEAS*	*CAR*	*EAC*	AvgErr	Bad 1.0	Params(M)
Baseline				0.479	2.47	12.60
AEAS	✓			0.470	2.43	13.30
AEAS+CAR	✓	✓		0.466	2.42	13.53
Full method	✓	✓	✓	**0.462**	**2.40**	13.53

4.4 Ablation Study

To further validate the efficacy of each module in our method and explore the optimal configuration of error-aware correlation, we conducted comprehensive ablation experiments on the SceneFlow dataset.

Model Components. As illustrated in Tab. 6, the utilization of an adaptive sampling strategy results in enhanced convergence and improved prediction accuracy. Compared to baseline model, our adaptive sampling approach enables a denser sampling space in accurately estimated regions and expands the sampling range in challenging areas, enabling more comprehensive information acquisition for updating the iterative direction. The incorporation of the confidence-aware refinement module allows for additional fine-tuning, refining the precision of the obtained disparity map.

Additionally, the introduction of the designed Error Aware Correlation module further elevates the accuracy of our model on the SceneFlow test set, albeit with an increase in computational complexity. However, its inclusion results in higher precision, ensuring overall improved performance, especially when dealing with ill-posed regions.

Number of sampling planes Since our proposed **E**rror **A**ware **C**orrelation (**EAC**) relies on the number of sampling planes, and the utilization of different planes can yield varying performance, we conduct an exploration of different settings to determine the most suitable configuration for our model. Tab. 7 presents the results of different correlation settings. We apply the same training settings on the synthetic SceneFlow training set and evaluate our model on the Scene-

Table 7: Exploration of plane numbers. Standard evaluation thresholds include a 2-pixel error rate for Middlebury [23] and a 3-pixel error rate for KITTI 2015 [19].

Model	*Planes*	SceneFlow AvgErr	Middlebury half	Middlebury quater	KITTI-2015
AEAS+CAR	-	0.466	7.37	6.65	6.07
Full Method	1	0.466	5.84	5.64	6.13
	3	0.465	**5.56**	6.93	6.03
	9	**0.462**	5.88	**5.09**	**5.86**

Table 8: Model generalization experiments. 2-pixel error rate for Middlebury-Half and 1-pixel error rate for ETH3D.

Model	SceneFlow AvgErr	SceneFlow Bad 1.0	Mid-H	ETH3D
RAFT-Stereo [14]	0.72	3.4	7.3	3.2
AEACV (RAFT-based)	**0.56**	**3.0**	**6.4**	**2.9**

Table 9: Quantitative evaluation on the KITTI-2015 leaderboard over non-occluded (noc) regions.

Model	Noc (%) D1-bg	D1-fg	D1-all
RAFT-Stereo [14]	1.58	3.05	1.82
AEACV (RAFT-based)	**1.52**	**2.72**	**1.72**

Flow test set, as well as the training sets of Middlebury and KITTI 2015. While the 1-plane and 3-plane configurations show marginal improvements in SceneFlow, the overall performance on Middlebury and KITTI 2015 is enhanced. The 9-plane setting ultimately achieves the best overall performance. The increase in the number of planes assists the model in obtaining a more precise motion direction under conditions of imprecise disparity estimation. Therefore, the final choice for our correlation configuration is the 9-plane setting.

4.5 Transferability of our method

Our error-aware correlation is built on the all-pairs cost volume, making it easily transferable to other cost-volume-based stereo-matching approaches. To further validate the efficacy and transferability of our designs across different cost-volume-based methods, we implement a new version of our method based on RAFT-Stereo [14], following the same training settings.

We initially train the AEACV (RAFT-based) method on the synthetic SceneFlow training set for 200k iterations. As shown in Tab. 8, we evaluate our method on the SceneFlow test set, as well as the Middlebury and ETH3D training sets. The results show that our method not only further promotes convergence on SceneFlow but also exhibits improved generalization capabilities. Additionally, we perform fine-tuning on KITTI using the same settings as we used. As demonstrated in Tab. 9. Compared with the RAFT-Stereo [14], our AEACV (RAFT-based) significantly improves the overall performance.

5 Conclusion and Discussion

Our proposed AEACV-Stereo achieves state-of-the-art results on the real world and synthetic datasets, and outperforms all published works on KITTI

2015, ETH3D, Middlebury and SceneFlow by leveraging Adaptive Error Aware Sampling module, Error Aware Correlation module, and Confidence Aware Refinement module, which achieve similar performance to previous SOTA (32 iterations) in only 10 iterations. Meanwhile, we also validate our method's flexibility and robustness by incorporating our module into RAFT-Stereo.

Our model only requires one-third of iterations to reach excellent accuracy. However, the extra warping computations still cause our model to have a similar inference time to previous iterative-based methods. We are considering introducing a novel way to further speed up each iteration and the entire inference time. Furthermore, we are considering transferring our methodology to Multi-View Stereo Matching, 3D-Reconstruction, and VSLAM.

References

1. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5410–5418 (2018) 1, 3, 8
2. Chen, Q., Ge, B., Quan, J.: Unambiguous pyramid cost volumes fusion for stereo matching. IEEE Transactions on Circuits and Systems for Video Technology (2023) 2
3. Cheng, X., Wang, P., Yang, R.: Learning depth with convolutional spatial propagation network. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2361–2379 (2019) 9
4. Cheng, X., Zhong, Y., Harandi, M., Dai, Y., Chang, X., Li, H., Drummond, T., Ge, Z.: Hierarchical neural architecture search for deep stereo matching. Advances in Neural Information Processing Systems **33**, 22158–22169 (2020) 2, 9
5. Duggal, S., Wang, S., Ma, W.C., Hu, R., Urtasun, R.: Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4384–4393 (2019) 3
6. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2495–2504 (2020) 3
7. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3273–3282 (2019) 1, 3, 9
8. Jing, J., Li, J., Xiong, P., Liu, J., Liu, S., Guo, Y., Deng, X., Xu, M., Jiang, L., Sigal, L.: Uncertainty guided adaptive warping for robust and efficient stereo matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3318–3327 (2023) 2, 4, 6, 7
9. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE international conference on computer vision. pp. 66–75 (2017) 3
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9

11. Kwon, O.H., Zell, E.: Image-coupled volume propagation for stereo matching. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 2510–2514. IEEE (2023) 2
12. Li, J., Wang, P., Xiong, P., Cai, T., Yan, Z., Yang, L., Liu, J., Fan, H., Liu, S.: Practical stereo matching via cascaded recurrent network with adaptive correlation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16263–16272 (2022) 1, 2, 3, 4, 6, 9, 10, 11
13. Liang, Z., Guo, Y., Feng, Y., Chen, W., Qiao, L., Zhou, L., Zhang, J., Liu, H.: Stereo matching using multi-level cost volume and multi-scale feature constancy. IEEE transactions on pattern analysis and machine intelligence **43**(1), 300–315 (2019) 3
14. Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International Conference on 3D Vision (3DV). pp. 218–227. IEEE (2021) 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 13
15. Liu, B., Yu, H., Qi, G.: Graftnet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13012–13021 (2022) 11
16. Liu, Z., Li, Y., Okutomi, M.: Global occlusion-aware transformer for robust stereo matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3535–3544 (2024) 4
17. Mao, Y., Liu, Z., Li, W., Dai, Y., Wang, Q., Kim, Y.T., Lee, H.S.: Uasnet: Uncertainty adaptive sampling network for deep stereo matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6311–6319 (2021) 4, 6, 7
18. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: 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. pp. 4040–4048 (2016) 2, 9
19. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3061–3070 (2015) 2, 9, 13
20. Nie, G.Y., Cheng, M.M., Liu, Y., Liang, Z., Fan, D.P., Liu, Y., Wang, Y.: Multi-level context ultra-aggregation for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3283–3291 (2019) 3
21. Park, J., Joo, K., Hu, Z., Liu, C.K., So Kweon, I.: Non-local spatial propagation network for depth completion. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 120–136. Springer (2020) 4
22. Ren, W., Liao, Q., Shao, Z., Lin, X., Yue, X., Zhang, Y., Lu, Z.: Patchmatch stereo++: Patchmatch binocular stereo with continuous disparity optimization. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2315–2325 (2023) 2
23. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: 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. pp. 31–42. Springer (2014) 2, 9, 13

24. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: 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. pp. 3260–3269 (2017) 2, 9
25. Shen, Z., Dai, Y., Rao, Z.: Cfnet: Cascade and fused cost volume for robust stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13906–13915 (2021) 3
26. Shen, Z., Dai, Y., Song, X., Rao, Z., Zhou, D., Zhang, L.: Pcw-net: Pyramid combination and warping cost volume for stereo matching. In: European Conference on Computer Vision. pp. 280–297. Springer (2022) 6, 7
27. Smith, L.N., Topin, N.: 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, pp. 369–386. SPIE (2019) 9
28. Tankovich, V., Hane, C., Zhang, Y., Kowdle, A., Fanello, S., Bouaziz, S.: Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14362–14372 (2021) 2, 11
29. Wang, T., Ma, C., Su, H., Wang, W.: Cspn: Multi-scale cascade spatial pyramid network for object detection. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1490–1494. IEEE (2021) 4
30. Weinzaepfel, P., Lucas, T., Leroy, V., Cabon, Y., Arora, V., Brégier, R., Csurka, G., Antsfeld, L., Chidlovskii, B., Revaud, J.: 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. pp. 17969–17980 (2023) 2, 10, 11
31. Wu, Z., Wu, X., Zhang, X., Wang, S., Ju, L.: Semantic stereo matching with pyramid cost volumes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7484–7493 (2019) 3
32. Xu, B., Xu, Y., Yang, X., Jia, W., Guo, Y.: Bilateral grid learning for stereo matching networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12497–12506 (2021) 3
33. Xu, G., Cheng, J., Guo, P., Yang, X.: Attention concatenation volume for accurate and efficient stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12981–12990 (2022) 1, 2, 3, 9
34. Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21919–21928 (2023) 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12
35. Xu, G., Wang, Y., Cheng, J., Tang, J., Yang, X.: Accurate and efficient stereo matching via attention concatenation volume. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 1, 3
36. Xu, G., Wang, Y., Cheng, J., Tang, J., Yang, X.: Accurate and efficient stereo matching via attention concatenation volume. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 3
37. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Yu, F., Tao, D., Geiger, A.: Unifying flow, stereo and depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 2, 11
38. Yang, G., Zhao, H., Shi, J., Deng, Z., Jia, J.: Segstereo: Exploiting semantic information for disparity estimation. In: Proceedings of the European conference on computer vision (ECCV). pp. 636–651 (2018) 3

39. Yao, C., Jia, Y., Di, H., Li, P., Wu, Y.: A decomposition model for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6091–6100 (2021) 3
40. Zeng, J., Yao, C., Yu, L., Wu, Y., Jia, Y.: Parameterized cost volume for stereo matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18347–18357 (2023) 1, 3, 4, 6, 8
41. Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.: Ga-net: Guided aggregation net for end-to-end stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 185–194 (2019) 9, 11
42. Zhang, F., Qi, X., Yang, R., Prisacariu, V., Wah, B., Torr, P.: Domain-invariant stereo matching networks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 420–439. Springer (2020) 11
43. Zhang, J., Wang, X., Bai, X., Wang, C., Huang, L., Chen, Y., Gu, L., Zhou, J., Harada, T., Hancock, E.R.: Revisiting domain generalized stereo matching networks from a feature consistency perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13001–13011 (2022) 11
44. Zhao, H., Zhou, H., Zhang, Y., Chen, J., Yang, Y., Zhao, Y.: High-frequency stereo matching network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1327–1336 (2023) 1, 2, 3, 4, 6, 10, 11
45. Zheng, D., Wu, X.M., Liu, Z., Meng, J., Zheng, W.s.: Diffuvolume: Diffusion model for volume based stereo matching. arXiv preprint arXiv:2308.15989 (2023) 2
46. Zheng, Z., Nie, N., Ling, Z., Xiong, P., Liu, J., Wang, H., Li, J.: Dip: Deep inverse patchmatch for high-resolution optical flow. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8925–8934 (2022) 2, 11