MoCha-Stereo: Motif Channel Attention Network for Stereo Matching

Ziyang Chen, Wei Long, He Yao, Yongjun Zhang ✉, Bingshu Wang, Yongbin Qin, Jia Wu

GUIZHOU UNIVERSITY

Project Page: https://github.com/ZYangChen/MoCha-Stereo

CVPR SEATTLE, WA JUNE 17-21, 2024

Introduction

Due to the fuzziness of geometric edges in certain channels, achieving accurate matching of stereo image edges is a challenging problem. MoCha-Stereo guides ordinary channels to focus on edge features through motif channels, achieving more accurate detail matching. Motif Channel refers to channel that composed of repeatedly occurring geometric contours. The regions delineated by the yellow border represent the magnified details.

- We introduce a novel stereo matching framework that incorporates repeated geometric contours. This architecture enables more accurate cost computation and disparity estimation through detail restoration of feature channels.
- We propose Motif Channel Attention (MCA) to mitigate imbalanced nonlinear transformations in network training. MCA optimize feature channels through motif channel projection instead of direct network optimization. Inspired by time-series motif mining, we capture motif channel using sliding windows.
- To achieve more precise matching cost computation for edge matching, we construct the Channel Affinity Matrix Profile (CAMP)-guided correlation volume. This volume is derived from the correlation matrix between normal and motif channels, then mapped onto the base correlation volume to produce a more rational cost volume called Motif Channel Correlation Volume (MCCV).
- To leverage the geometric information of the potential channels in the reconstruction error map, we develop Reconstruction Error Motif Penalty (REMP) to extract the motif channels from the error map, optimising the disparity based on the high and low-frequency signals.

Method

- Motif Channel Attention (MCA) for feature maps

$$f_{fre}^{mc}(s,h,w)=\sum_{c=1}^{N_c}\sum_{i=1}^{3}\sum_{j=1}^{3}(SW(s,h+i,w+j)\times f_{fre}(c,h,w))$$

- Channel Affinity Matrix Profile (CAMP) guided Correlation Volume

$$CAMP(s,c,h,w)=f^{mc}(s,h,w)\times f_{l,4}(c,h,w)$$

- Motif Channel Correlation Volume

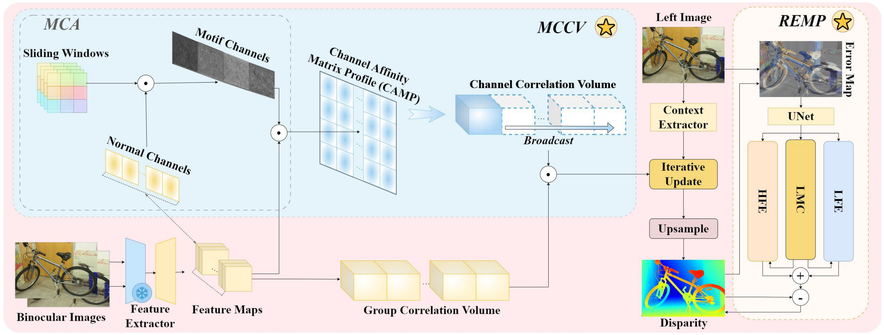
$$C(d,h,w)=\sum_{g=1}^{N_g}(C_g(d,h,w,g)\times C_c(d,h,w))$$

- Reconstruction Error Motif Penalty (REMP)

$$o=UNet(Concat(d_n^{'},E))$$

$$HFE(o)=o\odot LMC(o)$$

$$d_n=d_n^{'}-Conv(LFE(o)\odot(1-LMC(o))+HFE(o))$$



Experiment

Method	Lac-GwcNet[18]	UPFNet [5]	ACVNet [35]	DLNR [47]	IGEV-Stereo [36]	MoCha-Stereo (Ours)
EPE (px)↓	0.75	0.71	0.48	0.48	0.47	**0.41** (−12.77%)
Time (s)↓	0.65	**0.27**	0.48	0.30	0.37	0.34

Table 1. Quantitative evaluation on Scene Flow test set. The **best** result is bolded, and the second-best result is underscored. The variations in the performance of our method compared to the optimal results of other methods are indicated in red font.

Method	All				Reflective			
	Out-Noc (%)↓	Out-All (%)↓	Avg-Noc (px)↓	Avg-All (px)↓	**Out-Noc (%)↓**	Out-All (%)↓	Avg-Noc (px)↓	Avg-All (px)↓
GwcNet[12]	1.32	1.70	0.5	0.5	7.80	9.28	1.3	1.4
AcfNet[45]	1.17	1.54	0.5	0.5	6.93	8.52	1.8	1.9
RAFT-Stereo [17]	1.30	1.66	**0.4**	0.5	5.40	6.48	1.3	1.3
HITNet[30]	1.41	1.89	**0.4**	0.5	5.91	7.54	1.0	1.2
CREStereo[15]	1.14	1.46	**0.4**	0.5	6.27	7.27	1.4	1.4
Lac-GwcNet [18]	1.13	1.49	0.5	0.5	6.26	8.02	1.5	1.7
IGEV-Stereo [36]	1.12	1.44	**0.4**	**0.4**	4.35	5.00	1.0	1.1
MoCha-Stereo(Ours)	**1.06**(−5.36%)	**1.36**	**0.4**	**0.4**	**3.83**(−11.95%)	**4.50**	**0.8**	**0.9**

Table 2. Results on the KITTI-2012 leaderboard. Out-Noc represents the percentage of erroneous pixels in non-occluded areas, Out-All denotes the percentage of erroneous pixels in the entire image. Avg-Noc refers to the end-point error in non-occluded areas, Avg-All indicates the average disparity error across the entire image. Error threshold is 3 px.

Method	All pixels (%)↓			Noc pixels (%)↓		
	bg	fg	**all**	bg	fg	**all**
GwcNet[12]	1.74	3.93	2.11	1.61	3.49	1.92
RAFT-Stereo[17]	1.58	3.05	1.82	1.45	2.94	1.69
CREStereo[15]	1.45	2.86	1.69	1.33	2.60	1.54
Lac-GwcNet[18]	1.43	3.44	1.77	1.30	3.29	1.63
CFNet[26]	1.54	3.56	1.81	1.43	3.25	1.73
UPFNet[5]	1.38	2.85	1.62	1.26	2.70	1.50
CroCo-Stereo[34]	1.38	2.65	1.59	1.30	2.56	1.51
IGEV-Stereo[36]	1.38	2.67	1.59	1.27	2.62	1.49
DLNR[47]	1.60	2.59	1.76	1.45	**2.39**	1.61
MoCha-Stereo (Ours)	**1.36**	**2.43**	**1.53** −3.77%	**1.24**	2.42	**1.44** −3.36%

Table 3. Results on the KITTI-2015 leaderboard. Error threshold is 3 px. Background error is indicated by bg, and front-ground error by fg.

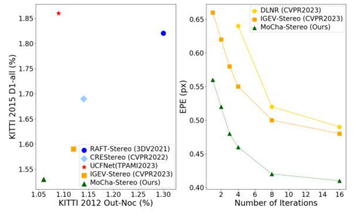


Figure 2. (a) Comparison with SOTA methods [15, 17, 28, 36] on KITTI 2012 [9] and 2015 leaderboards [22] (lower is better). (b) Performance evaluation of the Scene Flow test set [21] in comparison to IGEV-Stereo [36] and DLNR [47] as the number of iterations changes (lower EPE means better).

Conclusion

We propose MoCha-Stereo, a novel stereo matching framework. MoCha-Stereo aims to alleviate edge mismatch caused by the geometric structure blurring of channel features. MoCha-Stereo showcases robust cross-dataset generalization capabilities. It ranks **1st** on the KITTI-2015 and KITTI-2012 Reflective benchmarks and demonstrates SOTA performance on ETH3D, Middlebury, Scene Flow datasets and MVS domain. In the future, we plan to extend the motif channel attention mechanism to more processes in stereo matching, further enhancing the capability of algorithm for edge matching.