Paper ID 1223

ACM multimedia
Chengdu, China OCT 20-24 2021

BridgeNet: A Learning Network of Depth Map Super-Resolution and Monocular Depth Estimation

Qi Tang[1,2], Runmin Cong[1,2,4]*, Ronghui Sheng[1,2], Lingzhi He[1,2], Dan Zhang[3], Yao Zhao[1,2], and Sam Kwong[4]

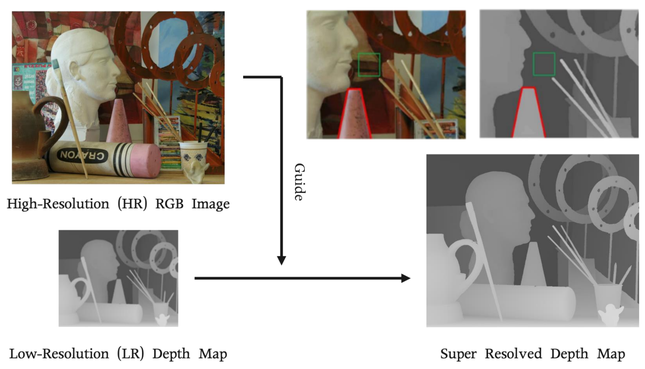
[1]Institute of Information Science, Beijing Jiaotong University, Beijing, China

[2]Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China

[3]UISEE Technology (Beijing) Co., Ltd. Beijing, China [4]City University of Hong Kong, Hong Kong, China

Section 1 (Problem):

- High-resolution color image has strong structural similarity with the depth map, thus it can provide some guidance information for depth SR. However, because there are still some differences between the two modalities, direct information transmission in the feature dimension or edge map dimension cannot achieve satisfactory result.



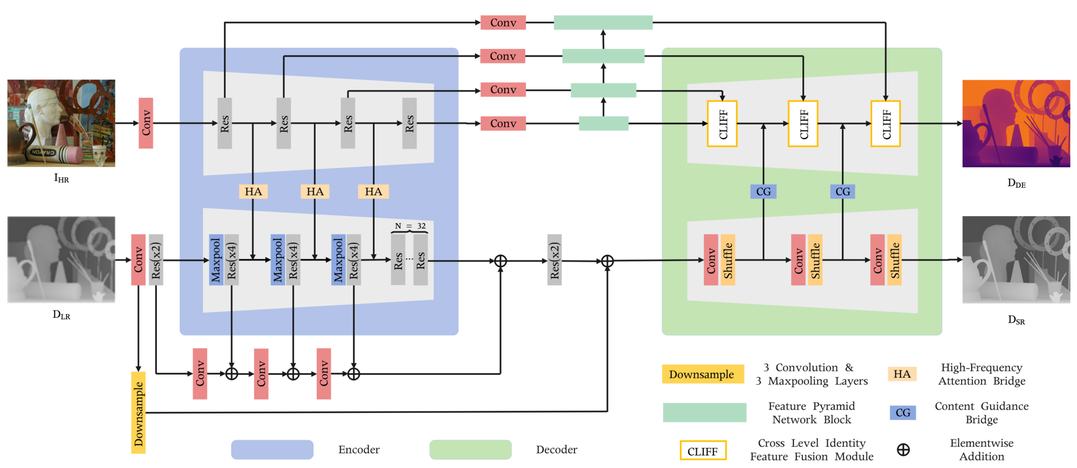
Section 2 (Motivation):

- Monocular depth estimation aims to map a scene from the photometric representation to the geometrical representation.
  1) training datasets for these two tasks can be shared.
  2) orientated by the task of monocular depth estimation, the features learned from RGB image are more suitable for guiding the depth SR, because it can realize this kind of cross-modality information transformation during training.

Section 3 (Contributions):

- We associate the depth map super-resolution and monocular depth estimation in a joint learning framework (BridgeNet) to boost the performance of depth map super-resolution.
- The high-frequency attention bridge in the feature encoding stage transfers the RGB high-frequency information learned from MDENet to DSRNet. Following the principle of simple task guiding difficult task, we propose the content guidance bridge to provide the content guidance learned from DSRNet for MDENet.
- Without the introduction of additional supervision labels, our method achieves competitive performance on benchmark datasets.

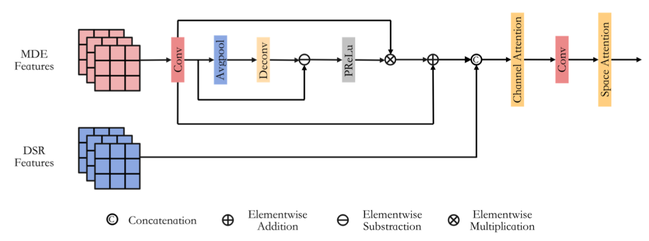
Section 4 (Framework):



Architecture of the proposed BridgeNet, which consists of a depth super-resolution subnetwork (DSRNet), a monocular depth estimation subnetwork (MDENet), a high-frequency attention bridge (HABdg), and a content guidance bridge (CGBdg).

Section 5 (Methodology):

High-Frequency Attention Bridge (HABdg)



$$F^i_{blurred} = deconv\left(avgpool(F^i_{MDE})\right)$$

$$A^i_{hf} = PRelu(F^i_{MDE} - F^i_{blurred})$$

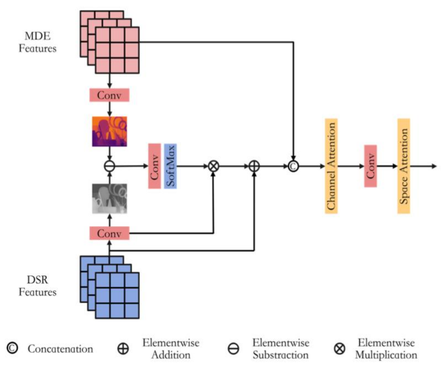
$$F^i_{hg} = F^i_{MDE} + A^i_{hf} \cdot F^i_{MDE}$$

$$F^i_{comp} = [F^i_{DSR}, F^i_{hg}]$$

$$F^i_{ha} = SA\left(conv_{1\times1}\left(CA(F^i_{comp})\right)\right)$$

We first learn the high-frequency attention from the encoder features of the MDENet. Then, it is used to weight the original features to obtain the refined guidance features. After cascading with the features of the corresponding layer of DSRNet, the final output features are obtained through the CA and SA mechanisms.

Content Guidance Bridge (CGBdg)



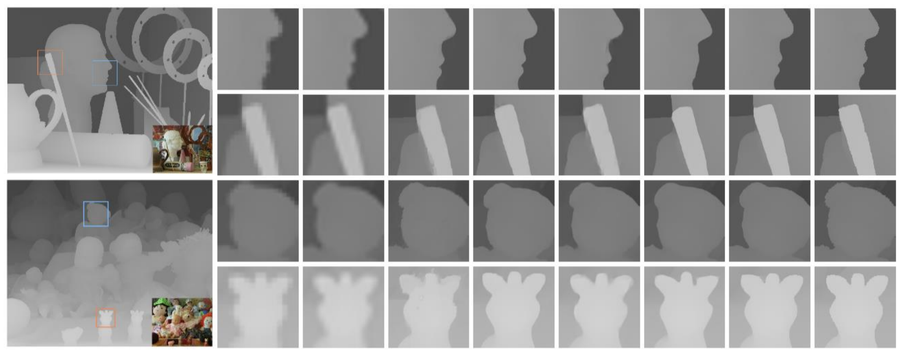
$$M^i_{DSR} = conv_{1\times1}(Fd^i_{DSR}) \qquad M^i_{MDE} = conv_{1\times1}(Fd^i_{MDE})$$

$$W^i_{diff} = softmax\left(conv_{1\times1}(M^i_{DSR} - M^i_{MDE})\right)$$

$$F^i_{cg} = Fd^i_{DSR} + W^i_{diff} * Fd^i_{DSR}$$

We first calculate the difference map between estimated depth map and super-resolved depth map, and then learn the difference. Applying the difference weight to the depth SR encoder features to generate the content guidance for the depth estimation branch. Finally, the depth estimation features and content guidance are concatenated, and fed into the CA and SA modules to produce the output features.

Section 6 (Experiments):



	Art			Books			Dolls			Laundry			Mobius			Reindeer		
	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16
CLMF [23]	0.76	1.44	2.87	0.28	0.51	1.02	0.34	0.60	1.01	0.50	0.80	1.67	0.29	0.51	0.97	0.51	0.84	1.55
JGF [22]	0.47	0.78	1.54	0.24	0.43	0.81	0.33	0.59	1.06	0.36	0.64	1.20	0.25	0.46	0.80	0.38	0.64	1.09
TGV [9]	0.65	1.17	2.30	0.27	0.42	0.82	0.33	0.70	2.20	0.55	1.22	3.37	0.29	0.49	0.90	0.49	1.03	3.05
CDLLC [37]	0.53	0.76	1.41	0.19	0.46	0.75	0.31	0.53	0.79	0.30	0.48	0.96	0.27	0.46	0.79	0.43	0.55	0.98
PB [2]	0.79	0.93	1.98	0.16	0.43	0.79	0.53	0.83	0.99	1.13	1.89	2.87	0.17	0.47	0.82	0.56	0.97	1.89
EG [38]	0.48	0.71	1.35	0.15	0.36	0.70	0.27	0.49	0.74	0.28	0.45	0.92	0.23	0.42	0.75	0.36	0.51	0.95
SRCNN [6]	0.63	1.21	2.34	0.25	0.52	0.97	0.29	0.58	1.03	0.40	0.87	1.74	0.25	0.43	0.87	0.35	0.75	1.47
ATGVNet [26]	0.65	0.81	1.42	0.43	0.51	0.79	0.41	0.52	**0.56**	0.37	0.89	0.94	0.38	0.45	0.80	0.41	0.58	1.01
MSG [16]	0.46	0.76	1.53	0.15	0.41	0.76	0.25	0.51	0.87	0.30	0.46	1.12	0.21	0.43	0.76	0.31	0.52	0.99
DGDIE [11]	0.48	1.20	2.44	0.30	0.58	1.02	0.34	0.63	0.93	0.35	0.86	1.56	0.28	0.58	0.98	0.35	0.73	1.29
DEIN [39]	0.40	0.64	**1.34**	0.22	0.37	0.78	0.22	0.38	0.73	0.23	0.36	0.81	0.20	0.35	0.73	0.26	0.40	0.80
CCFN [35]	0.43	0.72	1.50	0.17	0.36	0.69	0.25	0.46	0.75	0.24	0.41	**0.71**	0.23	0.39	0.73	0.29	0.46	0.95
GSRPT [5]	0.48	0.74	1.48	0.21	0.38	0.76	0.28	0.48	0.79	0.33	0.56	1.24	0.24	0.49	0.80	0.31	0.61	1.07
CTKT [32]	**0.25**	**0.53**	1.44	**0.11**	0.26	0.67	**0.16**	0.36	0.65	**0.16**	0.36	0.76	**0.13**	0.27	0.69	**0.17**	0.35	0.77
BridgeNet (Ours)	0.30	0.58	1.49	0.14	**0.24**	**0.51**	0.19	**0.34**	0.64	0.17	**0.34**	**0.71**	0.15	**0.26**	**0.54**	0.19	**0.31**	**0.70**

Summary/Conclusion:

- In this paper, we explore a joint learning framework that combines the depth map super-resolution and monocular depth estimation to boost the depth SR performance without adding any other supervision labels. The center of this paper is how to design the interaction between the two subnetworks (i.e., DSRNet and MDENet), thus we propose two bridges(i.e., HABdg and CGBdg). Comprehensive experiments show that our method achieves competitive performance, especially for the cases of large scaling factors. Moreover, our network architecture is highly portable and can provide a paradigm for associating the DSR and MDE tasks. Our code and models are released, which can be obtainded at https://rmcong.github.io/proj_BridgeNet.html.