Do End-to-end Stereo Algorithms Under-utilize Information?

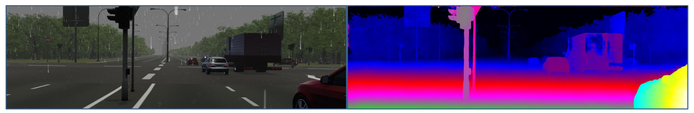
Changjiang Cai, Philippos Mordohai
Stevens Institute of Technology

Code: https://github.com/ccj5351/DAFStereoNets

3DV 2020 International Conference on 3D Vision
Online
November 25-28, 2020

1. Motivation

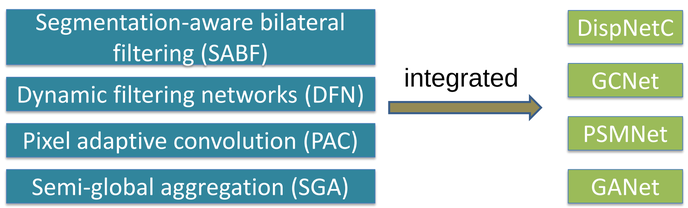
- Cost aggregation mechanisms under-utilize image information
  - ➢ Content-insensitive convolutions
  - ➢ Down- and up-sampling operations in the encoder-decoder architectures
  - ➢ Cost aggregation is not sensitive to pixel similarity, image edges or semantics
  - ➢ Over-smoothing near occlusion boundaries, erroneous predictions in thin structures and textureless regions
- E.g., GCNet on Virtual KITTI 2 validation set



Left Image | Disparity Map by GCNet

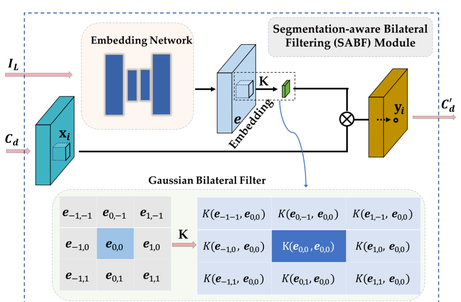
2. Proposal

- Our proposal leverages image context as a signal to dynamically guide the matching process
- Integrating four deep adaptive or guided filters into four existing 2D or 3D convolutional stereo networks



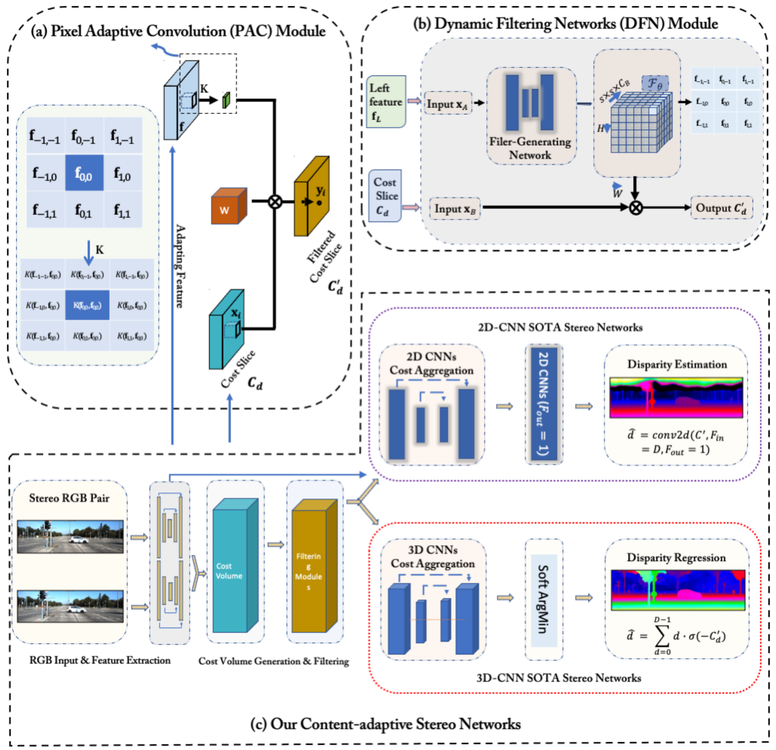
3. Deep Adaptive Filtering

- **SABF**
  - ➢ Embedding learning
  - ➢ SABF filter weights K



$$\mathbf{y}_i = \frac{\sum_{k\in\Omega(i)} \mathbf{x}_k \, K^{sabf}_{i,k}}{\sum_{k\in\Omega(i)} K^{sabf}_{i,k}}$$

$$K^{sabf}_{i,j} = \exp\left(-\frac{||\mathbf{p}_i - \mathbf{p}_j||^2}{2\sigma_s^2} - \frac{||\mathbf{e}_i - \mathbf{e}_j||^2}{2\sigma_r^2}\right)$$

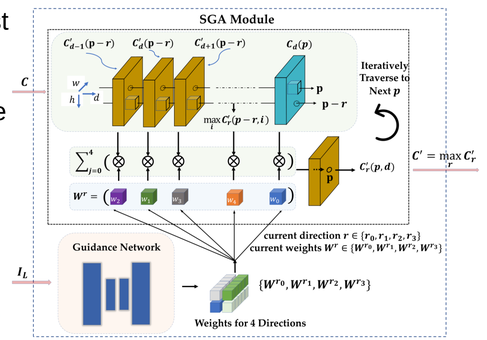


(c) Our Content-adaptive Stereo Networks

- **DFN** $G(u,v) = \mathcal{F}_\theta^{(u,v)}(\mathbf{x}_B(u,v))$
  - ➢ Filters F_θ are dynamically generated on input x_A
  - ➢ F_θ are applied to another input x_B
- **PAC** $\mathbf{y}_i = \sum_{j\in\Omega(i)} K^{pac}(\mathbf{f}_i, \mathbf{f}_j)\, \mathbf{W}[\mathbf{p}_i - \mathbf{p}_j]\, \mathbf{x}_i + \mathbf{b}$
  - ➢ PAC modifies convolution filter W by multiplying it with a position-specific filter K
- **SGA**
  - ➢ SGA aggregates the cost volume over the entire image pixels and each disparity d, in an iterative manner from pixel **p** − **r** to **p** in the direction **r**

$$\mathcal{C}'_{\mathbf{r}}(\mathbf{p},d) = \text{sum}\begin{cases} \mathbf{w}_0(\mathbf{p},\mathbf{r})\cdot\mathcal{C}(\mathbf{p},d) \\ \mathbf{w}_1(\mathbf{p},\mathbf{r})\cdot\mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d) \\ \mathbf{w}_2(\mathbf{p},\mathbf{r})\cdot\mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d-1) \\ \mathbf{w}_3(\mathbf{p},\mathbf{r})\cdot\mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d+1) \\ \mathbf{w}_4(\mathbf{p},\mathbf{r})\cdot\max_i \mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},i) \end{cases}$$

$$\mathcal{C}'(\mathbf{p},d) = \max_{\mathbf{r}} \mathcal{C}'_{\mathbf{r}}(\mathbf{p},d)$$



4. Experimental Results

- **Network Inference Runtime (ms) Comparison**

Filters	DispNetC	PSMNet	GANet	GCNet
W/O	**18.35**	**315.57**	**1894.70**	**146.83**
SABF	24.32	563.42	2488.72	379.37
DFN	28.33	432.32	2041.53	255.20
PAC	25.34	514.91	2383.44	334.73
SGA	489.60	823.00	-	655.18

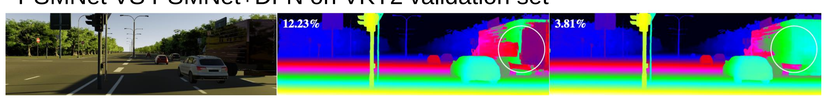
- **Evaluation on Virtual KITTI 2 Validation Set Val-S6**

Filters	DispNetC		PSMNet		GANet		GCNet	
	EPE(px)	≥3 px	EPE(px)	≥3 px	EPE(px)	≥3 px	EPE(px)	≥3 px
W/O	0.70	3.12	0.48	1.96	0.30	1.0563	0.59	2.25
SABF	0.69	3.00	0.44	1.73	**0.28**	**0.97**	0.56	2.23
DFN	**0.599**	**2.791**	**0.39**	**1.69**	0.29	1.0561	0.55	**2.14**
PAC	0.603	2.96	0.52	1.98	0.35	1.47	0.73	2.99
SGA	0.607	2.794	0.42	1.71	-	-	**0.53**	2.29

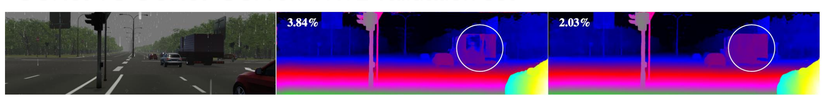
- **Evaluation on KITTI 2015 Validation Set**

Filters	DispNetC		PSMNet		GANet		GCNet	
	noc	all	noc	all	noc	all	noc	all
W/O	2.59	3.02	1.46	1.60	**0.97**	**1.10**	2.06	2.64
SABF	2.26	2.63	1.28	1.40	1.07	1.17	1.76	2.10
DFN	2.37	2.78	1.23	1.34	0.99	1.11	1.70	2.08
PAC	2.38	2.72	1.29	1.48	1.13	1.23	1.71	2.03
SGA	**1.90**	**2.18**	**1.17**	**1.32**	-	-	**1.69**	**1.91**

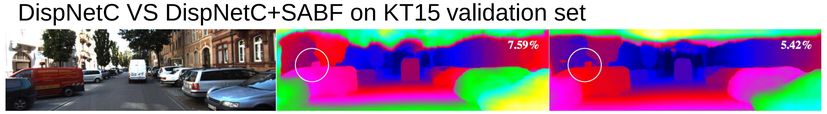
- **Qualitative Results: Input/Baselines/Ours**
  - ➢ PSMNet VS PSMNet+DFN on VKT2 validation set



  - ➢ GCNet VS GCNet+SGA on VKT2 validation set



  - ➢ DispNetC VS DispNetC+SABF on KT15 validation set



  - ➢ GANet VS GANet+PAC on KT15 validation set