



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





Changjiang Cai



Philippos Mordohai

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







Feature Extraction **Cost Volume** Generation

Cost Volume Aggregation



RGB Input & Feature Extraction



Cost Volume in 2D/3D Stereo Networks

Cost Volume Aggregation





Cost Volume Generation RGB Input & Feature Extraction

Cost Volume Aggregation

Cost Volume Generation RGB Input & Feature Extraction

Disparity Regression

Cost Volume Aggregation

Cost Volume Generation RGB Input & Feature Extraction

Disparity Regression

Cost Volume Aggregation

Cost Aggregation: Encoder-decoder

Cost Volume Generation RGB Input & Feature Extraction

Cost Volume Aggregation

- E.g., GCNet on Virtual KITTI 2 validation set

- > 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
- Content-insensitive convolutions

Do SOTA Stereo Networks Under-utilize Information? Cost aggregation mechanisms under-utilize image information

- structures and textureless regions

Deep Adaptive Filtering in End-to-end Stereo

bilateral filtering (SABF) Dynamic filtering networks (DFN) Pixel adaptive convolution (PAC)Semi-global aggregation (SGA)

dynamically guide the matching process convolutional stereo networks

¹N. Mayer et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. CVPR'16 ² A. Kendall et al. End-to-end learning of geometry and context for deep stereo regression. ICCV'17 ³ J.-R. Chang and Y.-S. Chen. Pyramid stereo matching network. CVPR'18 ⁴ F. Zhang et al. Ga-net: Guided aggregation net for end-to-end stereo matching. CVPR'19

• Our proposal can leverage image context as a signal to >Integrate four deep adaptive or guided filters into four existing 2D or 3D

Deep Adaptive Filtering: SABF⁵

⁵ A. W. Harley et al. Segmentation-aware convolutional networks using local attention masks. ICCV'17

Deep Adaptive Filtering: SABF⁵

⁵ A. W. Harley et al. Segmentation-aware convolutional networks using local attention masks. ICCV'17

$K(e_{-1-1}, e_{0,0})$	$K(e_{0,-1}, e_{0,0})$	$K(e_{1,-1}, e_{0,0})$
$K(e_{-1,0}, e_{0,0})$	$\mathrm{K}(oldsymbol{e}_{0,0}$, $oldsymbol{e}_{0,0})$	$K(e_{1,0}, e_{0,0})$
$K(e_{-1,1}, e_{0,0})$	$K(e_{0,1}, e_{0,0})$	$K(e_{1,1}, e_{0,0})$

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) Pixel adaptive convolution (PAC) Semi-global aggregation SGA

 $\left(\frac{||\mathbf{p}_{i} - \mathbf{p}_{j}||^{2}}{2\sigma^{2}} - \frac{||\mathbf{e}_{i} - \mathbf{e}_{j}||^{2}}{2\sigma_{\pi}^{2}}\right)$

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) Pixel adaptive convolution (PAC) Semi-global aggregation (SGA)

 $\left(\frac{||\mathbf{p}_i - \mathbf{p}_j||^2}{2\sigma_s^2} - \frac{||\mathbf{e}_i - \mathbf{e}_j||^2}{2\sigma_r^2}\right)$

⁶X. Jia et al. Dynamic filter networks. NeurIPS'16

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) Pixel adaptive convolution (PAC) Semi-global aggregation (SGA)

Deep Adaptive Filtering: DFN

• First, filters F_{θ} in the DFN are dynamically generated by a separate filter generating network conditioned on an input x_A .

• Second, filters F_{θ} are applied to another input x_{B} via the dynamic filtering layer DFN response depends on input content and also its spatial position

$G(u, v) = \mathcal{F}_{\theta}^{(u,v)}(\mathbf{x}_B(u, v))$

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN Pixel adaptive convolution (PAC) Semi-global aggregation (SGA)

Deep Adaptive Filtering: DFN

$G(u, v) = \mathcal{F}_{\theta}^{(u,v)}(\mathbf{x}_B(u, v))$

• First, filters F_{θ} in the DFN are dynamically generated by a separate filter generating network conditioned on an input x_A .

• Second, filters F_{θ} are applied to another input x_{B} via the dynamic filtering layer DFN response depends on input content and also its spatial position

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) Pixel adaptive convolution (PAC) Semi-global aggregation (SGA)

Deep Adaptive Filtering: PAC⁷

, ,	(a)) Pi	.X(el Ac	lapt
		f _1,-	-1	f _{0,-1}	f _{1,-1}
		f _1,	0	f _{0,0}	f _{1,0}
		f _1,	1	f _{0,1}	f _{1,1}
				↓ K	
	К(f _1-1	, f _{0,0})	K((f _{0,-1} , f _{0,0})) K(f _{1,-}
	К(f _1,0,	f _{0,0})	K	$f(\mathbf{f}_{0,0},\mathbf{f}_{0,0})$	K(f _{1,0}
	K(f _1,1,	f _{0,0})	K	(f _{0,1} , f _{0,0})	<i>K</i> (f _{1,2}
1					

 Conventional convolution filter W is spatially invariant • PAC modifies this convolution filter W at each position by multiplying it with a position-specific filter K (i.e., the Gaussian kernel) • The adapting features **f** are the deep features extracted from the left image

$\mathbf{y}_i = \sum K^{pac}(\mathbf{f}_i, \mathbf{f}_j) \mathbf{W}[\mathbf{p}_i - \mathbf{p}_j] \mathbf{x}_i + \mathbf{b}$

⁷ H. Su et al. Pixel-adaptive convolutional neural networks. CVPR'19

(a) Pixel Adapt					
	f _1,-	-1	f _{0,-1}	f _{1,-1}	
	f _1,	,0	f _{0,0}	f _{1,0}	
	f _1,	,1	f _{0,1}	f _{1,1}	
			↓ K		
K(f _1-1,	, f _{0,0})	K(f _{0,-1} , f _{0,0}) K(f _{1,-}	
K(f _1,0,	f _{0,0})	K	$(\mathbf{f}_{0,0}$, $\mathbf{f}_{0,0})$	<i>K</i> (f _{1,}	
K(f _1,1,	f _{0,0})	K	(f _{0,1} , f _{0,0})	<i>K</i> (f _{1,}	

$j \in \Omega(i)$ Conventional convolution filter W is spatially invariant position-specific filter K (i.e., the Gaussian kernel)

tive Convolution (PAC) Module

• PAC modifies this convolution filter W at each position by multiplying it with a • The adapting features **f** are the deep features extracted from the left image

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) convolution (PAC) Semi-global aggregation (SGA)

/	, (a)) Pi	X	el Ac	lapt
/		f_ _1,-	-1	f _{0,-1}	f _{1,-1}
		f _1,	0	f _{0,0}	f _{1,0}
		f _1,	1	f _{0,1}	f _{1,1}
				↓ K	
	К(f _1-1	, f _{0,0})	K([f _{0,-1} , f _{0,0}]) K(f _{1,-}
	K(f _1,0,	f _{0,0})	K	$({f f}_{0,0},{f f}_{0,0})$	K(f _{1,}
1	K(f _1,1,	f _{0,0})	K	(f _{0,1} , f _{0,0})	K(f _{1,2}

position-specific filter K (i.e., the Gaussian kernel)

• The adapting features **f** are the deep features extracted from the left image

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) convolution (PAC) Semi-global aggregation (SGA)

(a) Pixel Adapt					
	f _1,-	-1	f _{0,-1}	f _{1,-1}	
	f _1,	,0	f _{0,0}	f _{1,0}	
	f _1,	,1	f _{0,1}	f _{1,1}	
			↓ K		
<i>K</i> (f ₋₁₋₁	, f _{0,0})	K((f _{0,-1} , f _{0,0})) K(f _{1,-}	
K(f _1,0,	, f _{0,0})	K	$f(\mathbf{f}_{0,0},\mathbf{f}_{0,0})$	<i>K</i> (f _{1,}	
K(f _1,1,	f _{0,0})	K	(f _{0,1} , f _{0,0})	<i>K</i> (f _{1,2}	

 Conventional convolution filter W is spatially invariant • PAC modifies this convolution filter W at each position by multiplying it with a position-specific filter K (i.e., the Gaussian kernel) • The adapting features **f** are the deep features extracted from the left image

tive Convolution (PAC) Module

$\mathbf{y}_i = \sum K^{pac}(\mathbf{f}_i, \mathbf{f}_j) \mathbf{W}[\mathbf{p}_i - \mathbf{p}_j] \mathbf{x}_i + \mathbf{b}$

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) convolution (PAC) Semi-global aggregation (SGA)

⁸ F. Zhang et al. Ga-net: Guided aggregation net for end-to-end stereo matching. CVPR'19

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) Pixel adaptive convolution (PAC)

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

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) Pixel adaptive convolution (PAC)

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

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) Pixel adaptive convolution (PAC)

 $\begin{pmatrix} \mathbf{w}_0(\mathbf{p},\mathbf{r}) \cdot \mathcal{C}(\mathbf{p},d) \\ \mathbf{w}_1(\mathbf{p},\mathbf{r}) \cdot \mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d) \end{cases}$ $C'_{\mathbf{r}}(\mathbf{p},d) = \operatorname{sum} \left\langle \mathbf{w}_{2}(\mathbf{p},\mathbf{r}) \cdot C'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d-1) \right\rangle$ $\mathbf{w}_3(\mathbf{p},\mathbf{r})\cdot \mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d+1)$ $\mathbf{w}_4(\mathbf{p},\mathbf{r})\cdot\max_{\mathbf{r}}\mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},i)$

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) Pixel adaptive convolution (PAC)

 $(\mathbf{w}_0(\mathbf{p},\mathbf{r})\cdot \mathcal{C}(\mathbf{p},d))$ $\mathbf{w}_1(\mathbf{p},\mathbf{r})\cdot \mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d)$ $\mathcal{C}'_{\mathbf{r}}(\mathbf{p},d) = \operatorname{sum} \left\{ \mathbf{w}_{2}(\mathbf{p},\mathbf{r}) \cdot \mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d-1) \right\}$ $\mathbf{w}_3(\mathbf{p},\mathbf{r})\cdot \mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d+1)$ $\mathbf{w}_4(\mathbf{p},\mathbf{r})\cdot\max_i \mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},i)$

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) Pixel adaptive convolution (PAC) Semi-global aggregation

 $(\mathbf{w}_0(\mathbf{p},\mathbf{r})\cdot \mathcal{C}(\mathbf{p},d))$ $\mathbf{w}_1(\mathbf{p},\mathbf{r})\cdot \mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d)$ $\mathcal{C}'_{\mathbf{r}}(\mathbf{p},d) = \operatorname{sum} \langle \mathbf{w}_2(\mathbf{p},\mathbf{r}) \cdot \mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d-1) \rangle$ $\mathbf{w}_3(\mathbf{p},\mathbf{r})\cdot \mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d+1)$ $\mathbf{w}_4(\mathbf{p},\mathbf{r})\cdot\max_{\mathbf{r}}\mathcal{C}'_{\mathbf{r}}(\mathbf{p}-\mathbf{r},i)$

Iteratively **Traverse to** Next *p*

 $C' = \max C'_r$

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) Pixel adaptive convolution (PAC) Semi-global aggregation

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

Deep Adaptive Filtering Stereo Networks tive Convolution (PAC) Module ----take PAC filter as an example 1, **f**_{0,0}) o, **f**_{0,0}) 05 C'_d

(a)) Pi	Xe	el Ac	lapt
	f_ _1,-	-1	f _{0,-1}	f _{1,-1}
	f _1,	0	f _{0,0}	f _{1,0}
	f _1,	1	f _{0,1}	f _{1,1}
			↓ K	
К(f _1-1	, f _{0,0})	K(f _{0,-1} , f _{0,0}) K(f _{1,-}
$K(\mathbf{f}_{-1,0},$	f _{0,0})	K	$(\mathbf{f}_{0,0}$, $\mathbf{f}_{0,0})$	K(f _{1,0}
K(f _1,1,	f _{0,0})	K	(f _{0,1} , f _{0,0})	<i>K</i> (f _{1,2}

Segmentation-aware bilateral filtering (SABF) Dynamic filtering networks (DFN) Pixel adaptive convolution (PAC) Semi-global aggregation (SGA)

Deep Adaptive Filtering Stereo Networks tive Convolution (PAC) Module -----intered a .₁, **f**_{0,0}) Fea 2D CNNs ,0, **f**_{0,0}) 05 N 9 C'_d **Disparity Estimation Cost Aggregation** CNNs sice 8 0 , **f**_{0,0}) $(F_{ou}$ $\int \mathcal{O}^{\mathcal{O}} C_d$ $\widehat{d} = conv2d(C', F_{in})$ $= D, F_{out} = 1$ **Stereo RGB Pair 3D CNNs Disparity Regression Cost Aggregation** Soft Cost Filtering Volume Modules \triangleright gMin $\cdot \sigma(-C'_d)$ **RGB Input & Feature Extraction Cost Volume Generation & Filtering** (b) Our Content-adaptive Stereo Networks

, (a)) Pi	X	el Ac	lapti
	f_ _1,-	-1	f _{0,-1}	f _{1,-1}
	f_ 1,	0	f _{0,0}	f _{1,0}
	f_ 1,	1	f _{0,1}	f _{1,1}
			↓ K	
K(f _1-1	, f _{0,0})	K((f _{0,-1} , f _{0,0}) K(f _{1,-}
K(f _1,0,	f _{0,0})	K	$(\mathbf{f}_{0,0}$, $\mathbf{f}_{0,0})$	<i>K</i> (f _{1,0}
K(f _1,1,	f _{0,0})	K	(f _{0,1} , f _{0,0})	<i>K</i> (f _{1,1}

1	(a) Pixel Adapti						
		f_ _1,-	-1	f _{0,-1}]	f _{1,—1}	
		f _1,	0	f _{0,0}		f _{1,0}	
		f_ 1,	1	f _{0,1}		f _{1,1}	
				ŀ	$\boldsymbol{\zeta}$		
K	(f _1−1	, f _{0,0})	K((f _{0,-1} , f _{0,}	.0)	K(f _{1,-}	
K	ζ(f _ _{1,0} ,	f _{0,0})	K	(f_{0,0} , f_{0,0}))	К(f _{1,}	(
K	ζ(f _ _{1,1} ,	f _{0,0})	K	(f _{0,1} , f _{0,0})	<i>K</i> (f _{1,}	1

1	(a) Pixel Adapti					
		f _1,-	-1	f _{0,-1}	f	1,—1
		f _1,	0	f _{0,0}	1	f _{1,0}
		f _1,	1	f _{0,1}	1	f _{1,1}
				V K		
	К(f _1-1	, f _{0,0})	K(f _{0,-1} , f _{0,0})	К(f _{1,-}
	$K(\mathbf{f}_{-1,0},$	f _{0,0})	K	$({f f}_{0,0}$, ${f f}_{0,0})$		<i>K</i> (f _{1,0}
	K(f _1,1,	f _{0,0})	K	(f _{0,1} , f _{0,0})		<i>K</i> (f _{1,1}
1						

tive Convolution (PAC) Module

(a)) Pixe	el Ac	lapti
	f _1,-1	f _{0,-1}	f _{1,-1}
	f _1,0	f _{0,0}	f _{1,0}
	f _1,1	f _{0,1}	f _{1,1}
		↓ K	

$K(\mathbf{f}_{-1-1}, \mathbf{f}_{0,0})$	$K(\mathbf{f}_{0,-1}, \mathbf{f}_{0,0})$	<i>K</i> (f _{1,} _
$K(\mathbf{f}_{-1,0}, \mathbf{f}_{0,0})$	K(f _{0,0} , f _{0,0})	<i>K</i> (f _{1,0}
$K(\mathbf{f}_{-1,1}, \mathbf{f}_{0,0})$	$K(\mathbf{f}_{0,1}, \mathbf{f}_{0,0})$	<i>K</i> (f _{1,1}

(a)) Pixe	el Ac	lapti
	f _1,-1	f _{0,-1}	f _{1,-1}
	f _1,0	f _{0,0}	f _{1,0}
	f _1,1	f _{0,1}	f _{1,1}
		↓ K	

$K(\mathbf{f}_{-1-1}, \mathbf{f}_{0,0})$	$K(\mathbf{f}_{0,-1}, \mathbf{f}_{0,0})$	<i>K</i> (f _{1,} _
$K(\mathbf{f}_{-1,0}, \mathbf{f}_{0,0})$	K(f _{0,0} , f _{0,0})	<i>K</i> (f _{1,0}
$K(\mathbf{f}_{-1,1}, \mathbf{f}_{0,0})$	$K(\mathbf{f}_{0,1}, \mathbf{f}_{0,0})$	<i>K</i> (f _{1,1}

(b) Our Content-adaptive Stereo Networks

(a) Pixel Adapti						
	f _1,-1	f _{0,-1}	f _{1,-1}			
	f _1,0	f _{0,0}	f _{1,0}			
	f _1,1	f _{0,1}	f _{1,1}			
		↓ K				

$K(\mathbf{f}_{-1-1}, \mathbf{f}_{0,0})$	$K(\mathbf{f}_{0,-1}, \mathbf{f}_{0,0})$	<i>K</i> (f _{1,} _
$K(\mathbf{f}_{-1,0}, \mathbf{f}_{0,0})$	K(f _{0,0} , f _{0,0})	<i>K</i> (f _{1,0}
$K(\mathbf{f}_{-1,1}, \mathbf{f}_{0,0})$	$K(\mathbf{f}_{0,1}, \mathbf{f}_{0,0})$	<i>K</i> (f _{1,1}

ive Convolution (PAC) Module

, , , , , , , , , , , , , , , , , , ,	((a) Pixel Adapt							
		f _1,-	-1	f _{0,-1}	f _{1,-1}			
		f _1,	0	f _{0,0}	f _{1,0}			
		f _1,	1	f _{0,1}	f _{1,1}			
				↓ K				
	K(f _1-1	, f _{0,0})	K([f _{0,-1} , f _{0,0}]) K(f _{1,-}			
	$K(\mathbf{f}_{-1,0},$	f _{0,0})	K	$({f f}_{0,0},{f f}_{0,0})$	K(f _{1,0}			
	K(f _1,1,	f _{0,0})	K	(f _{0,1} , f _{0,0})	<i>K</i> (f _{1,2}			
``								

Datasets

- Scene Flow (SF)
- KITTI 2015 (KT15)
- Virtual KITTI 2 (VKT2)
- VKT2 or KT15

• a synthetic clone of the real KITTI dataset Pre-training on SF, and finetuning on

• a real dataset of street views, containing 200 training stereo image pairs with sparsely labeled disparity from LiDAR data

• a synthetic dataset of 35k training images with dense ground truth disparity maps

KITTI 2015 (KT15)

Scene Flow (SF)

Driving

The order of the filtering modules according to increasing inference runtime is DFN < PAC < SABF < SGA

DispNetC

Network Inference Runtime Comparison

Runtime Comparison (in ms)

■ W/O ■ SABF ■ DFN ■ PAC ■ SGA

Evaluation on Virtual KITTI 2 Val-S6

Evaluation on KITTI 2015

Bad-3 Error on KITTI 2015 Validation Set

Evaluation on Synthetic Dataset: VKT2 PSMNet VS PSMNet + DFN

Evaluation on Real Dataset: KT15 • DispNetC VS DispNetC + SABF

Summary

- Novel deep adaptive filtering architectures for end-to-end stereo matching segmentation-aware bilateral filtering (SABF) o dynamic filtering networks (DFN) o pixel adaptive convolution (PAC) semi-global aggregation (SGA)
- Further progress is possible by leveraging image context as a signal to dynamically guide the matching process
- SGA typically achieves the highest accuracy among them, at the cost of more parameters and runtime
- Integrating even the smaller filtering modules leads to 10% decreases in error
- Code is available at https://github.com/ccj5351/DAFStereoNets