STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

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

Changjiang Cai

Philippos Mordohai

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

Deep Stereo Networks

Cost Volume in 2D/3D Stereo Networks

Cost Volume in 2D/3D Stereo Networks

Disparity Regression

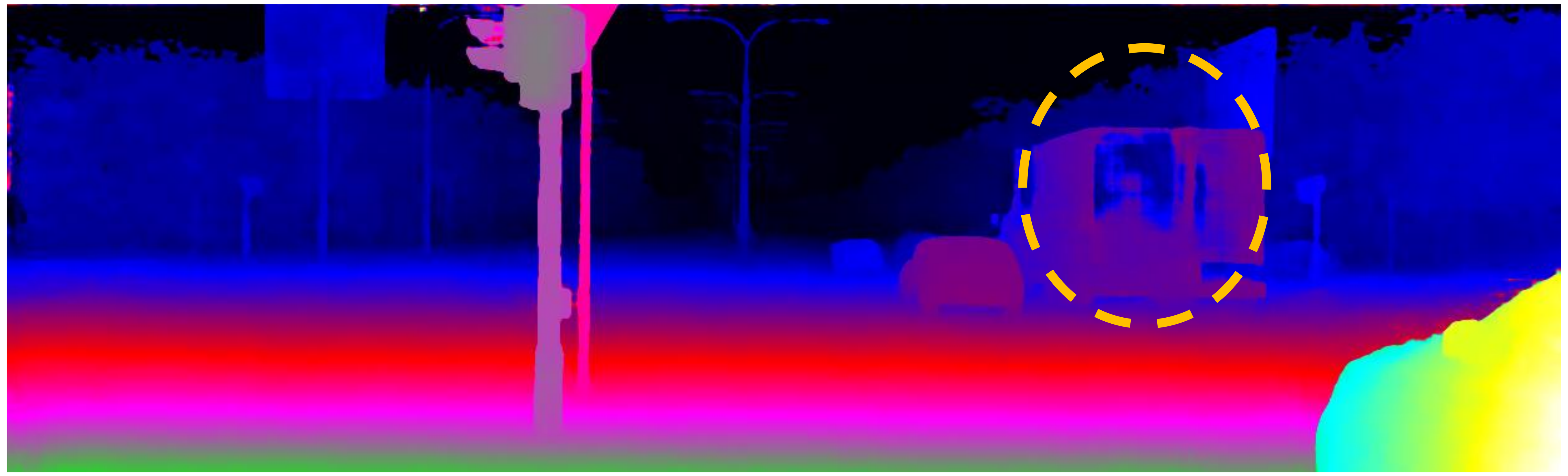
Disparity Regression

Disparity Regression

Cost Aggregation: Encoder-decoder

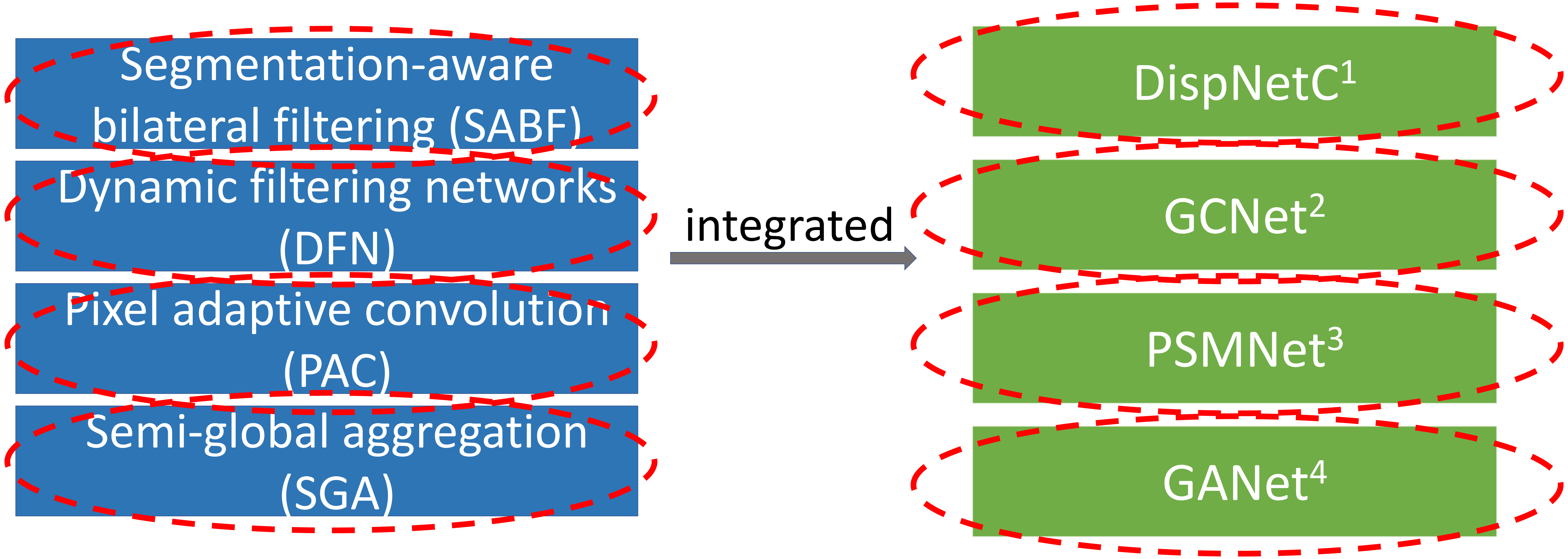
Do SOTA Stereo Networks Under-utilize Information?

- 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



Deep Adaptive Filtering in End-to-end Stereo

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



[1] N. Mayer et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. CVPR'16

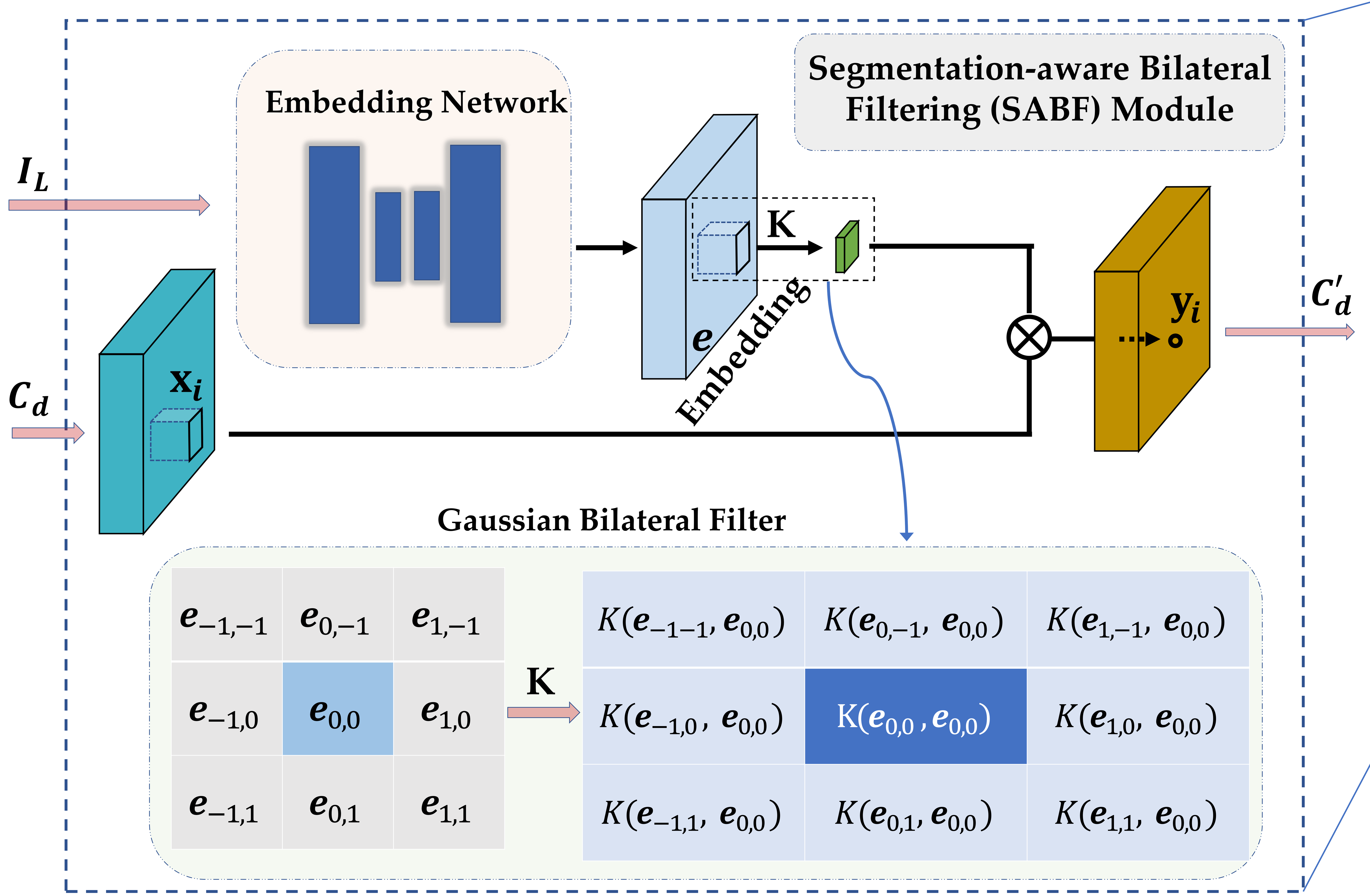
[2] A. Kendall et al. End-to-end learning of geometry and context for deep stereo regression. ICCV'17

[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

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

Deep Adaptive Filtering: SABF[5]



EBF Module

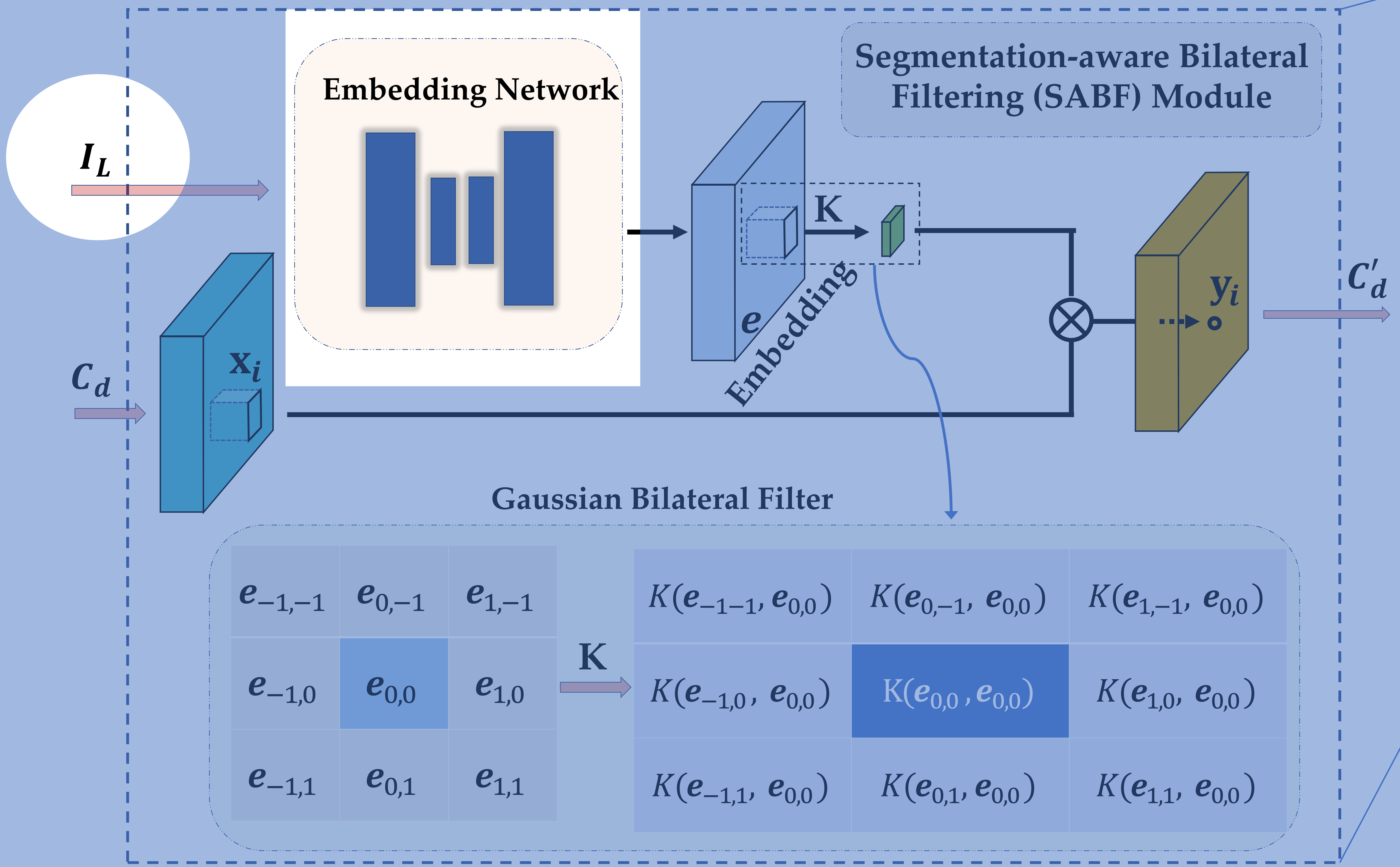
- Segmentation-aware bilateral filtering (SABF)
- Dynamic filtering networks (DFN)
- Pixel adaptive convolution (PAC)
- Semi-global aggregation (SGA)

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

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

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

Deep Adaptive Filtering: SABF[5]



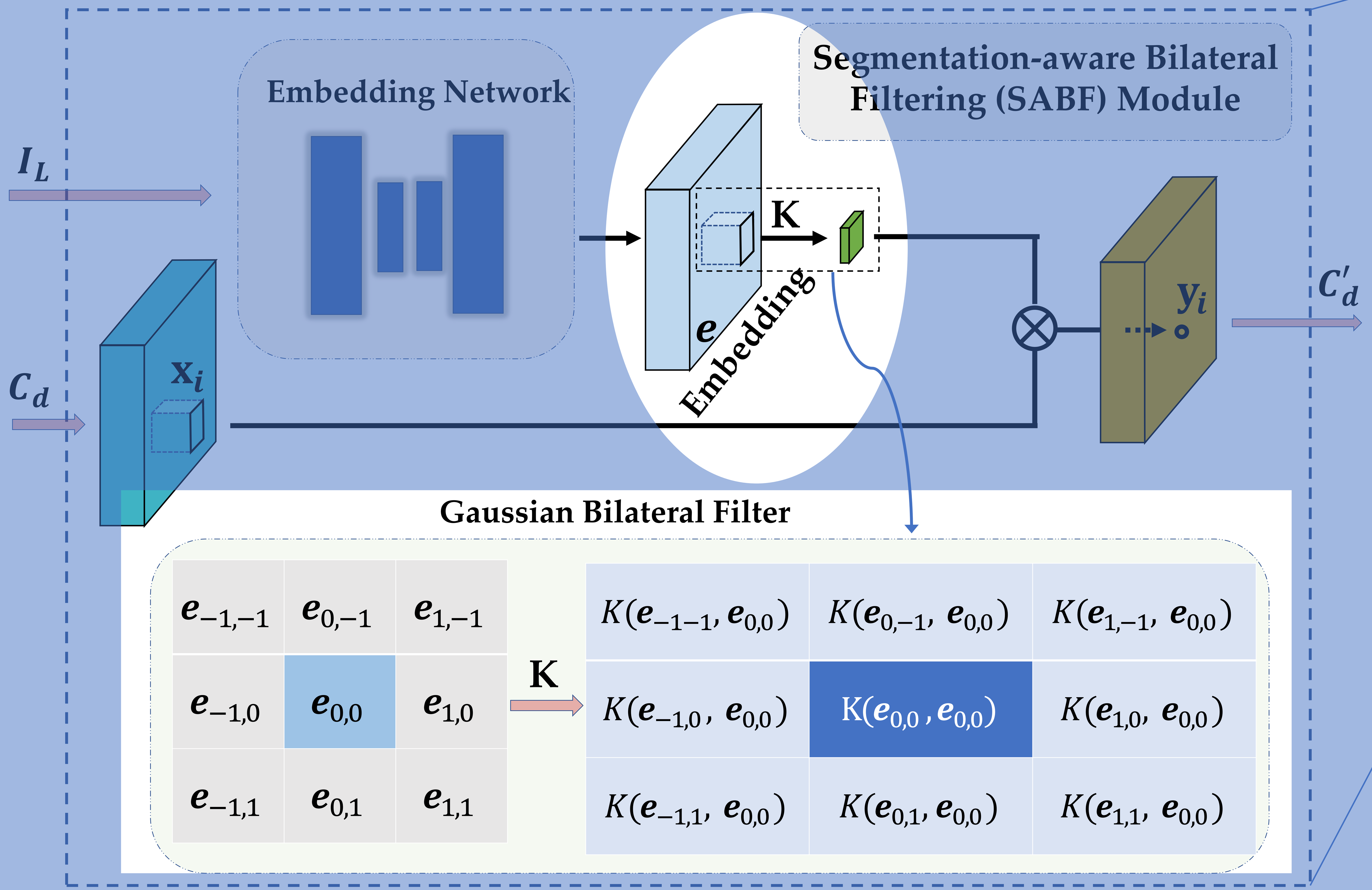
EBF Module

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

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

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

Deep Adaptive Filtering: SABF



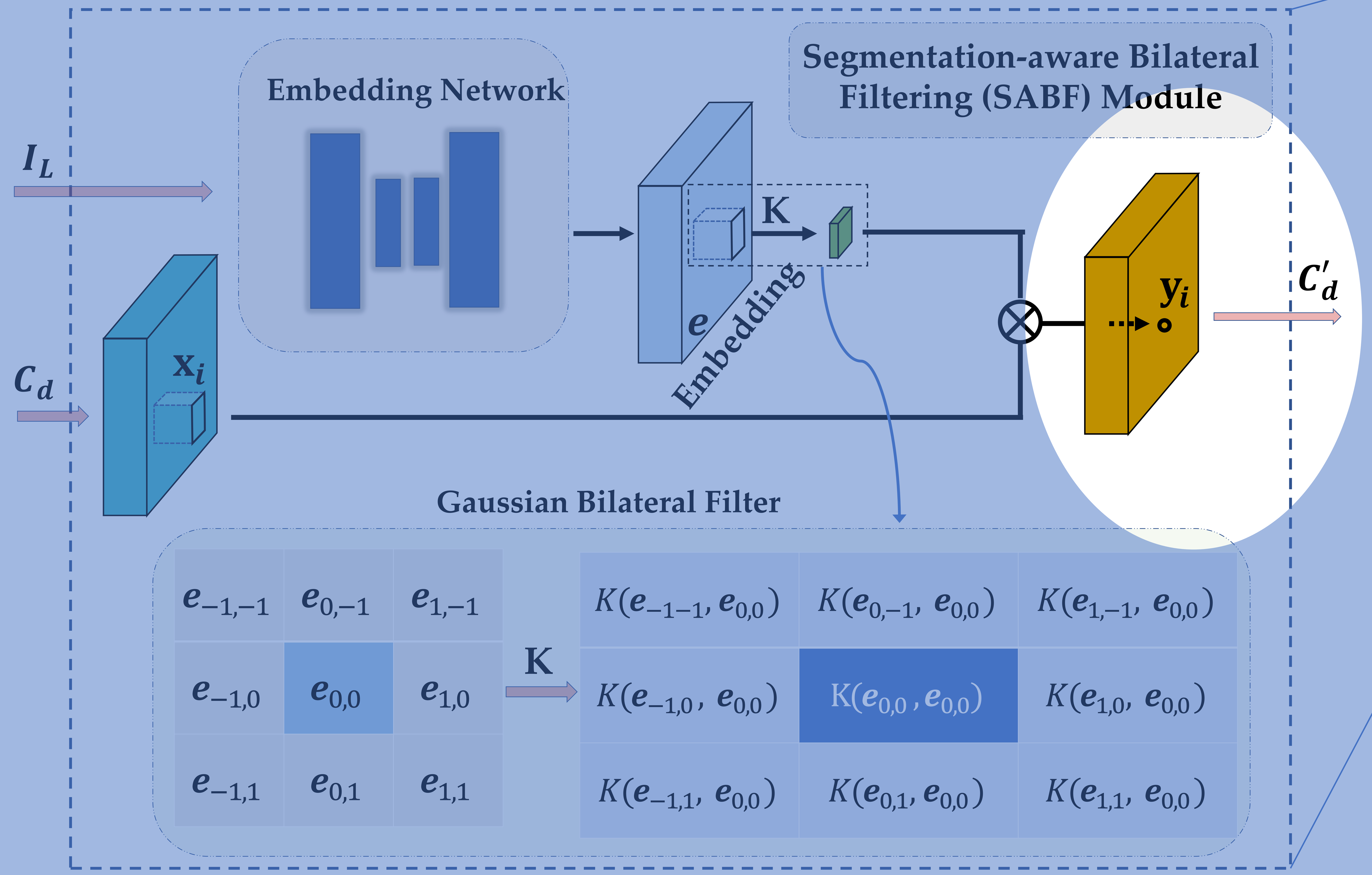
EBF Module

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

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

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

Deep Adaptive Filtering: SABF



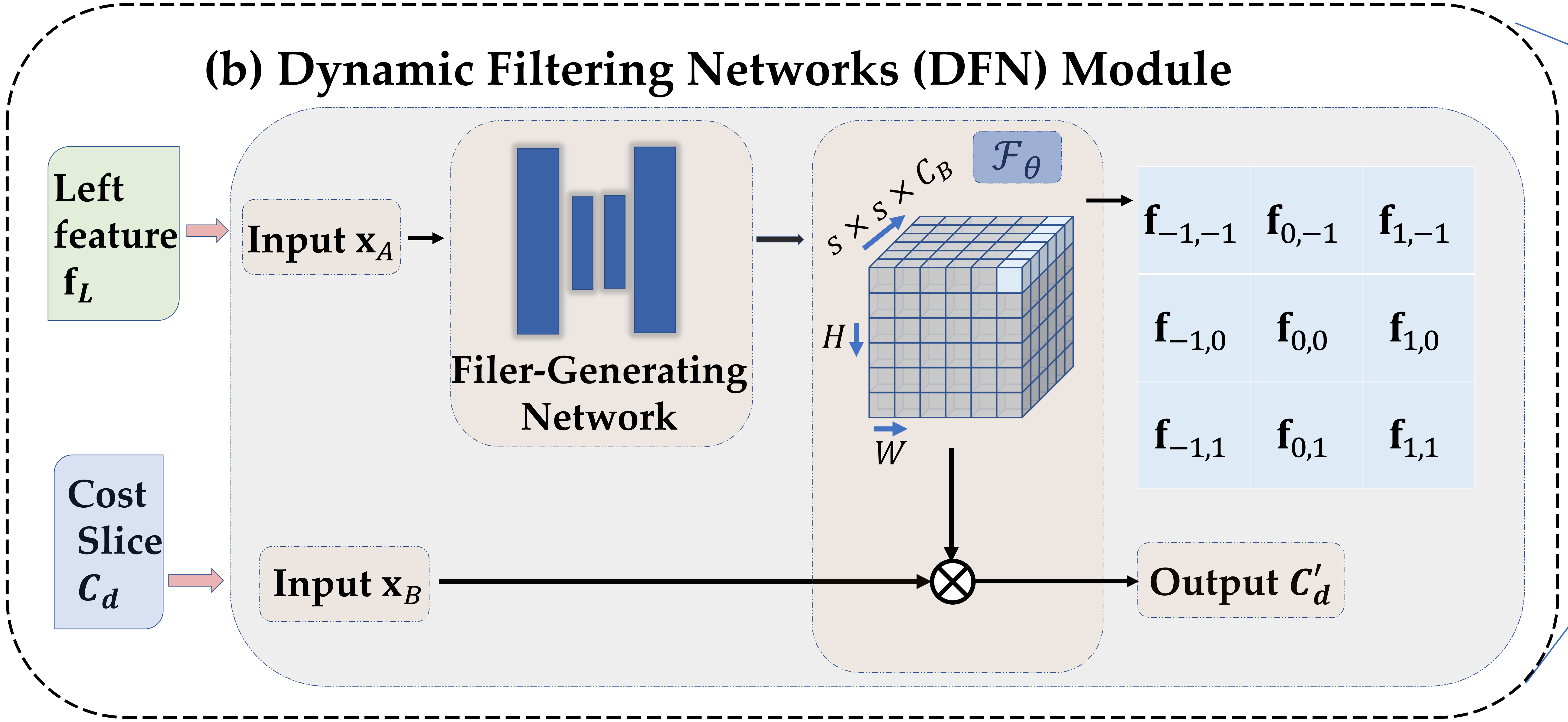
EBF Module

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

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

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

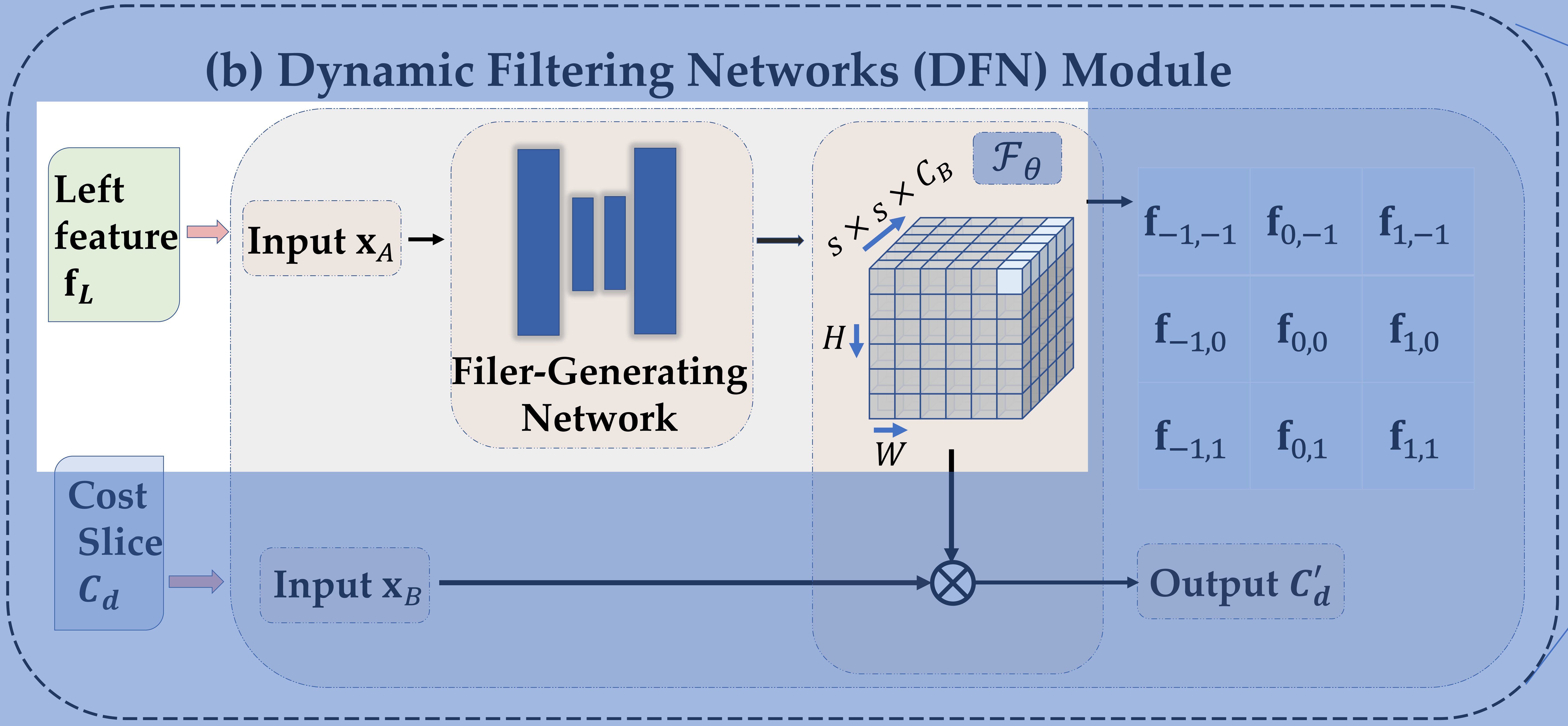
Deep Adaptive Filtering: DFN[6]



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

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

Deep Adaptive Filtering: DFN

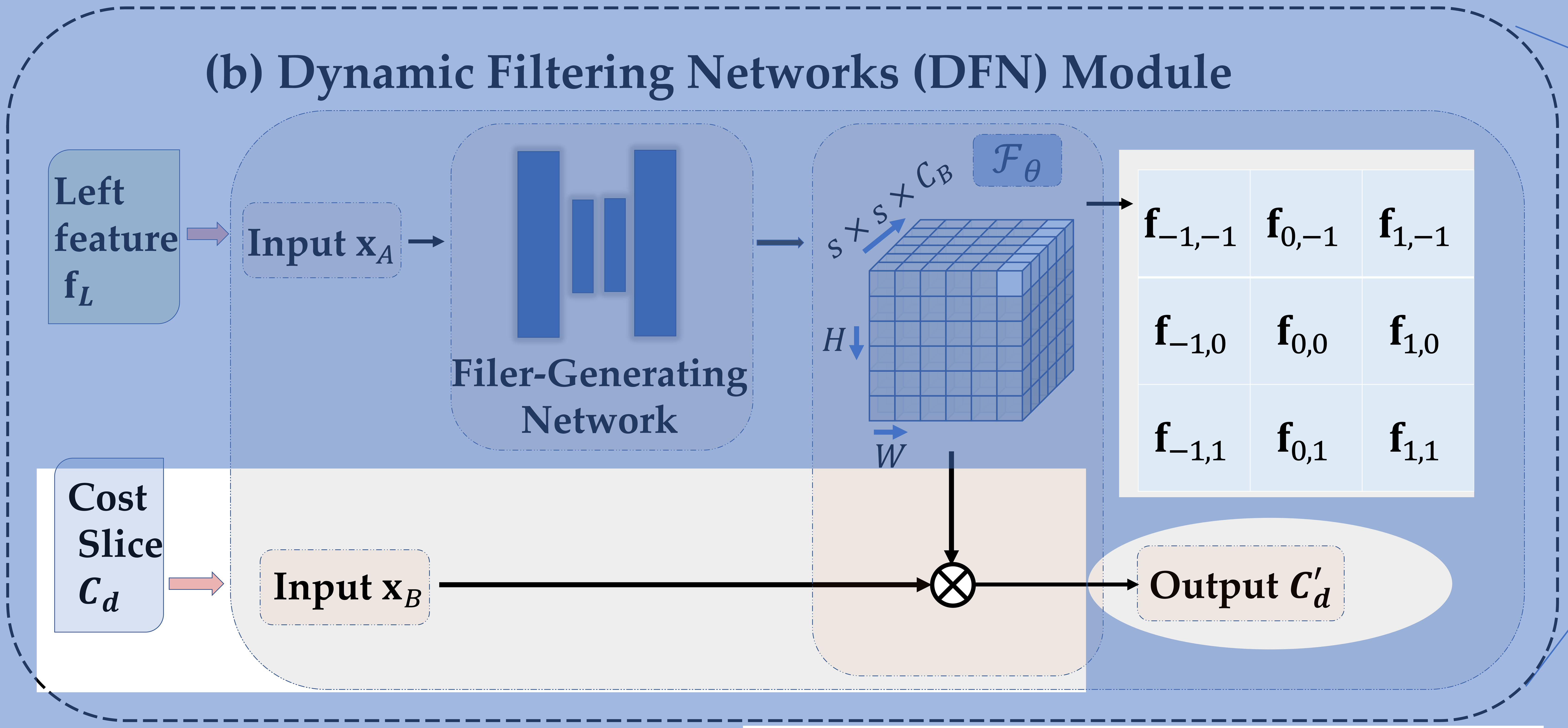


- Segmentation-aware bilateral filtering (SABF)
- Dynamic filtering networks (DFN)
- Pixel adaptive convolution (PAC)
- Semi-global aggregation (SGA)

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

- 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

Deep Adaptive Filtering: DFN



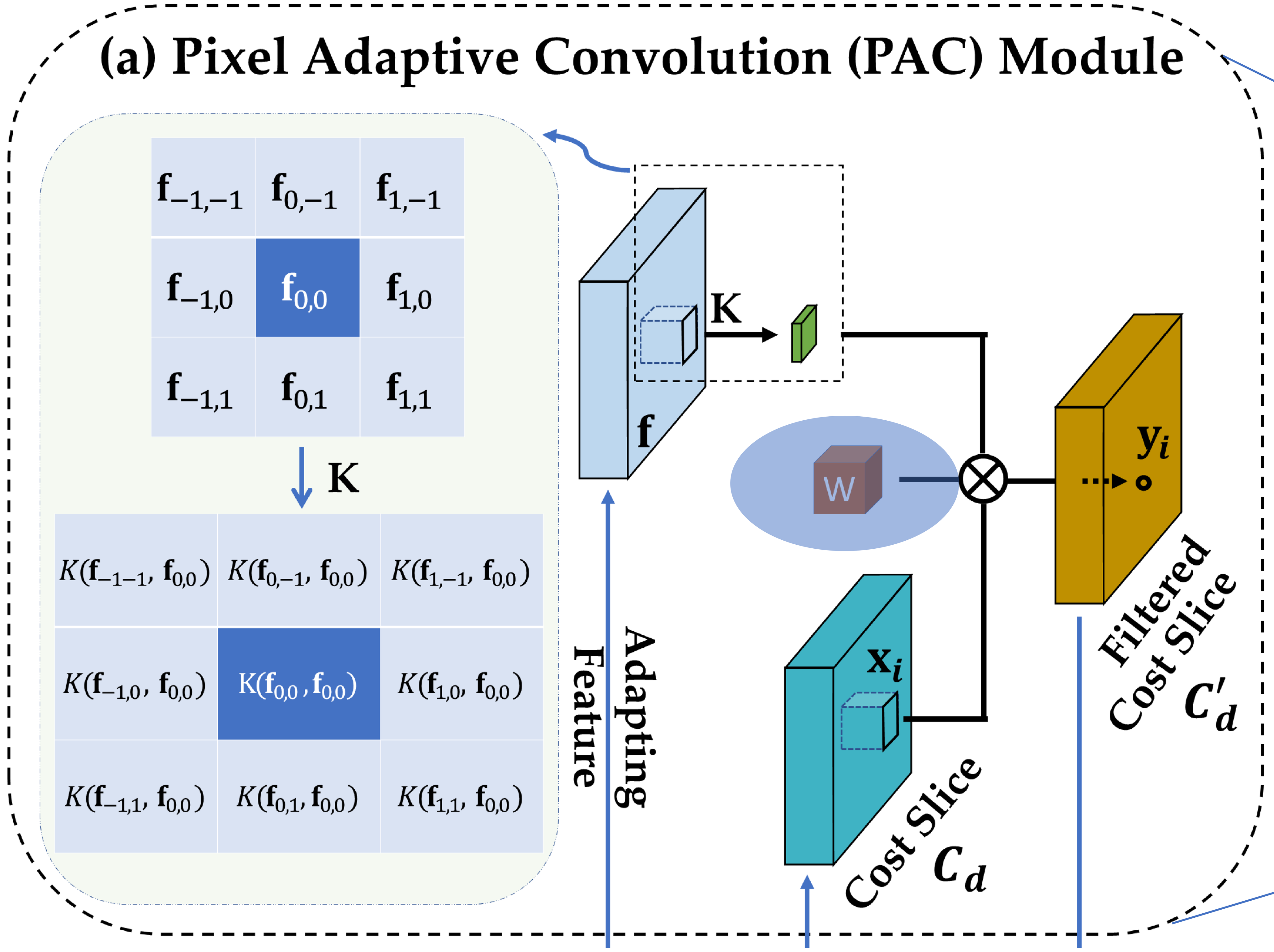
- Segmentation-aware bilateral filtering (SABF)
- Dynamic filtering networks (DFN)
- Pixel adaptive convolution (PAC)
- Semi-global aggregation (SGA)

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

- 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

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

Deep Adaptive Filtering: PAC[7]

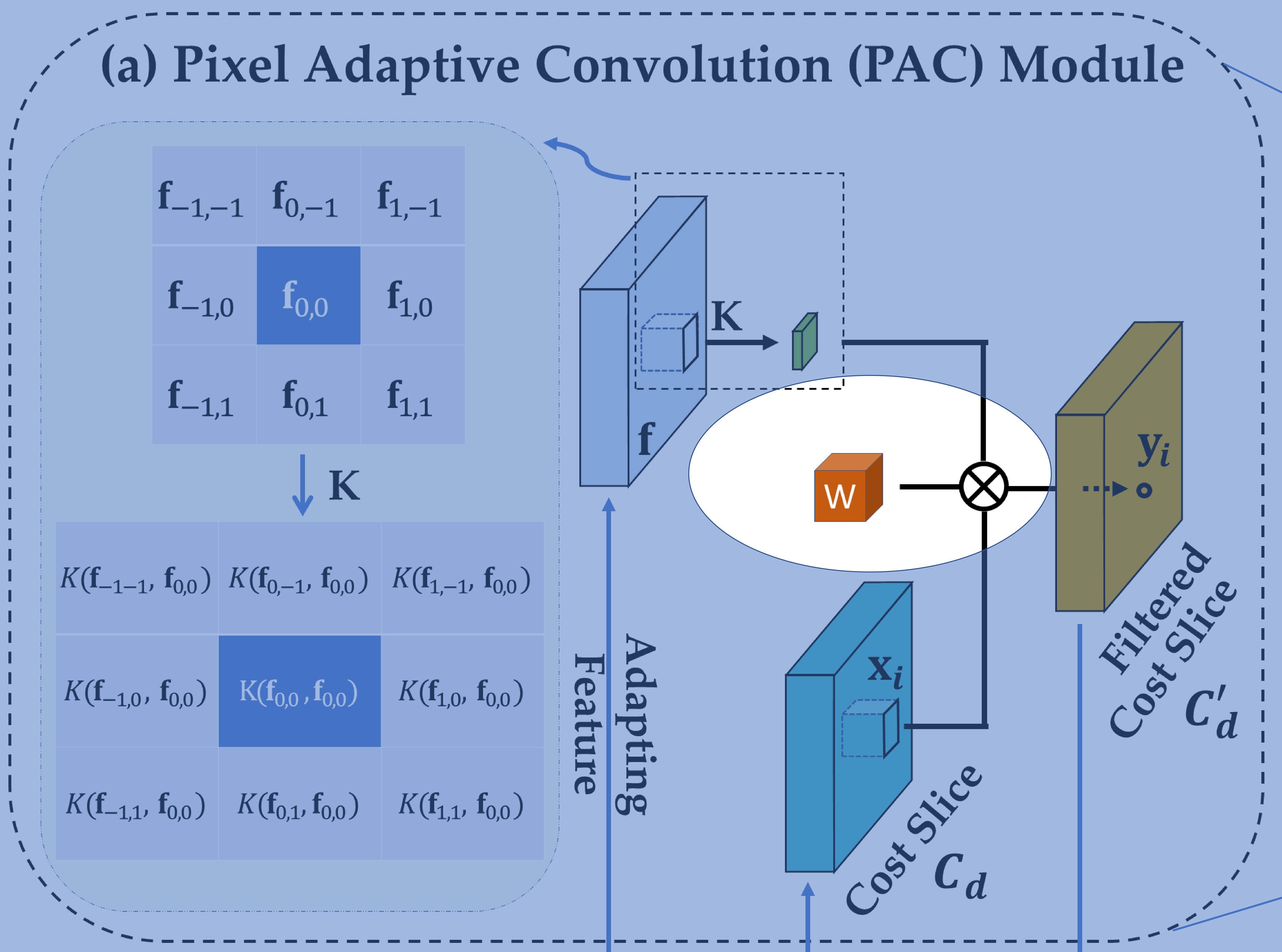


- Segmentation-aware bilateral filtering (SABF)
- Dynamic filtering networks (DFN)
- Pixel adaptive convolution (PAC)
- Semi-global aggregation (SGA)

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

- 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

Deep Adaptive Filtering: PAC

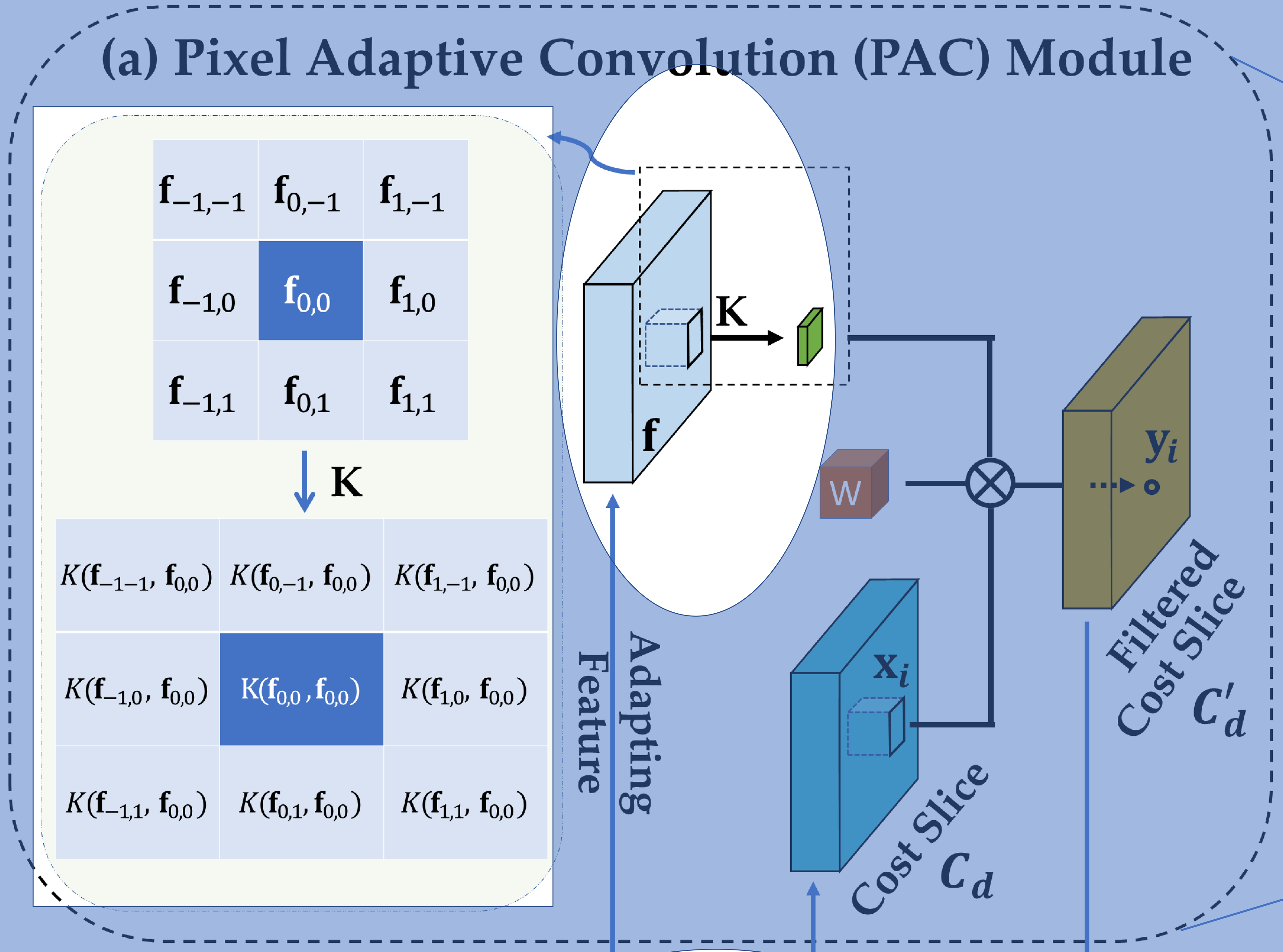


- Segmentation-aware bilateral filtering (SABF)
- Dynamic filtering networks (DFN)
- Pixel adaptive convolution (PAC)
- Semi-global aggregation (SGA)

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

- 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

Deep Adaptive Filtering: PAC

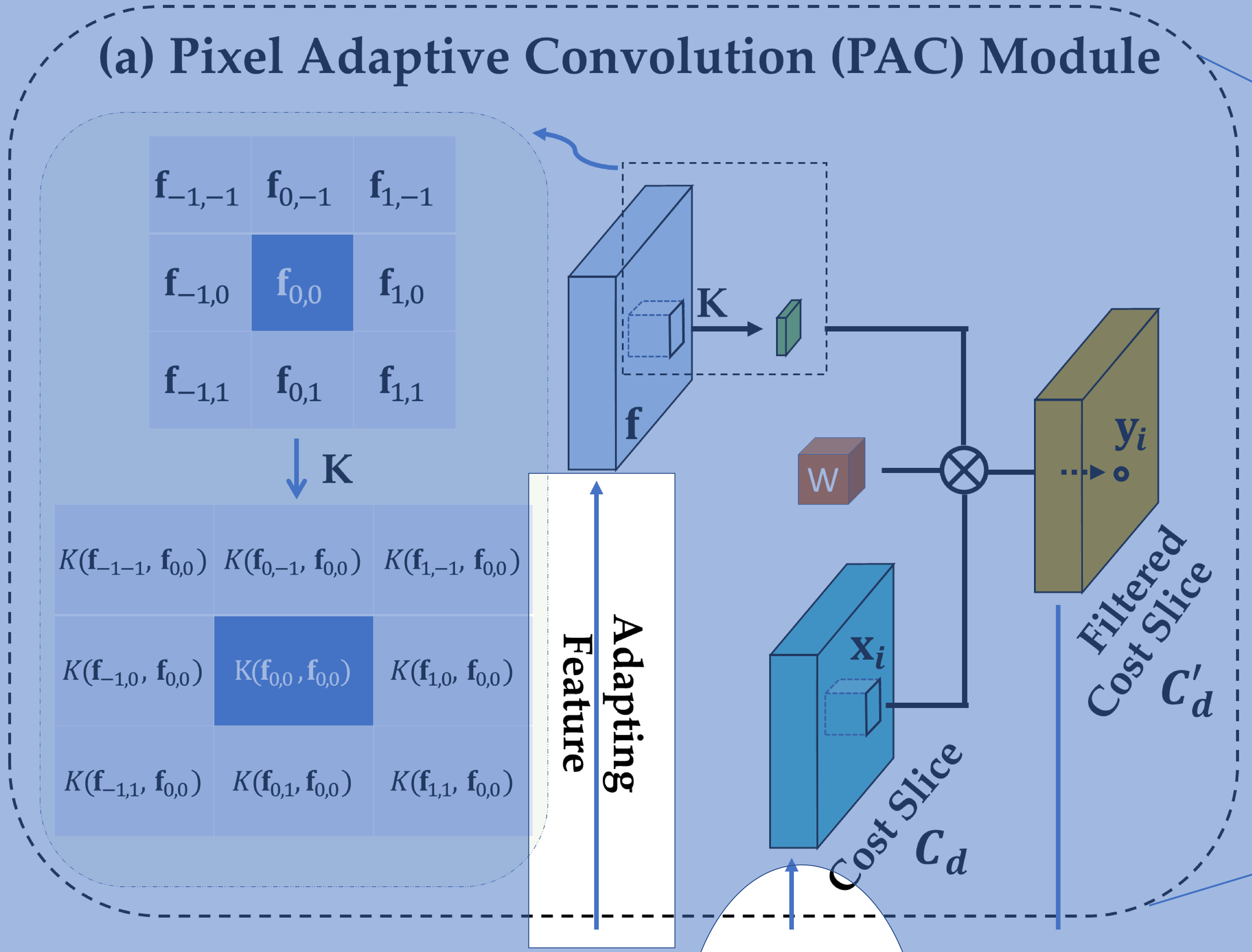


- Segmentation-aware bilateral filtering (SABF)
- Dynamic filtering networks (DFN)
- Pixel adaptive convolution (PAC)
- Semi-global aggregation (SGA)

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

- 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

Deep Adaptive Filtering: PAC



Segmentation-aware bilateral filtering (SABF)

Dynamic filtering networks (DFN)

Pixel adaptive convolution (PAC)

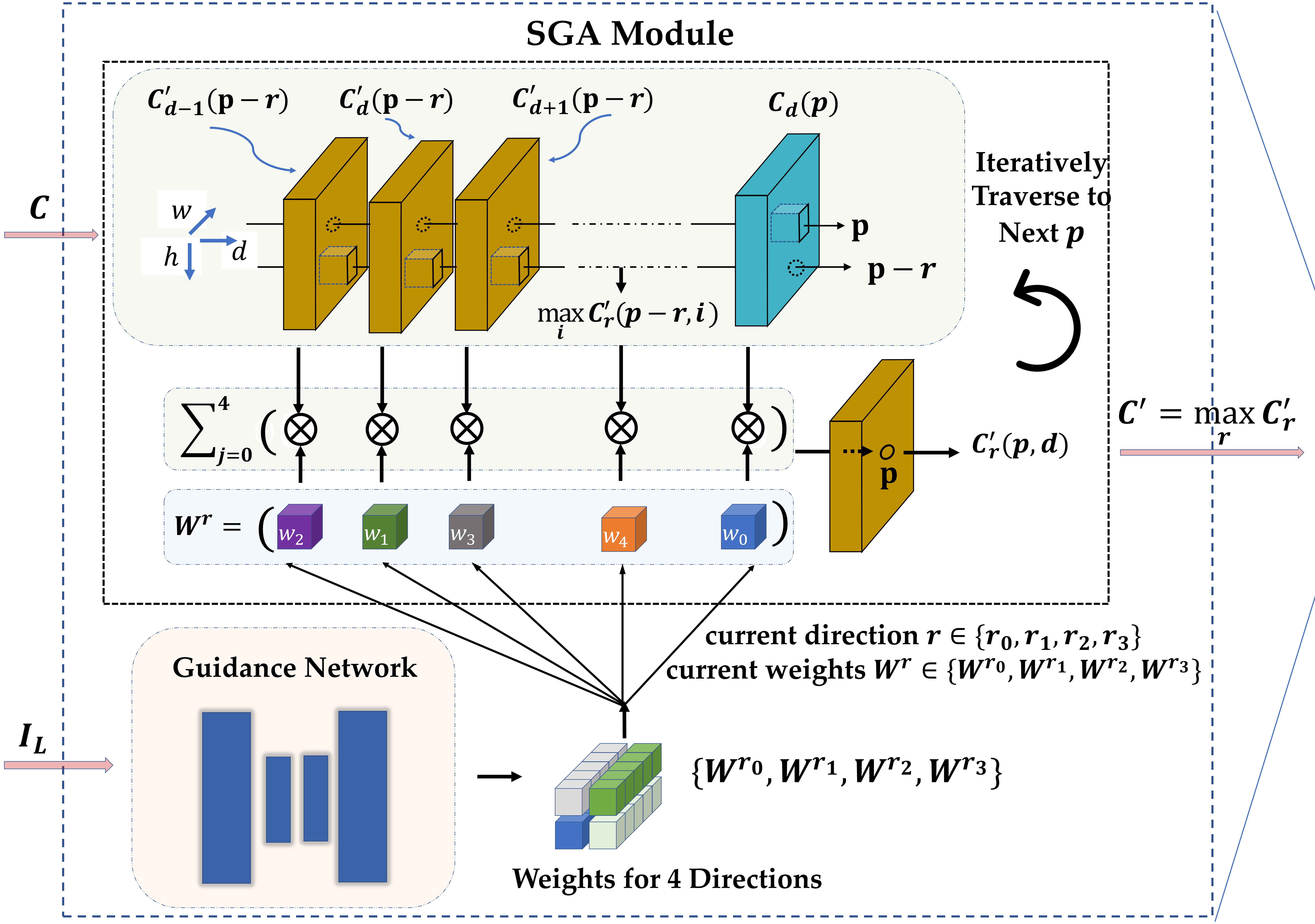
Semi-global aggregation (SGA)

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

- 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

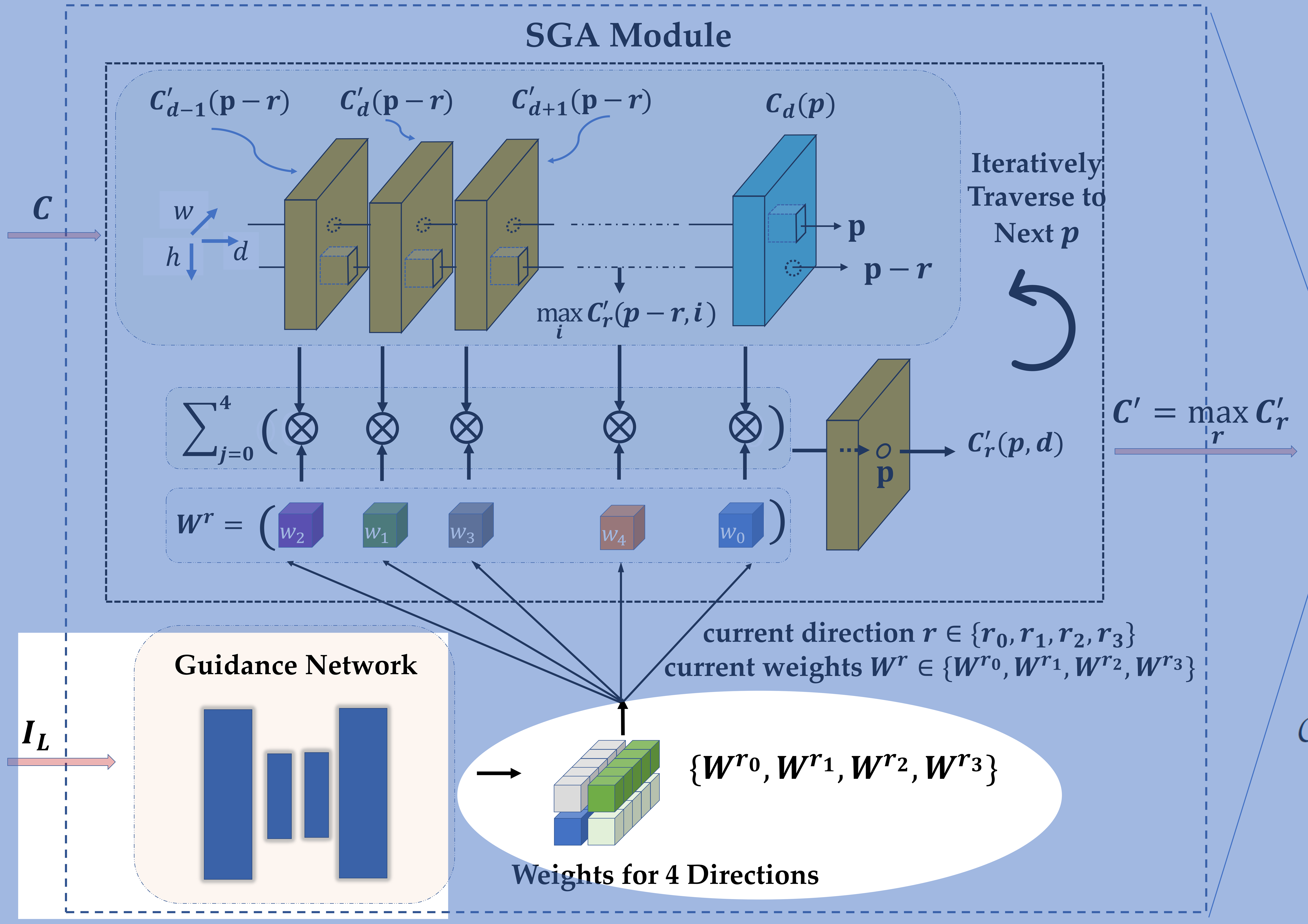
[8] F. Zhang et al. Ga-net: Guided aggregation net for end-to-end stereo matching. CVPR'19

Deep Adaptive Filtering: SGA[8]



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

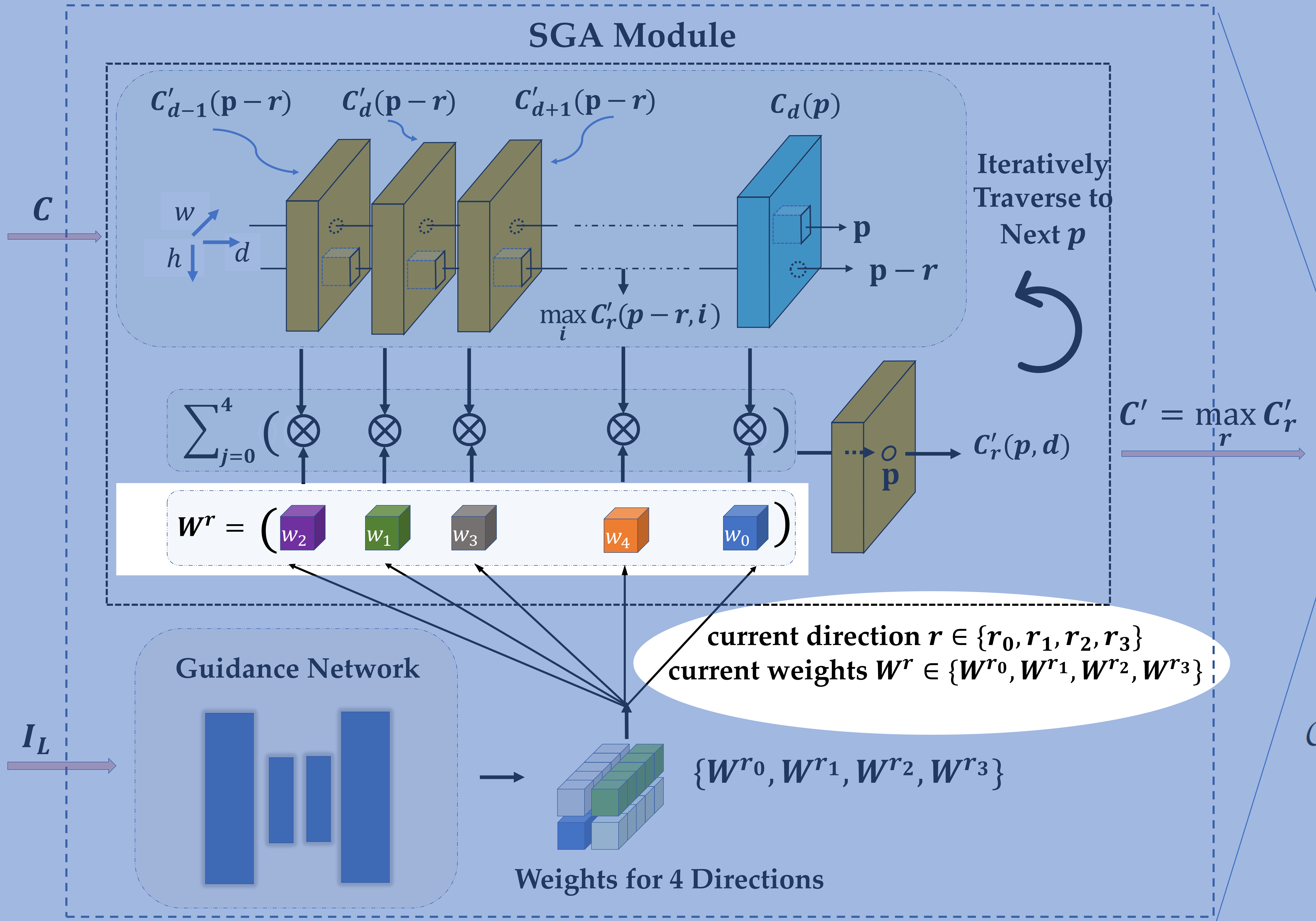
Deep Adaptive Filtering: SGA



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

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

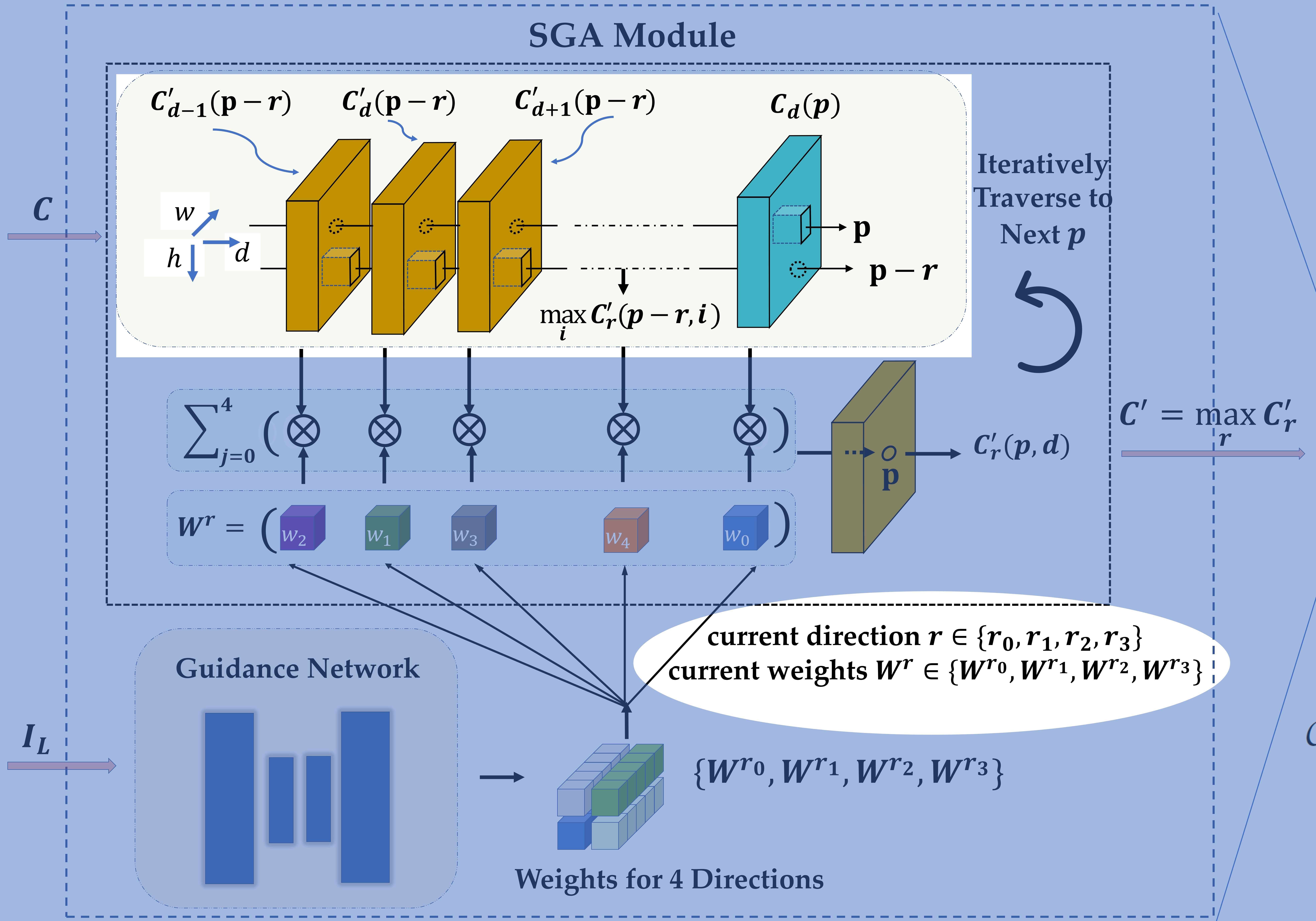
Deep Adaptive Filtering: SGA

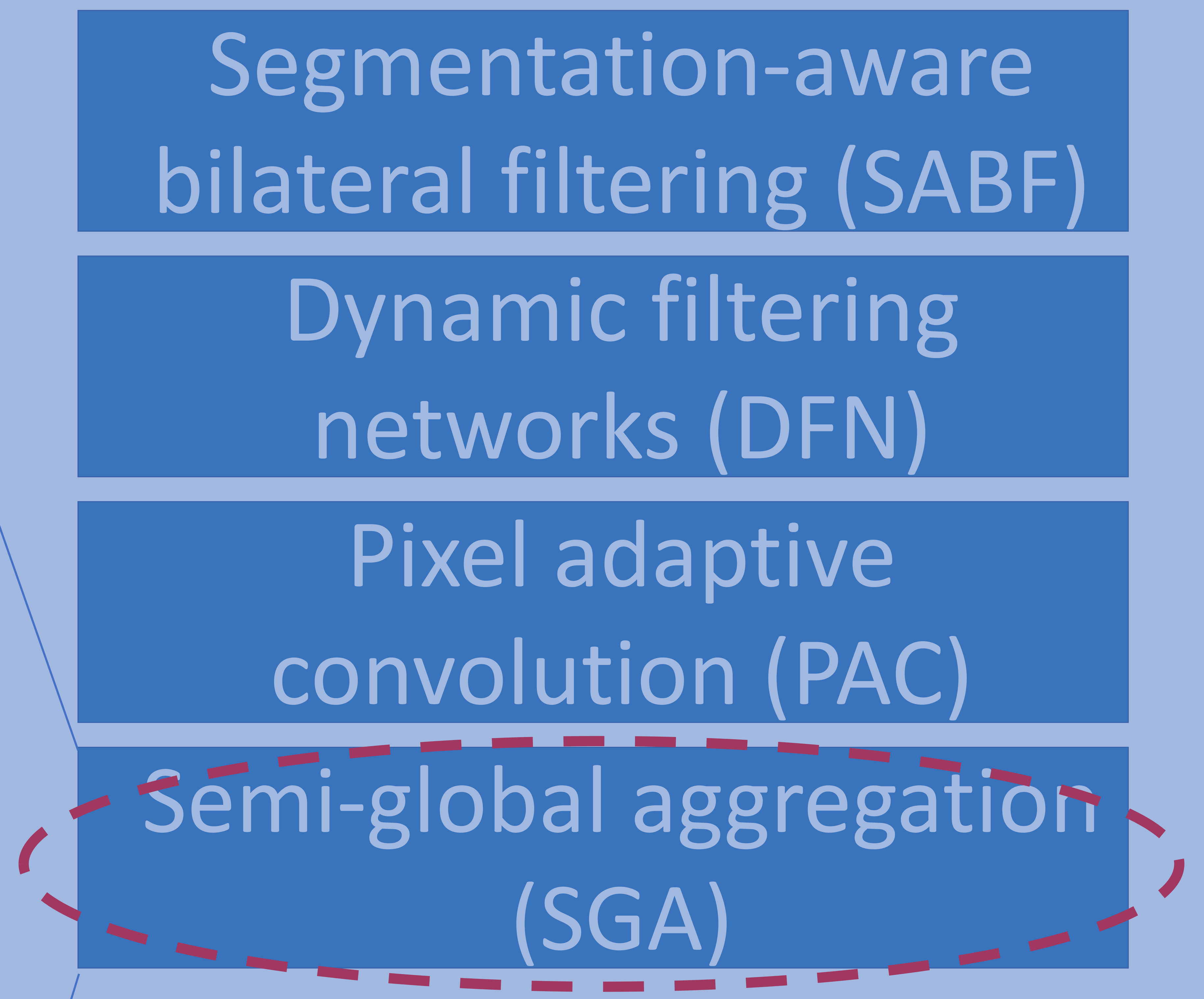


- Segmentation-aware bilateral filtering (SABF)
- Dynamic filtering networks (DFN)
- Pixel adaptive convolution (PAC)
- Semi-global aggregation (SGA)

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

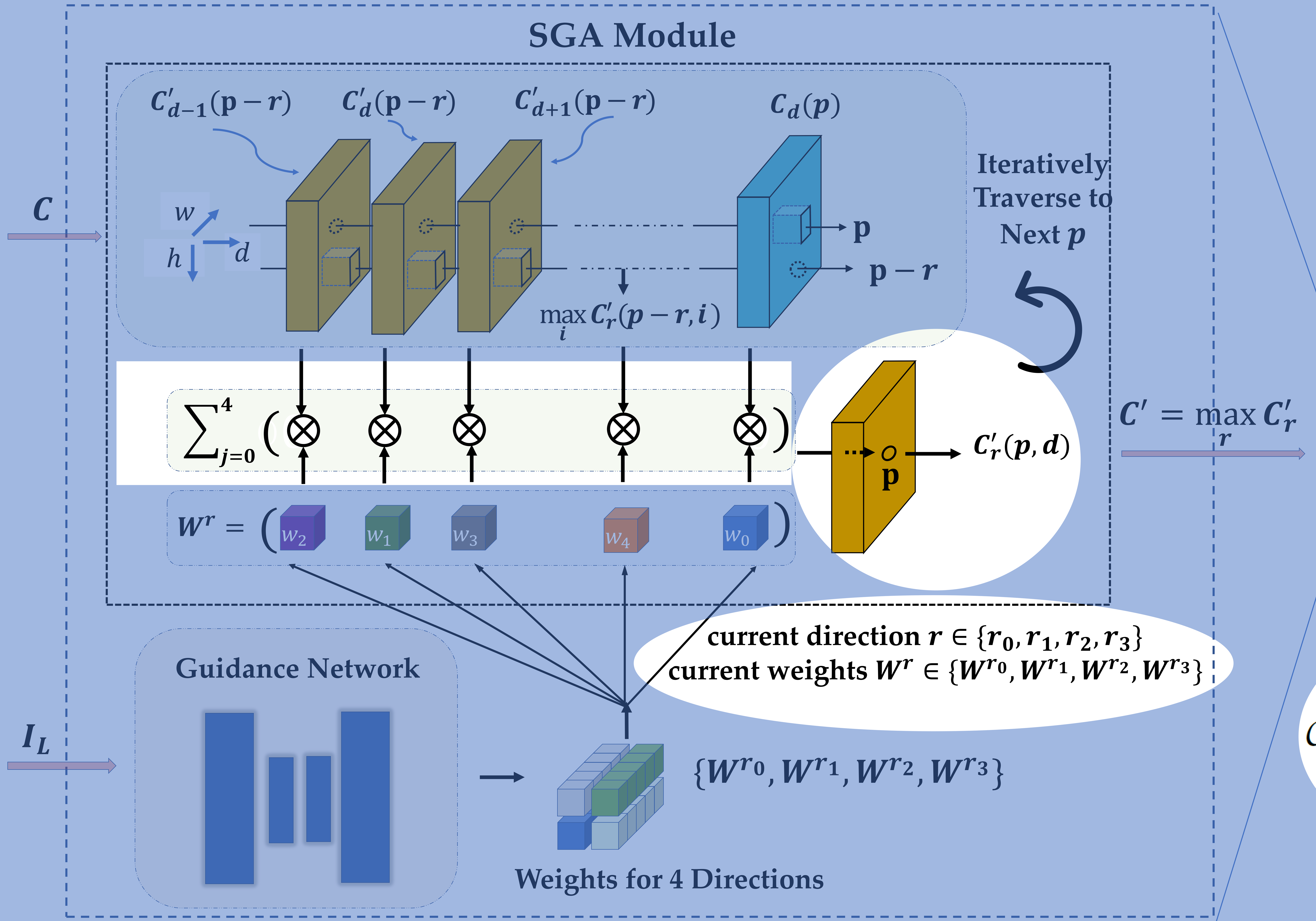
Deep Adaptive Filtering: SGA





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

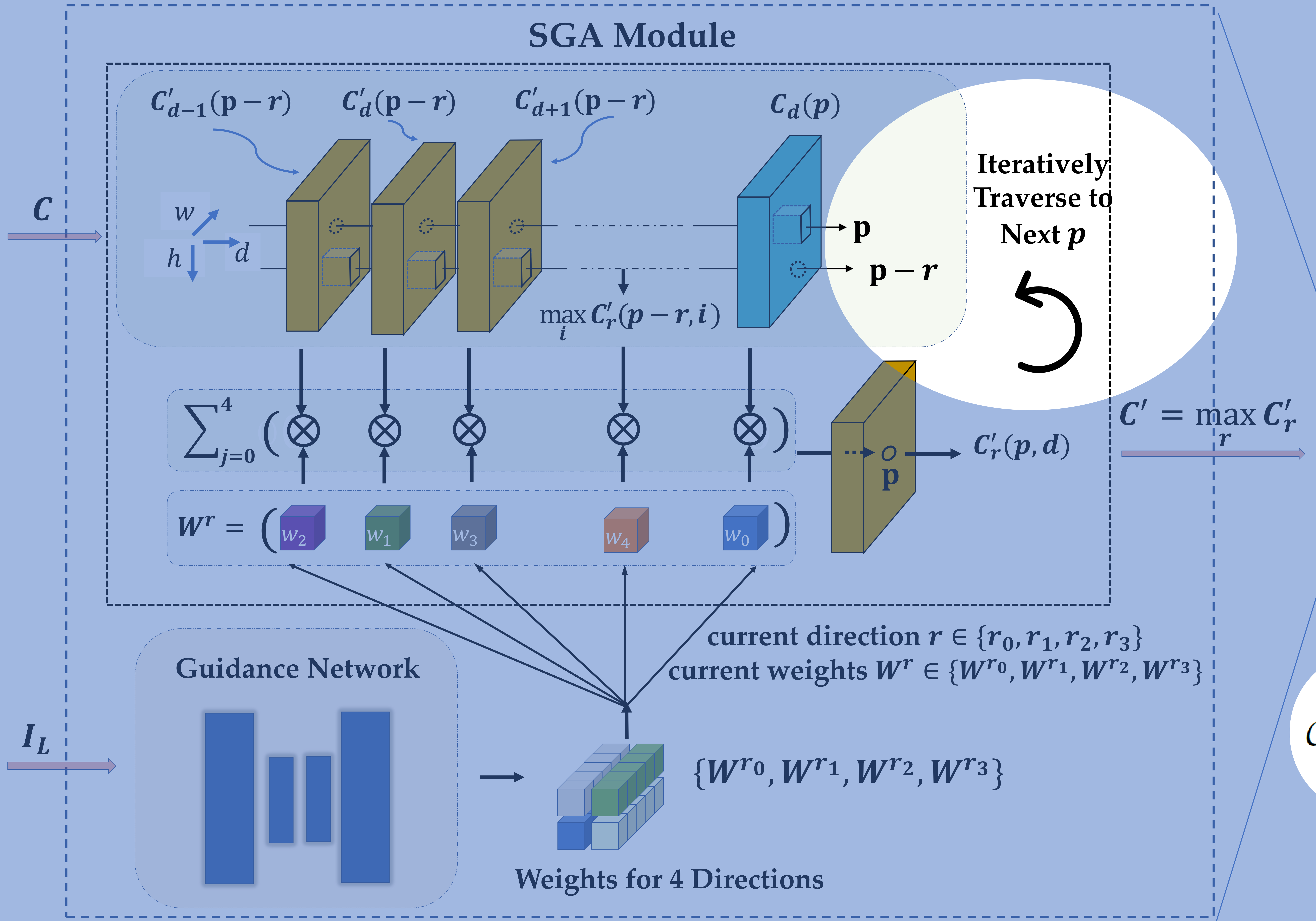
Deep Adaptive Filtering: SGA



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

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

Deep Adaptive Filtering: SGA



Segmentation-aware bilateral filtering (SABF)

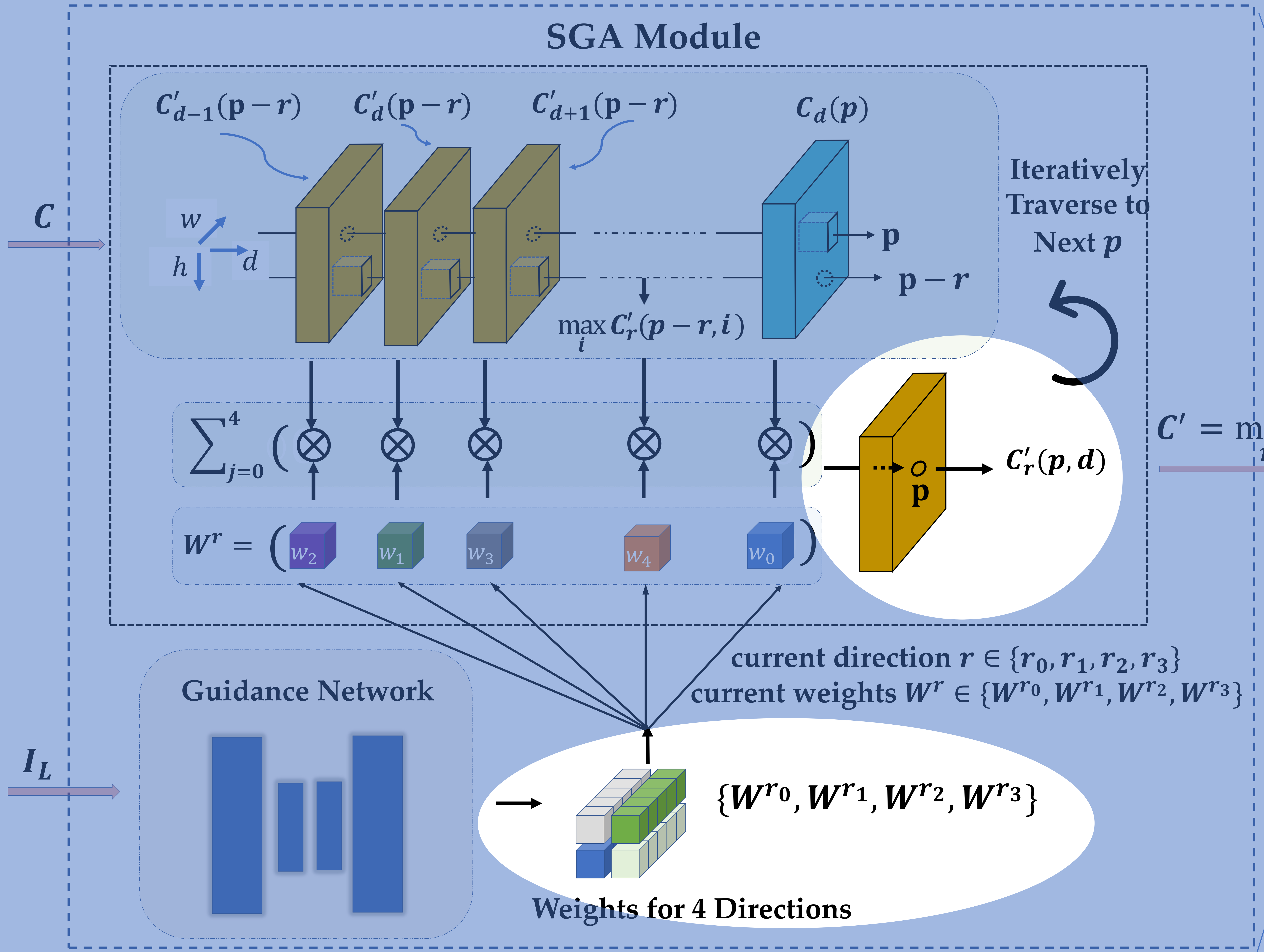
Dynamic filtering networks (DFN)

Pixel adaptive convolution (PAC)

Semi-global aggregation (SGA)

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

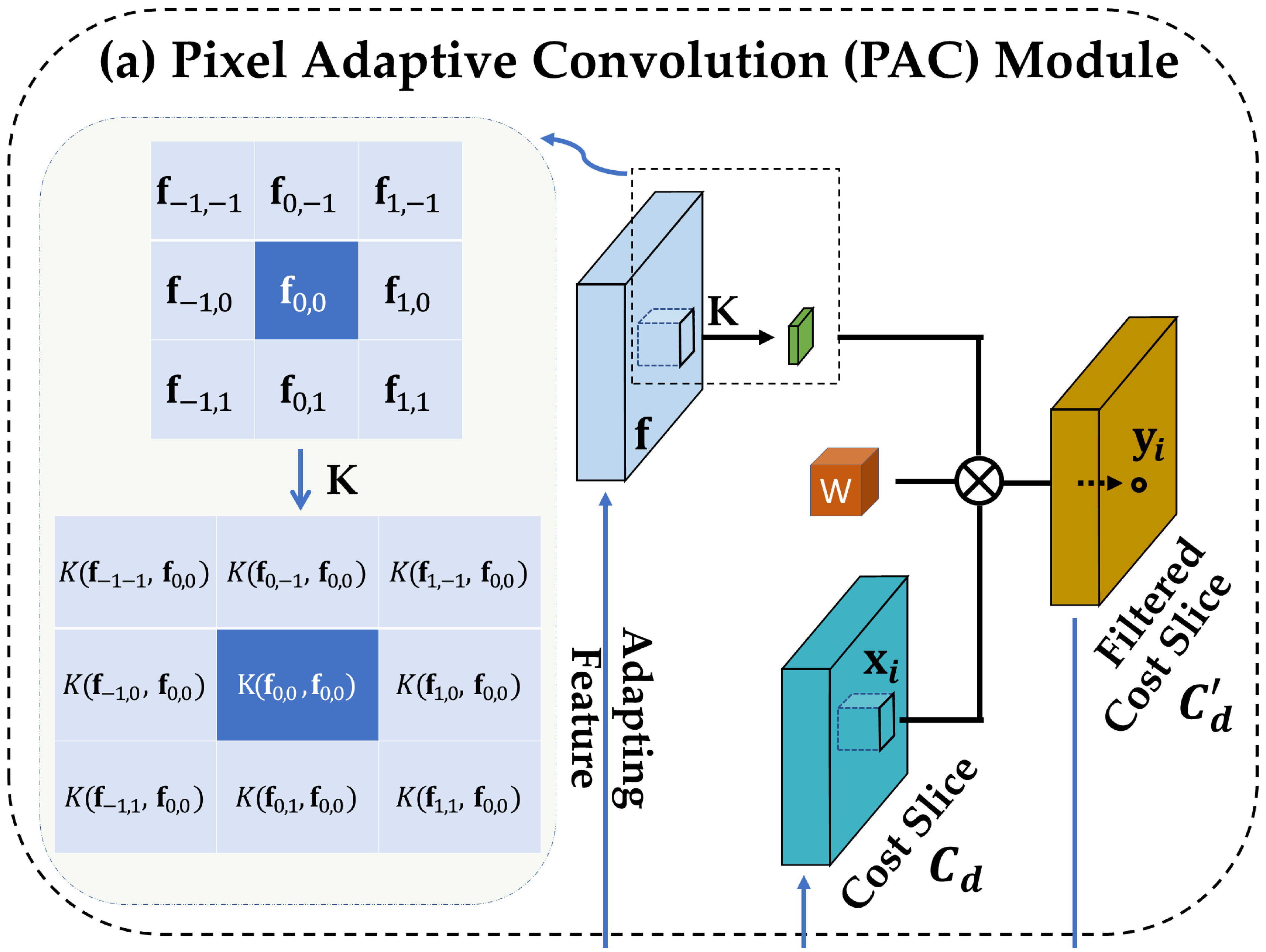
Deep Adaptive Filtering: SGA



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

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

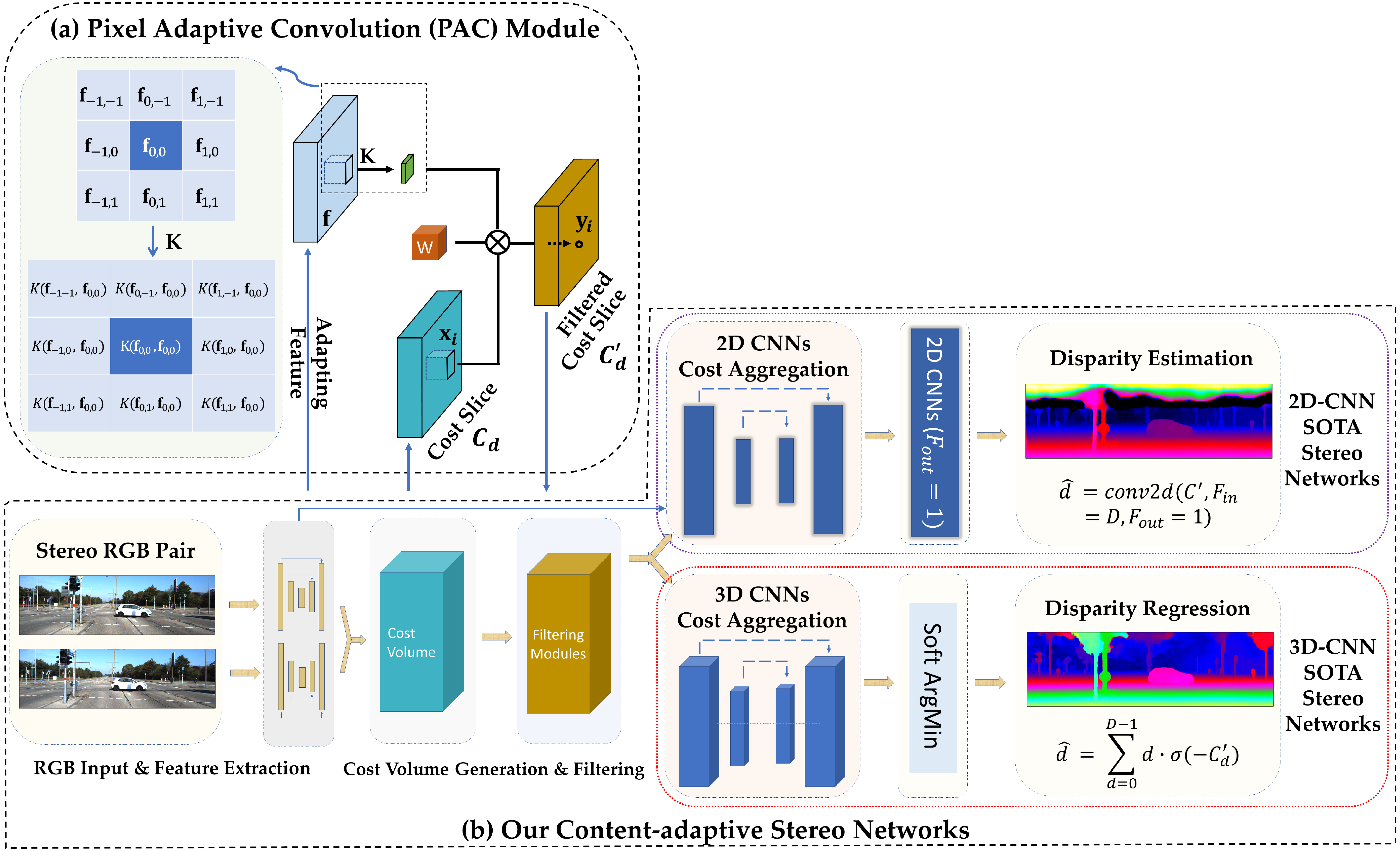
Deep Adaptive Filtering Stereo Networks



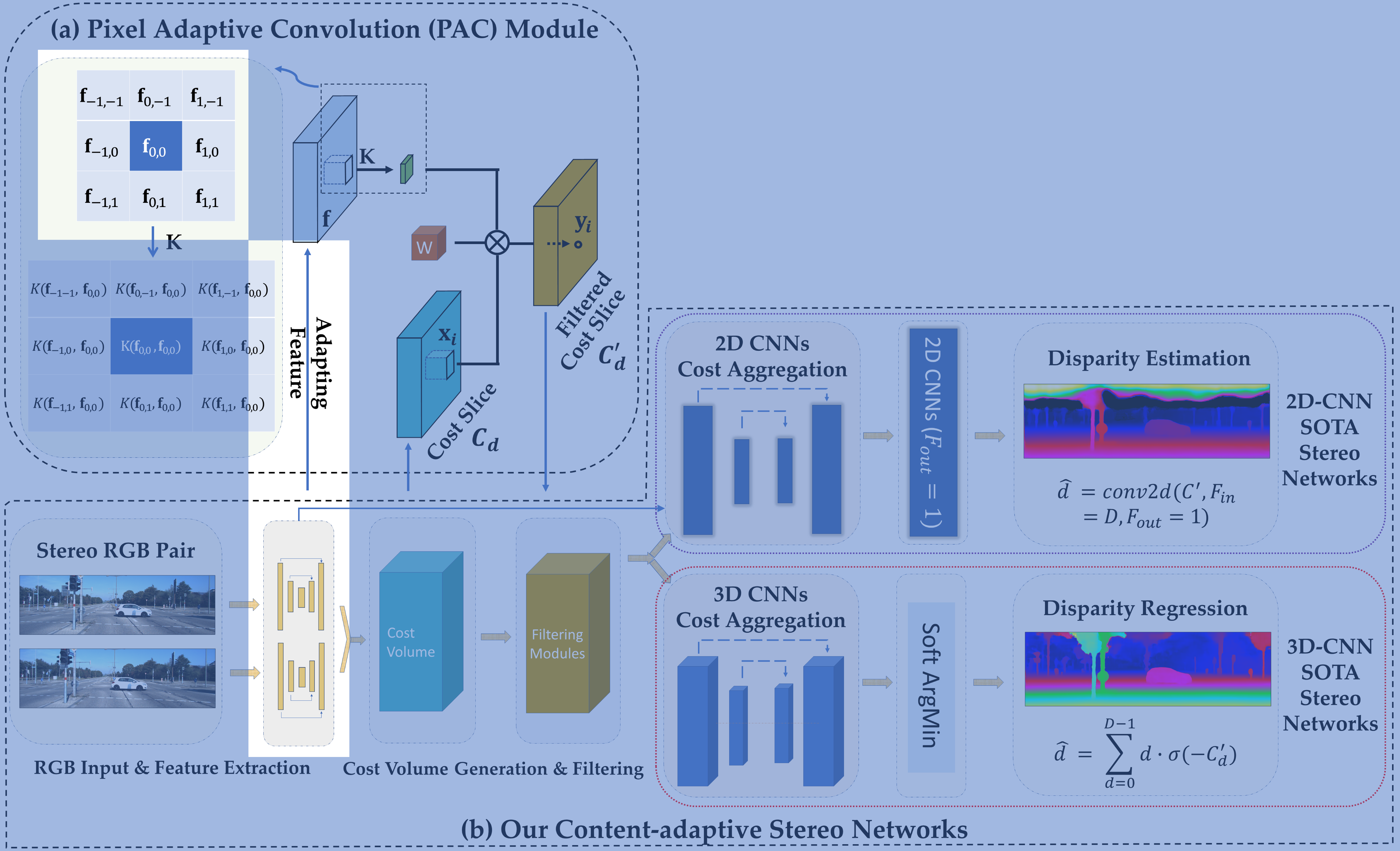
take PAC filter
as an example

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

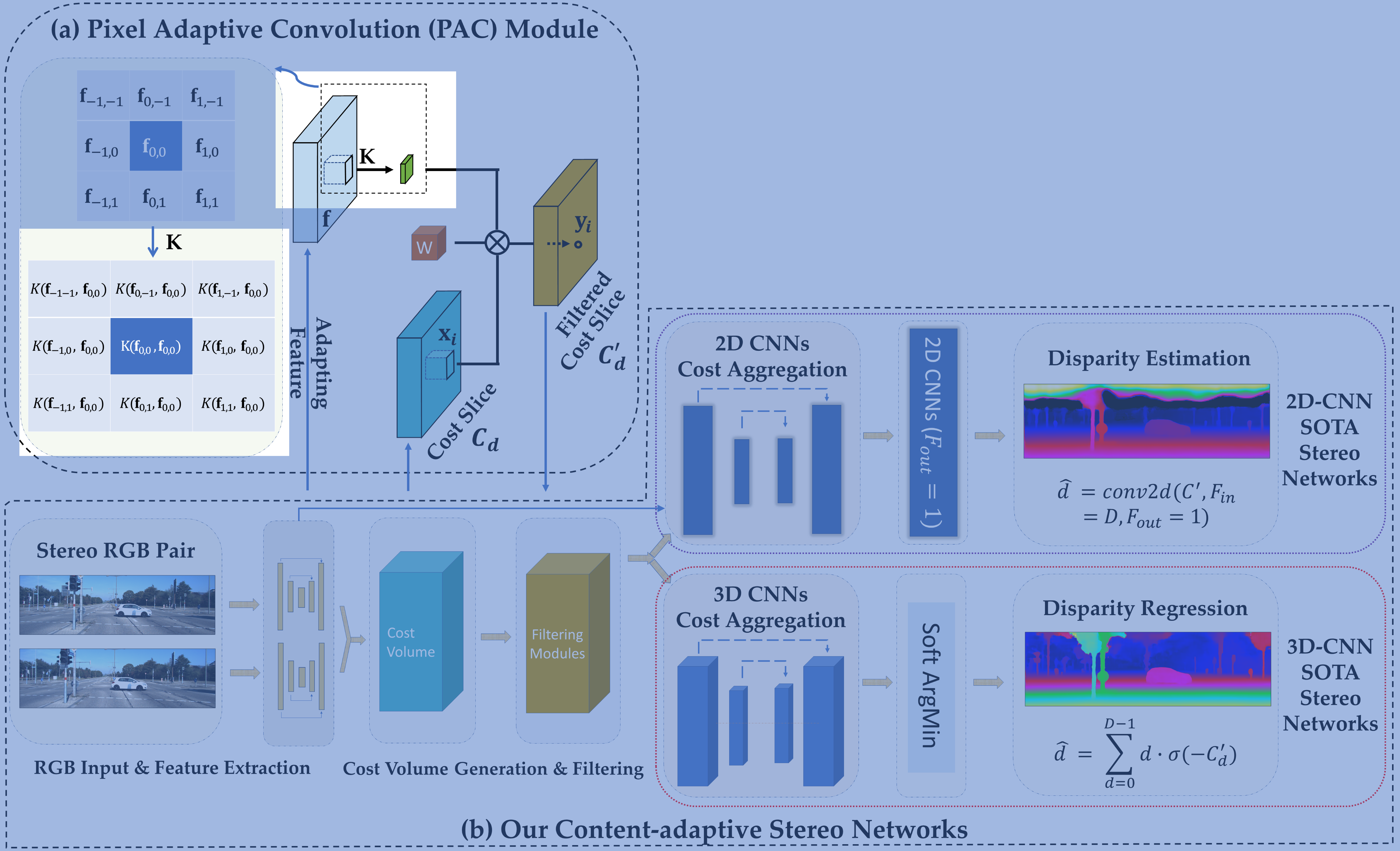
Deep Adaptive Filtering Stereo Networks



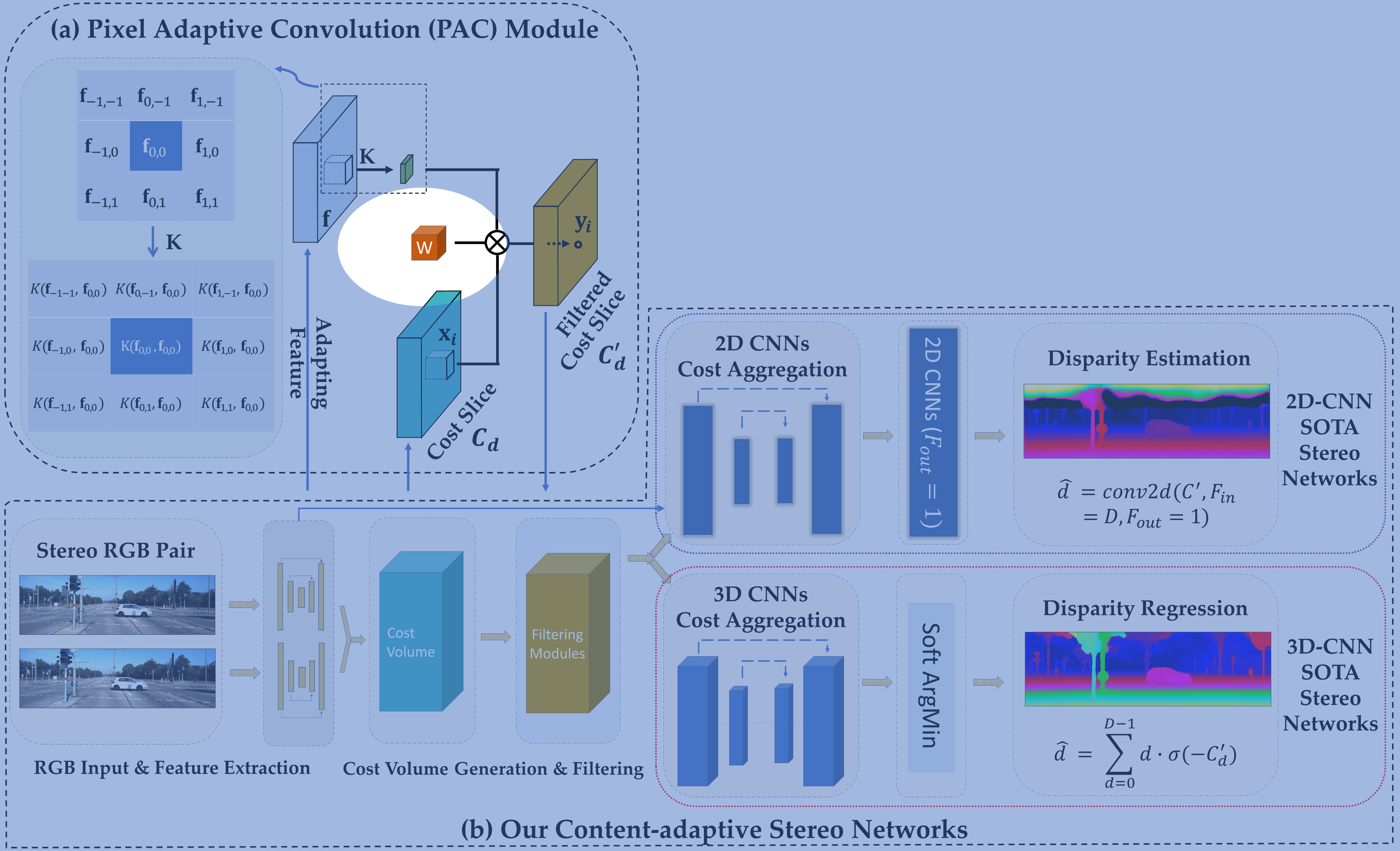
Deep Adaptive Filtering Stereo Networks



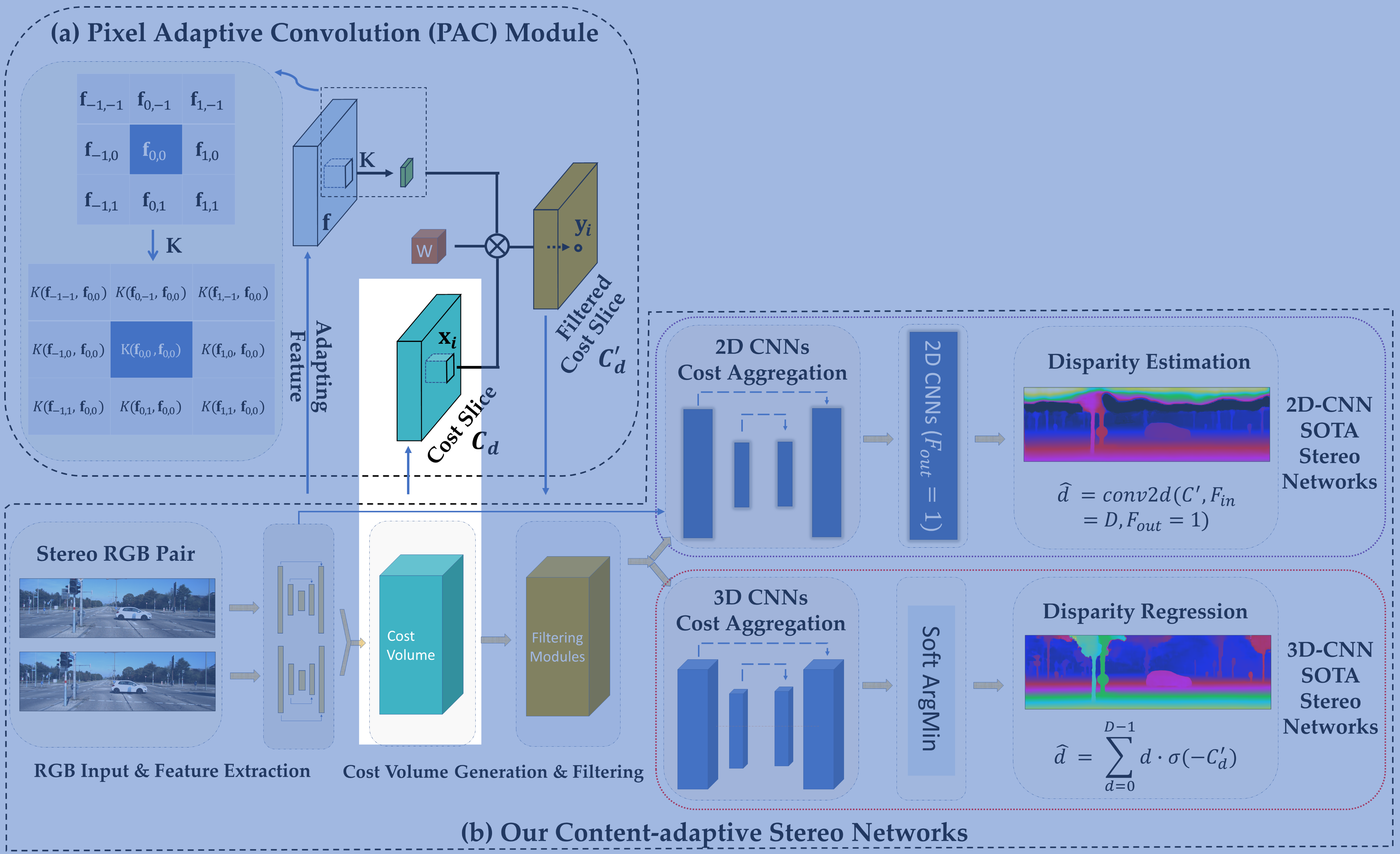
Deep Adaptive Filtering Stereo Networks



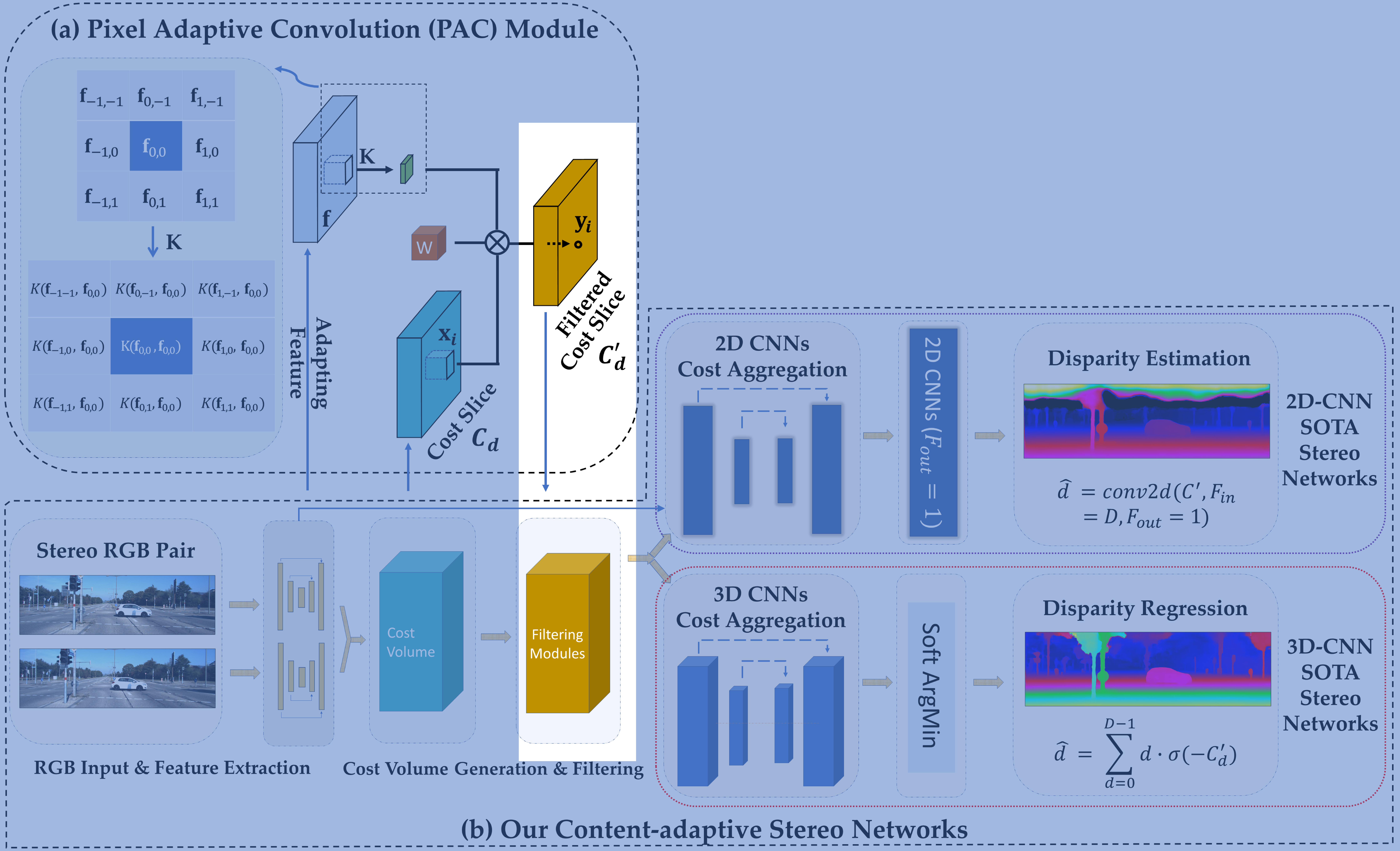
Deep Adaptive Filtering Stereo Networks



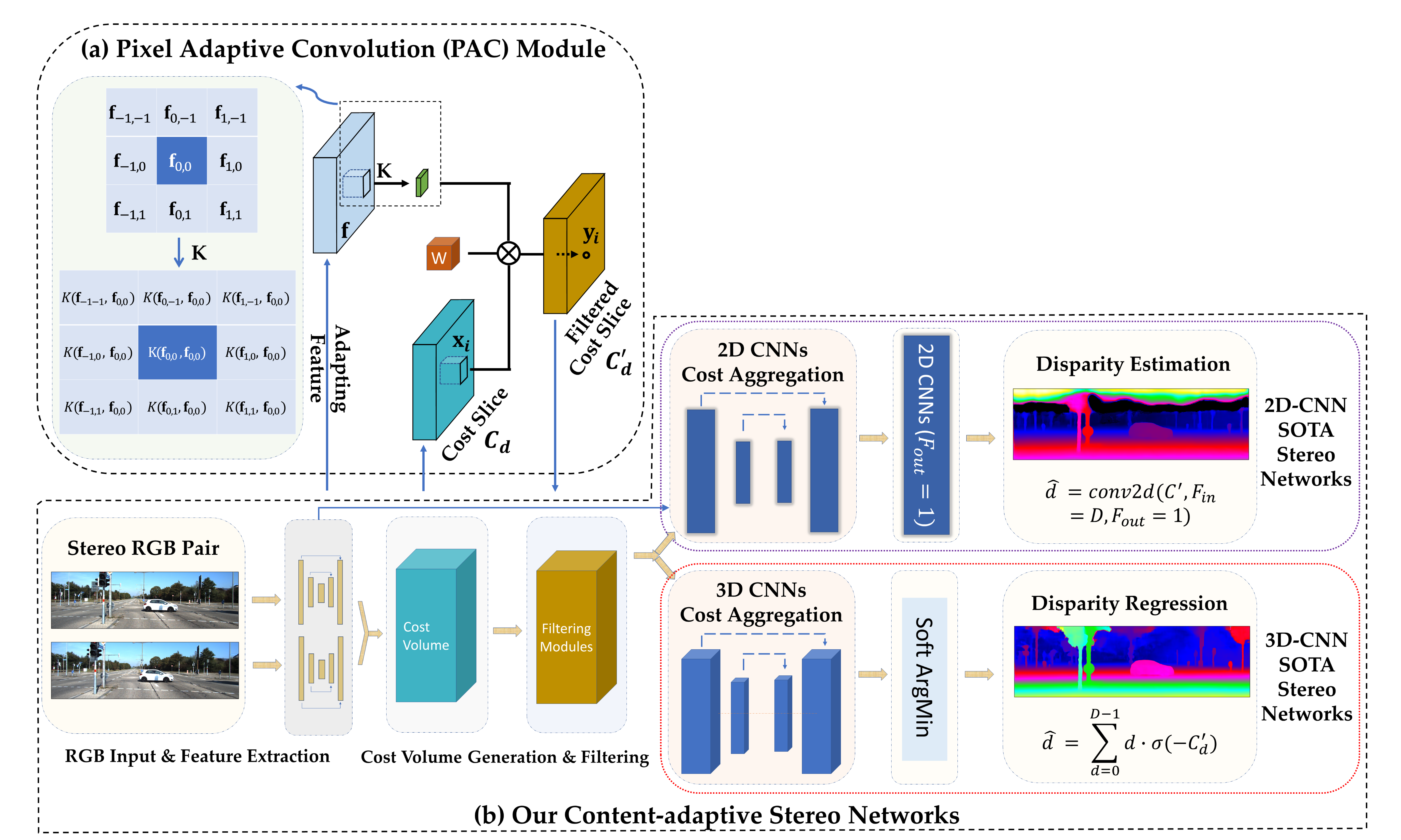
Deep Adaptive Filtering Stereo Networks



Deep Adaptive Filtering Stereo Networks

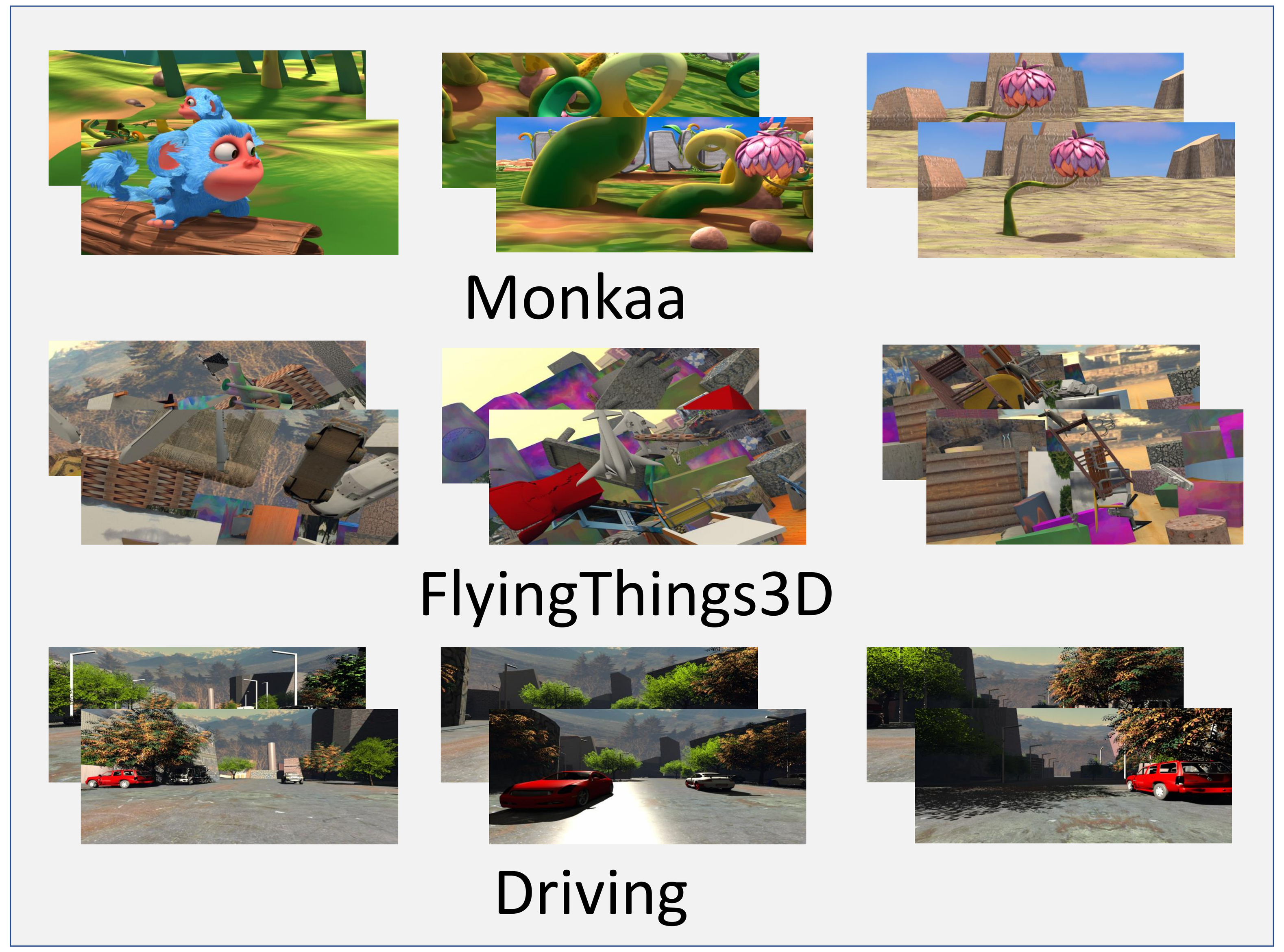


Deep Adaptive Filtering Stereo Networks



Datasets

- Scene Flow (SF)
  - a synthetic dataset of 35k training images with dense ground truth disparity maps
- KITTI 2015 (KT15)
  - a real dataset of street views, containing 200 training stereo image pairs with sparsely labeled disparity from LiDAR data
- Virtual KITTI 2 (VKT2)
  - a synthetic clone of the real KITTI dataset
- Pre-training on SF, and finetuning on VKT2 or KT15

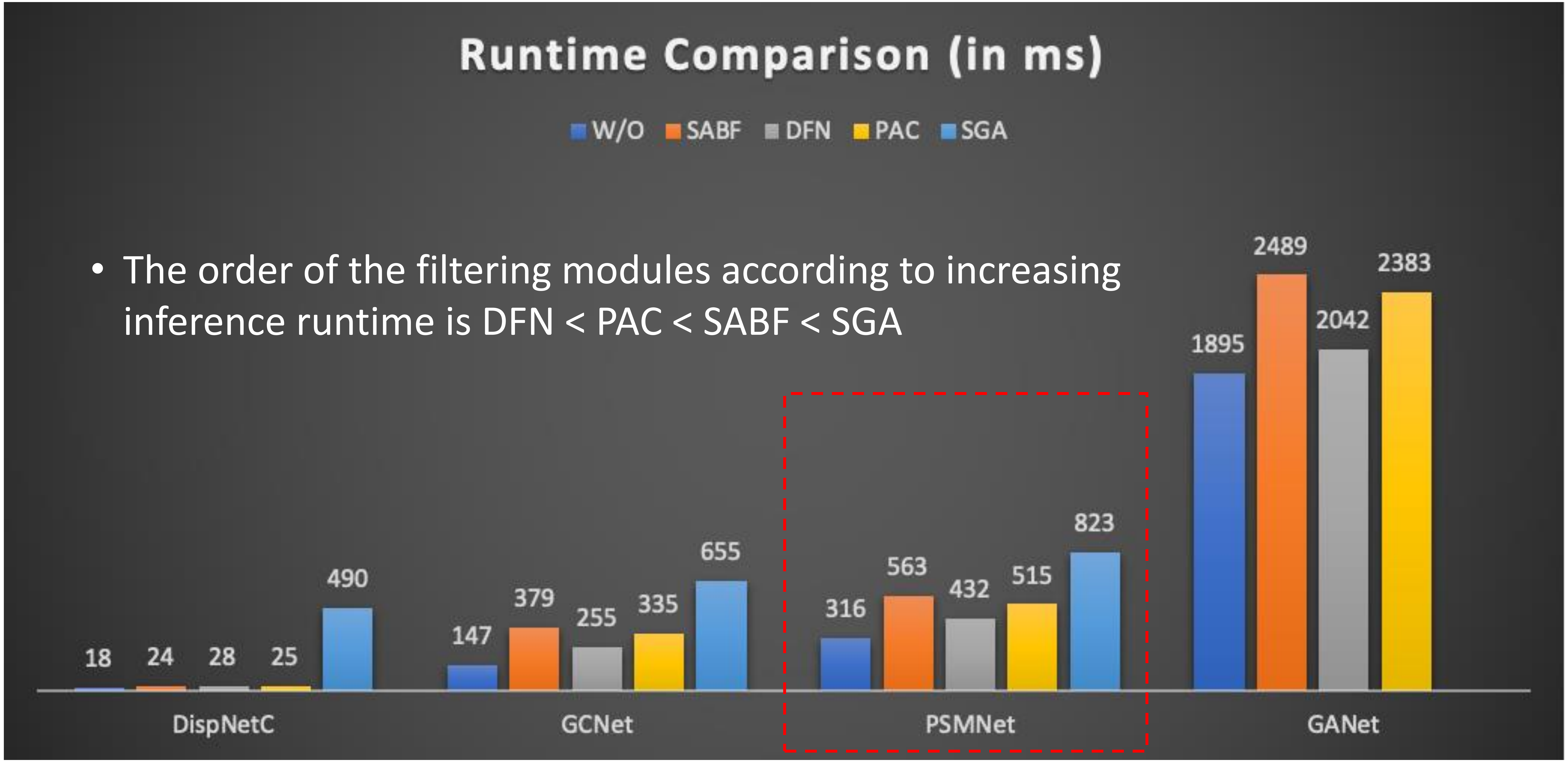


Scene Flow (SF)

KITTI 2015 (KT15)

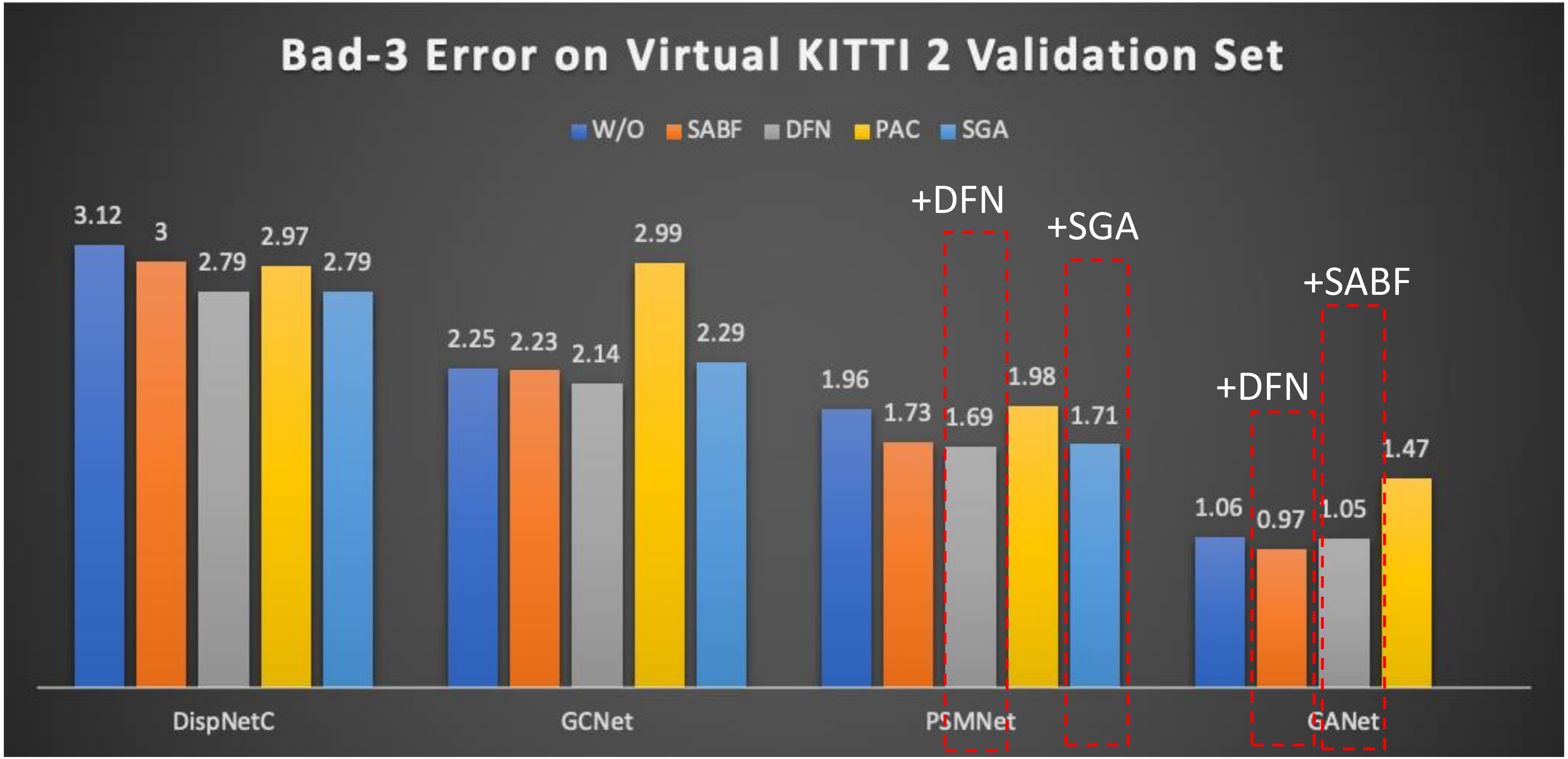
Virtual KITTI 2 (VKT2)

Network Inference Runtime Comparison

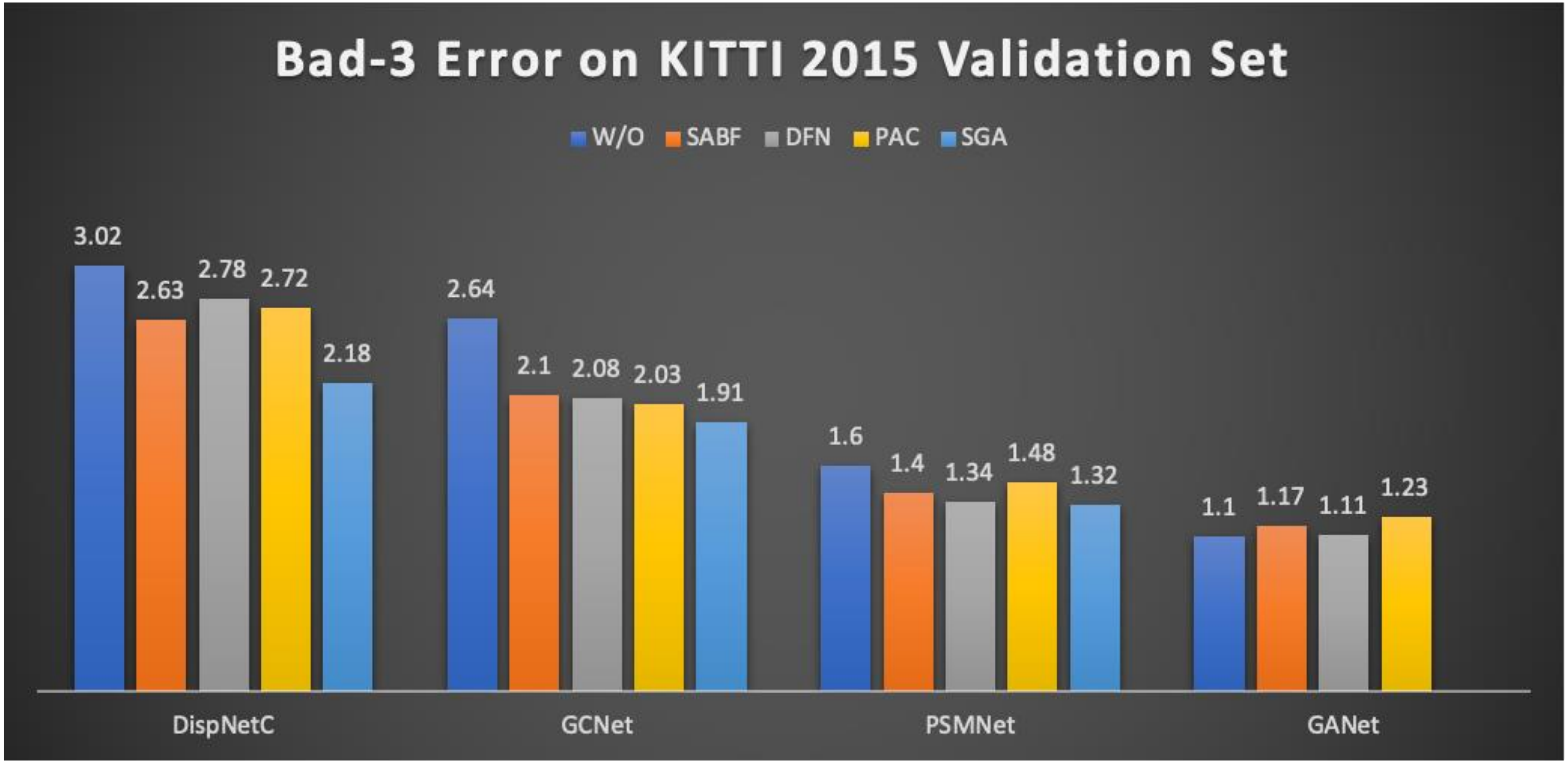


- The order of the filtering modules according to increasing inference runtime is DFN < PAC < SABF < SGA

Evaluation on Virtual KITTI 2 Val-S6



Evaluation on KITTI 2015



Evaluation on Synthetic Dataset: VKT2

- PSMNet VS PSMNet + DFN

- GCNet VS GCNet + SGA

Evaluation on Real Dataset: KT15

- DispNetC VS DispNetC + SABF

- GANet VS GANet + PAC

Summary

- Novel deep adaptive filtering architectures for end-to-end stereo matching
  - segmentation-aware bilateral filtering (SABF)
  - dynamic filtering networks (DFN)
  - 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