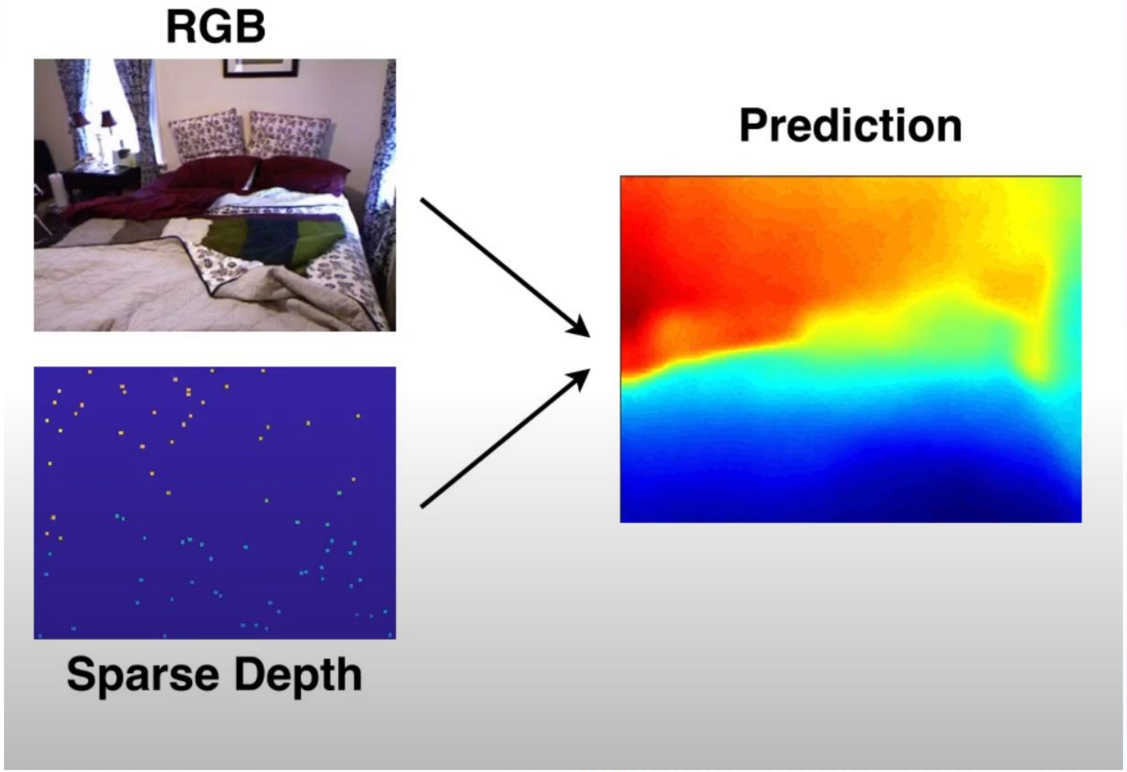
SemAttNet: Towards Attention-based Semantic Aware Guided Depth Completion

Danish Nazir, Marcus Liwicki, Didier Stricker, Muhammad Zeshan Afzal

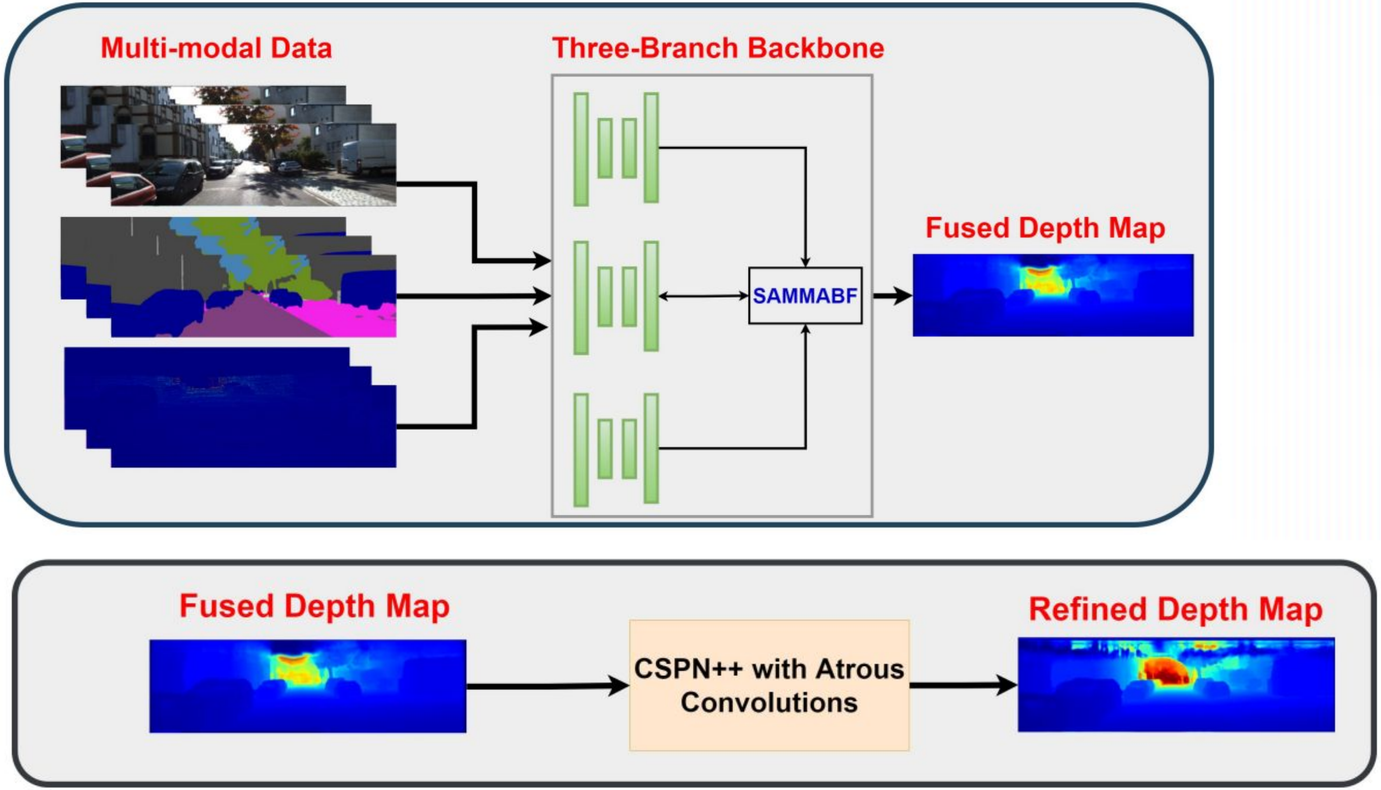
Presented by: Glanda Darie-Teofil

Research Question

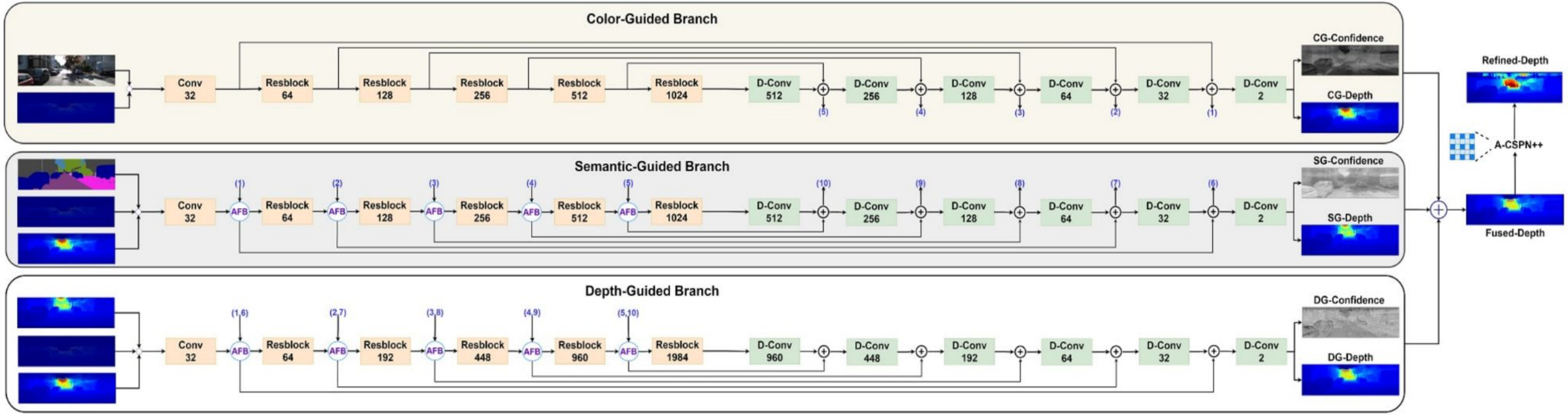
Depth Completion



SemAttNet Architecture



SemAttNet Architecture



SAMMAFB

$$\mathbf{A_c}(\mathbf{F}) = \sigma(MLP(AvgPool(\mathbf{F})) + MLP(MaxPool(\mathbf{F})))$$
$$= \sigma\left(\mathbf{W_1}\left(\mathbf{W_0}\left(\mathbf{F^c_{avg}}\right)\right) + \mathbf{W_1}\left(\mathbf{W_0}\left(\mathbf{F^c_{max}}\right)\right)\right)$$

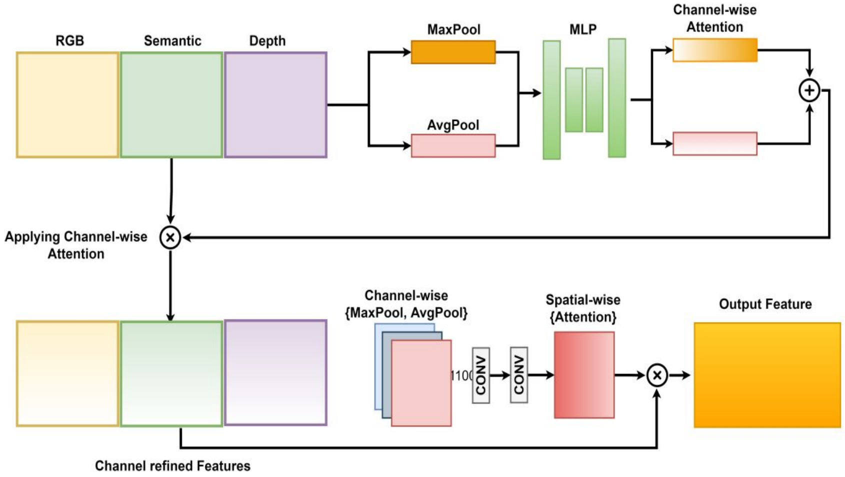
$$\mathbf{F'} = \mathbf{A_c}(\mathbf{F}) \otimes \mathbf{F}.$$

apply
channel-wise attention

$$\mathbf{A_s}\left(\mathbf{F'}\right) = \sigma\left(Conv\left(\left[\mathbf{F^c_{avg}}; \mathbf{F^c_{max}}\right]\right)\right)$$

$$\mathbf{F''} = \mathbf{A_s}(\mathbf{F'}) \otimes \mathbf{F'}$$

apply
spatial-wise attention



Color-guided Branch

$$\mathbf{X_{cg}} = [C_B; D_B] \in \mathbb{R}^{B\times 4\times H\times W}$$

Encoder $\longrightarrow$ $\phi_{\mathbf{cg}} \in \mathbb{R}^{B\times 1024\times \bar{H}\times W}$

$$\phi_{\mathbf{cg}} = f(\mathbf{W_{cg}}\mathbf{X_{cg}} + b_{cg})$$

Decoder

$C_{cg} \in \mathbb{R}^{B\times 1\times H\times W}$ $\longrightarrow$ confidence map

$D_{cg} \in \mathbb{R}^{B\times 1\times \bar{H}\times W}$ $\longrightarrow$ depth map

$$D_{cg}, C_{cg} = g(\mathbf{V_{cg}}\phi_{\mathbf{cg}} + c_{cg})$$

$$\mathbf{L_{cg}} = argmin_{D_{cg}} ||D^{gt} - D_{cg}||^2$$

Semantic-guided Branch

$$\mathbf{X_{sg}} = \left[D^B_{cg}; S_B; D_B\right] \in \ \mathbb{R}^{B\times 4\times H\times W}$$

Encoder $\longrightarrow$ $\phi_{\mathbf{sg}} \in \mathbb{R}^{B\times 1024\times H\times W}$

$$\phi_{\mathbf{sg}} = f(\mathbf{W^T_{sg}}.\mathbf{X_{sg}} + b_{sg})$$

Decoder

$C_{sg} \ \in \mathbb{R}^{B\times 1\times H\times W}$ $\longrightarrow$ confidence map

$D_{sg} \in \mathbb{R}^{B\times 1\times \bar{H}\times W}$ $\longrightarrow$ depth map

$$D_{sg}, C_{sg} = g(\mathbf{V^T_{sg}}.\phi_{\mathbf{sg}} + c_{sg})$$

$$\mathbf{L_{sg}} = argmin_{D_{sg}} \ || \ D^{gt} - D_{sg}||^2$$

Depth-guided Branch

$$\mathbf{X_{dg}} = \left[D_{cg}^{B}; D_{sg}^{B}; D_{B}\right] \in \mathbb{R}^{B\times 3\times H\times W}$$

Encoder $\longrightarrow$ $\phi_{\mathbf{dg}} \in \mathbb{R}^{B\times 1984\times H\times W}$

$$\phi_{\mathbf{dg}} = f(\mathbf{W_{dg}^{T}}.\mathbf{X_{dg}} + b_{dg})$$

Decoder

$C_{dg} \in \mathbb{R}^{B\times 1\times H\times W}$ $\longrightarrow$ confidence map

$D_{dg} \in \mathbb{R}^{B\times 1\times H\times W}$ $\longrightarrow$ depth map

$$D_{dg}, C_{dg} = g(\mathbf{V_{dg}^{T}}.\phi_{\mathbf{dg}} + c_{dg})$$

$$\mathbf{L_{dg}} = argmin_{D_{dg}} \,||\, D^{gt} - D_{dg}||^{2}$$

Multi-Modal Depth Fusion

$$D_f = \frac{e^{C_{cg}} \cdot D_{cg} + e^{C_{sg}} \cdot D_{sg} + e^{C_{dg}} \cdot D_{dg}}{e^{C_{cg}} + e^{C_{sg}} + e^{C_{dg}}}$$

$$\mathbf{L_{fused}} = argmin_{D_f} \, || \, D^{gt} - D_f ||^2$$

$$\mathbf{L_{total}} = \lambda_{cg} L_{cg} + \lambda_{sg} L_{sg} + \lambda_{dg} L_{dg} + L_{fused}$$

Implementation Details

- Pytorch is used for implementation
- Adam optimizer
- Weight decay is 10^-6
- Perform random cropping, flipping and color jitter on the dataset
- Batch size for the three-branch backbone is 8
- Initial learning rate is 0.00128
- Initial weight of 0.2 is assigned to $\lambda_{cg}, \lambda_{sg}$ and λ_{dg} coefficients
- Three-branch backbone is trained for 60 epochs
- CSPN++ with Atrous convolutions is trained for 95 epochs

Experiments

Method	RMSE mm	MAE mm	iRMSE 1/km	iMAE 1/km
TWISE [45]	840.20	**195.58**	2.08	**0.82**
DSPN [31]	766.74	220.36	2.47	1.03
DLiDAR [12]	758.38	226.50	2.56	1.15
FuseNet [20]	752.88	221.19	2.34	1.14
ACMNet [17]	744.91	206.09	2.08	0.90
CSPN++ [19]	743.69	209.28	2.07	0.90
NLSPN [4]	741.68	199.59	**1.99**	0.84
GuideNet [8]	736.24	218.83	2.25	0.99
FCFRNet [18]	735.81	217.15	2.20	0.98
PENet [13]	730.08	210.55	2.17	0.94
RigNet [14]	713.44	204.55	2.16	0.92
SemAttNet	**709.41**	205.49	2.03	0.90

Conclusions

- They propose a novel three-branch backbone for sparse depth completion, which counters the sensitivity of image-guided methods to optical changes (e.g., shadows and reflections).
- They present a novel SAMMAFB block to actively fuse the color, semantic, and depth modalities at multiple stages in their three-branch backbone.
- SemAttNet's ability to mitigate sensitivity to optical changes while achieving superior results signifies its potential in advancing depth completion techniques for various real-world applications, such as autonomous driving, robotics, 3D reconstruction, and augmented reality.
- Extensive experimental results show that their model achieves state-of-the-art results on the outdoor KITTI depth completion dataset.

Thank you!