Real-time Monocular Depth Estimation with Extremely Light-Weight Neural Network

Paper ID: 1954

Mian-Jhong Chiu Wei-Chen Chiu Hua-Tsung Chen Jen-Hui Chuang

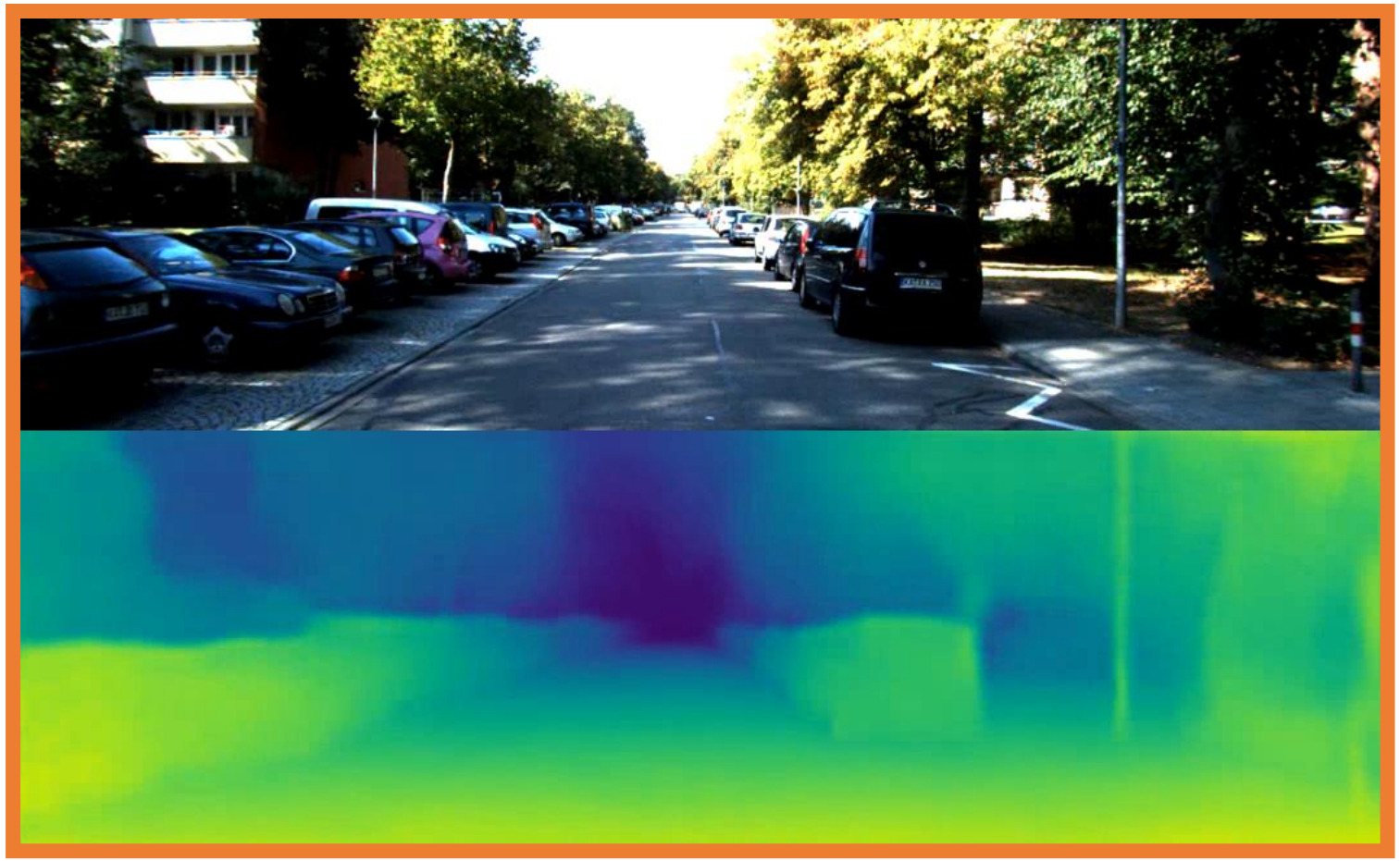
National Chiao Tung University

Introduction

- Autonomous Driving
- Augmented Reality
- Robotics
- 3D Modeling

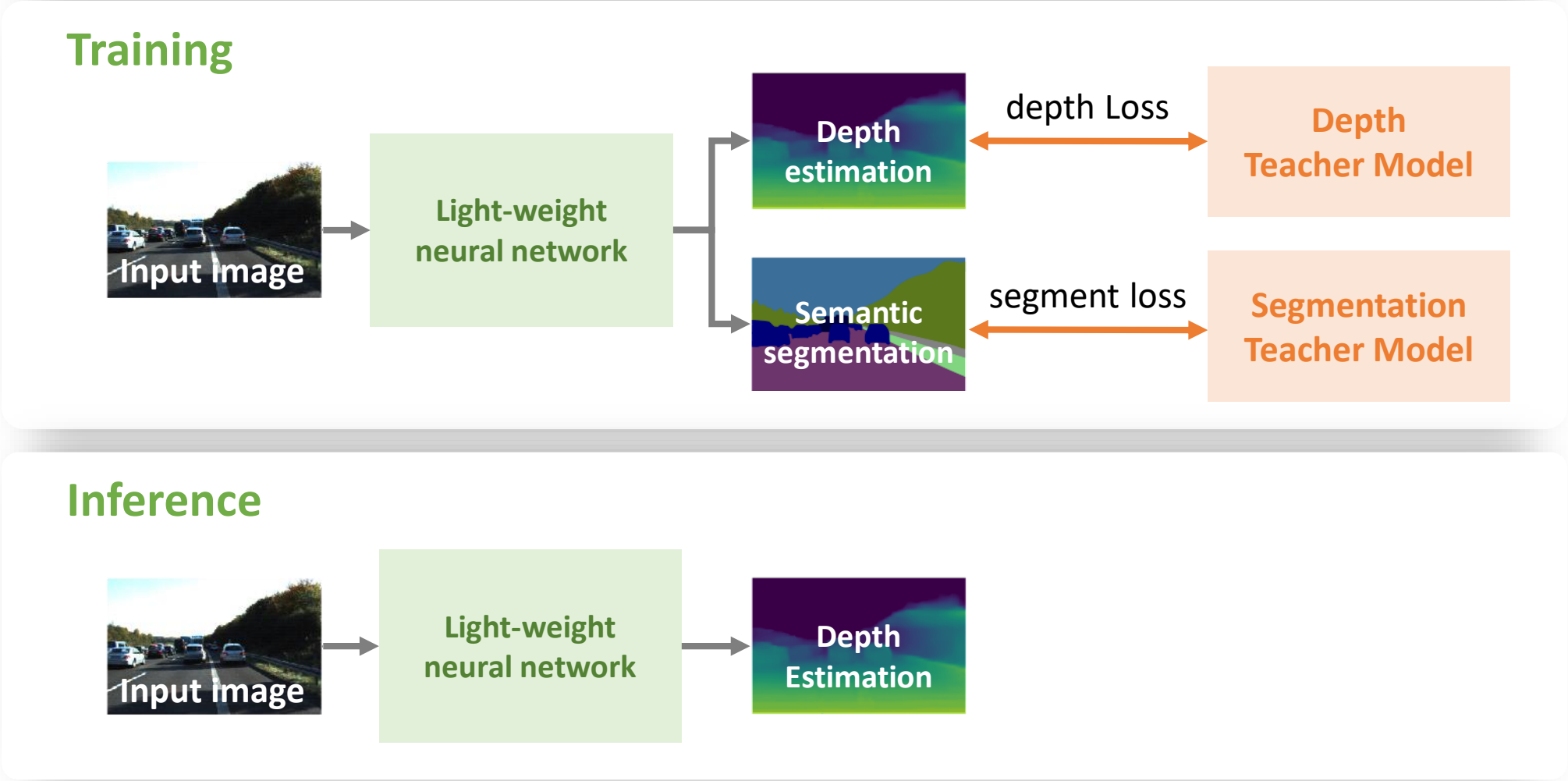
Introduction

Method	RMSE	Params
Elkerdawy et al.	5.891	5.9 M
Poggi et al.	6.030	1.9 M
Nekrasov et al.	3.453	2.99 M
Ours	**3.871**	**0.32 M**



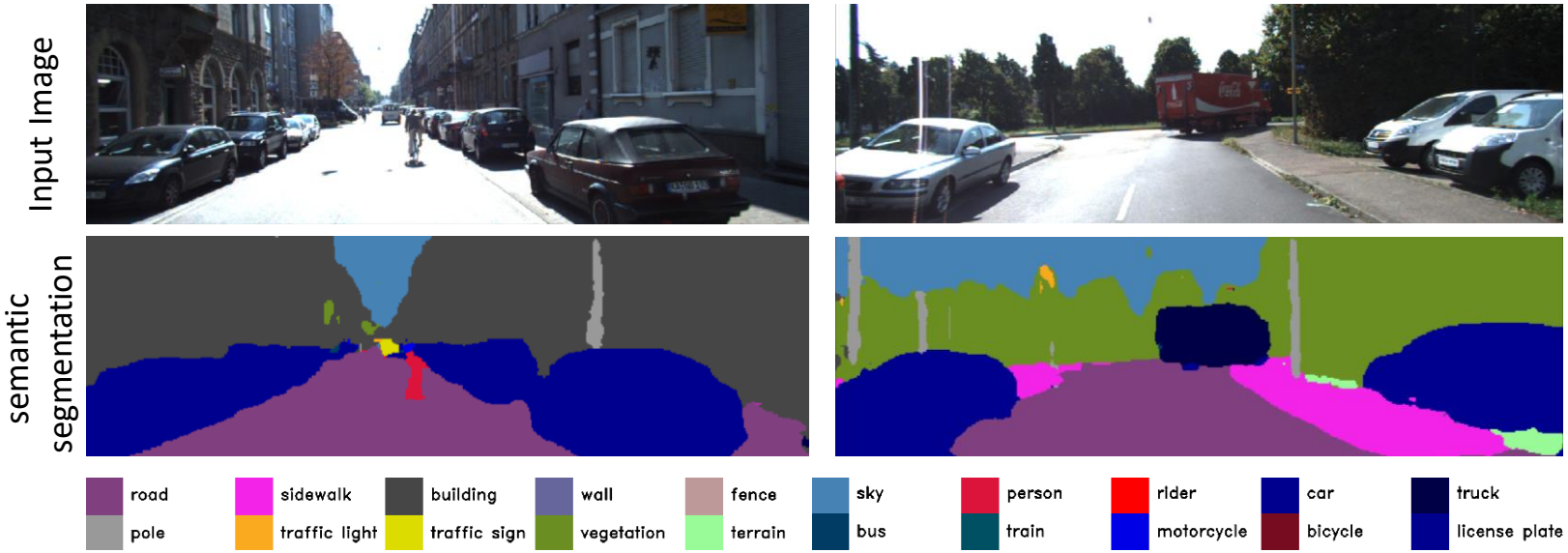
our depth prediction results

Pipeline Overview



Teacher Models (Semantic Segmentation)

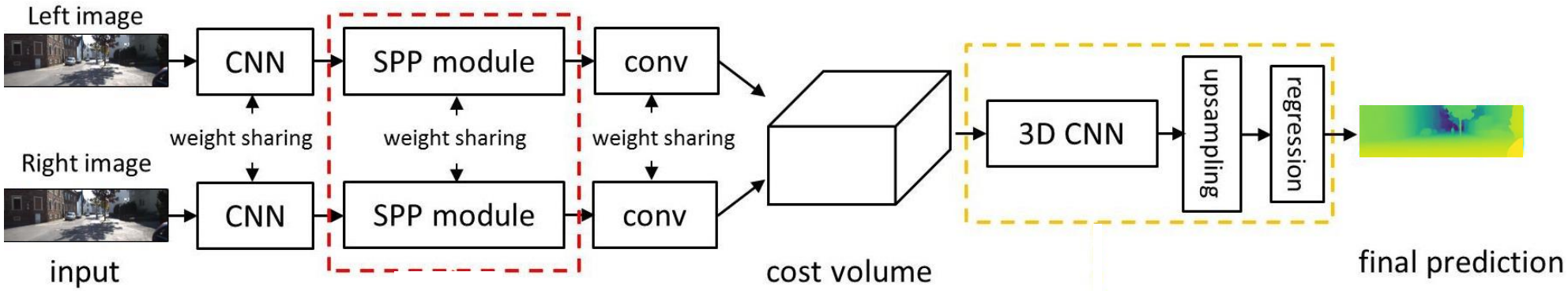
- Semantic segmentation
  - Using DeepLabV3 [11] to generate semantic segmentation as training ground truth



[11] Chen et al., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs., TPAMI 2018

Teacher Models (Depth Estimation)

- Depth Estimation
  - Use Pyramid Stereo Matching Network (PSMNet) [12] proposed by Chang et al. to generate dense disparity map

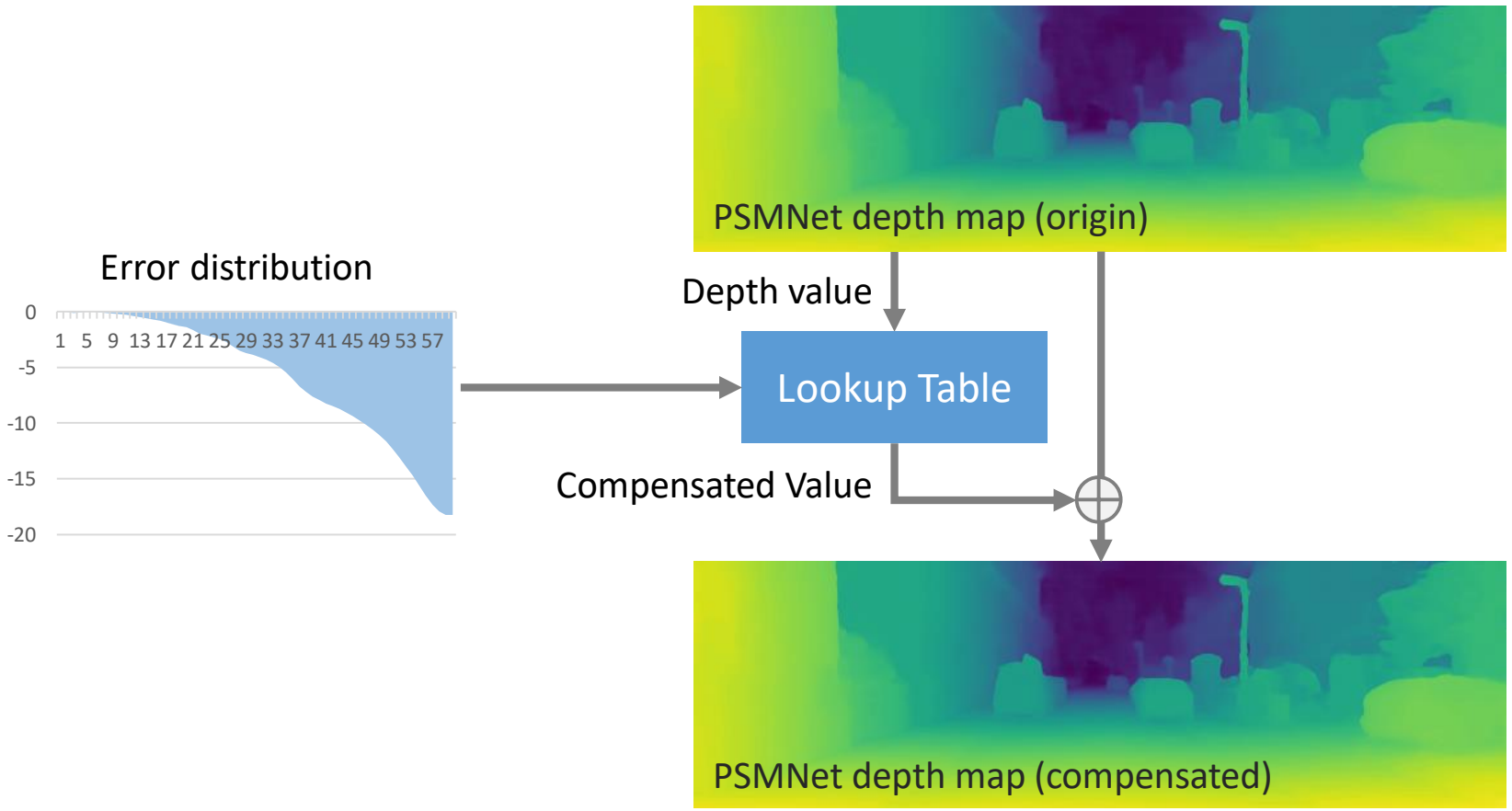


Input image | Generated depth map

[12] Chang et al., Pyramid Stereo Matching Network. CVPR 2018

Teacher Models (Depth Estimation)

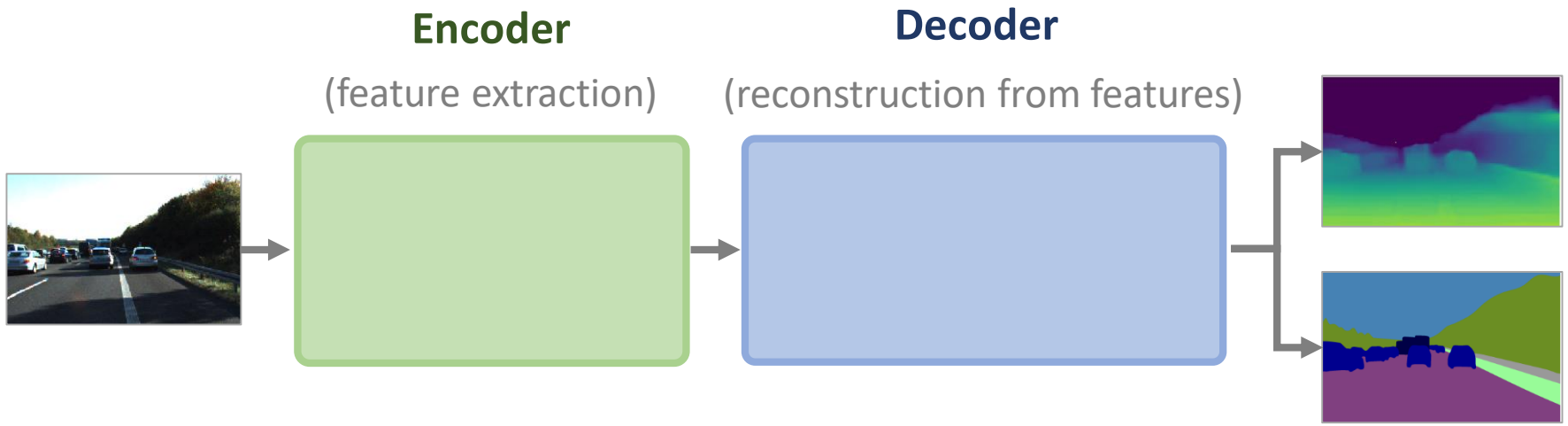
- Depth Estimation
  - Use Pyramid Stereo Matching Network (PSMNet) [12] proposed by Chang et al. to generate dense disparity map



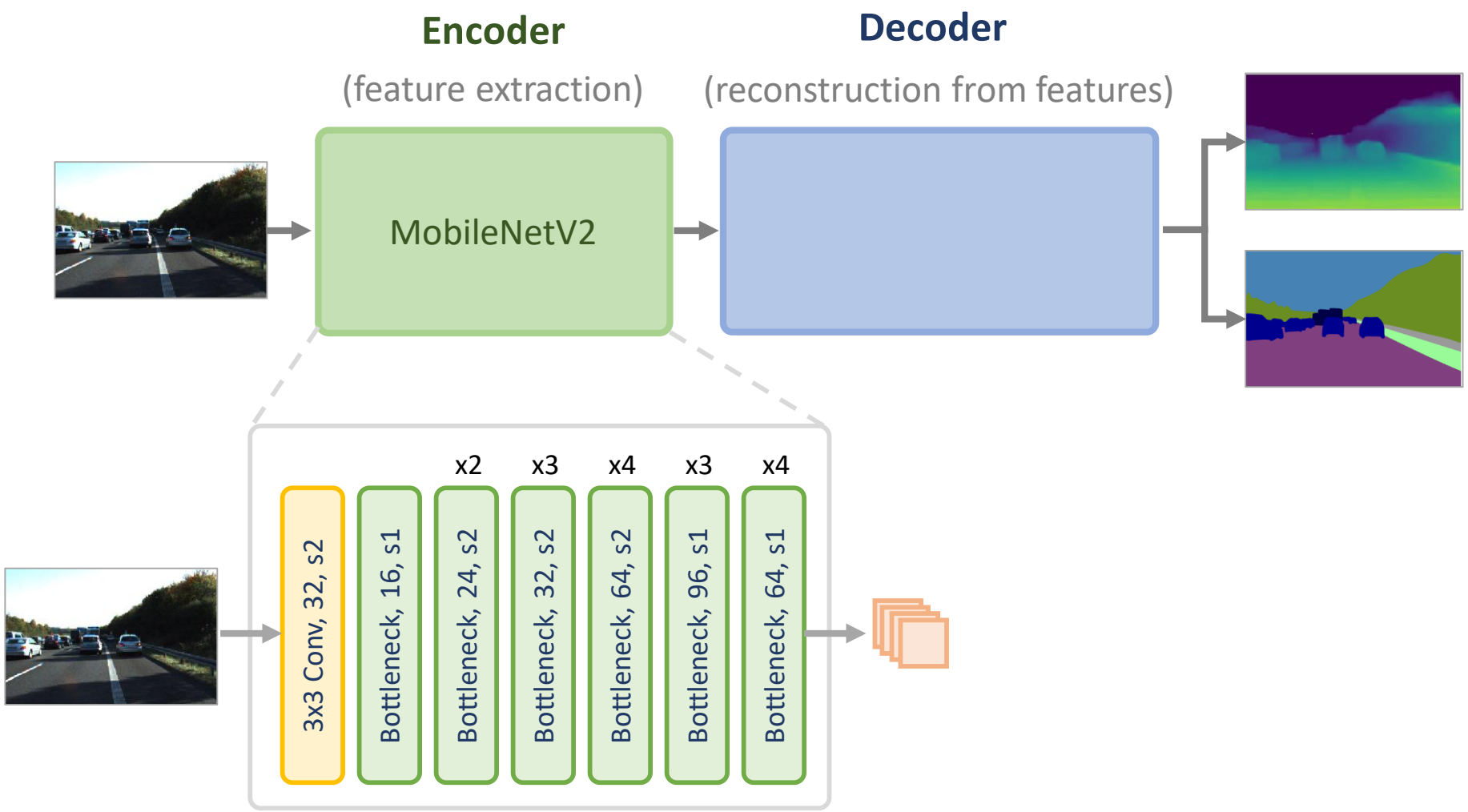
Min	5.64
Max	69.67
RMSE	0.96

Min	5.67
Max	87.32
RMSE	0.41

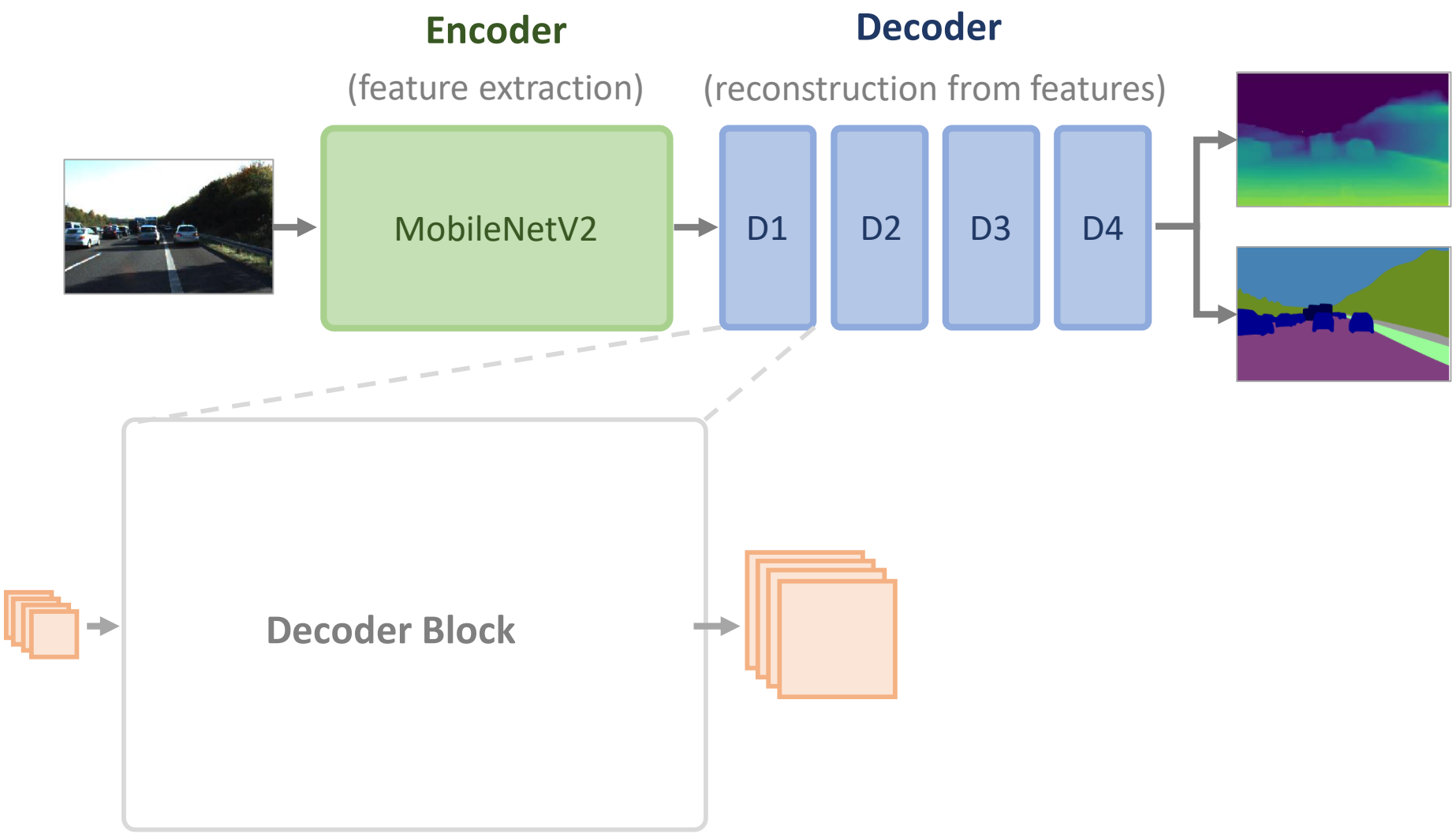
Network Architecture



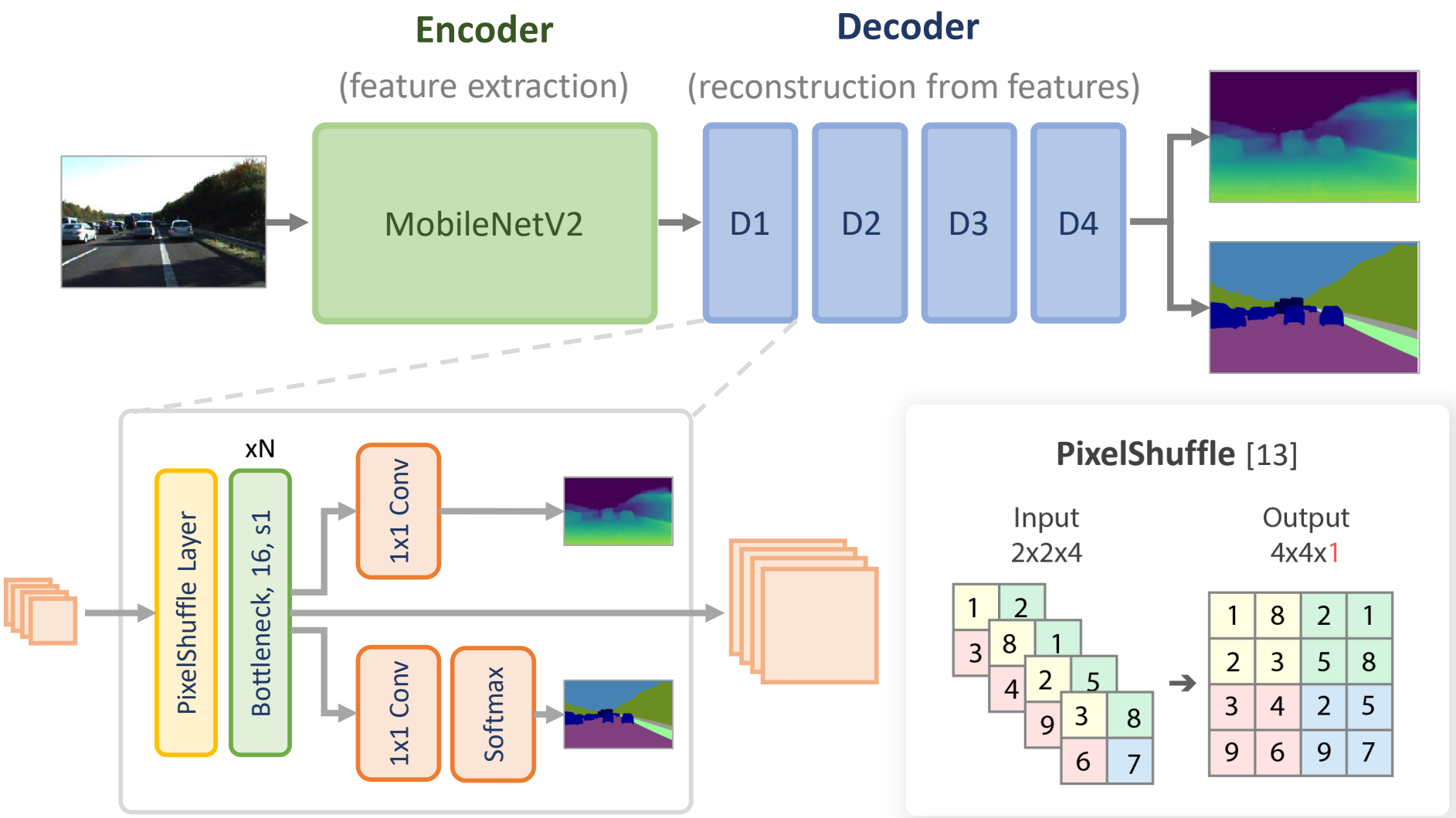
Network Architecture



Network Architecture

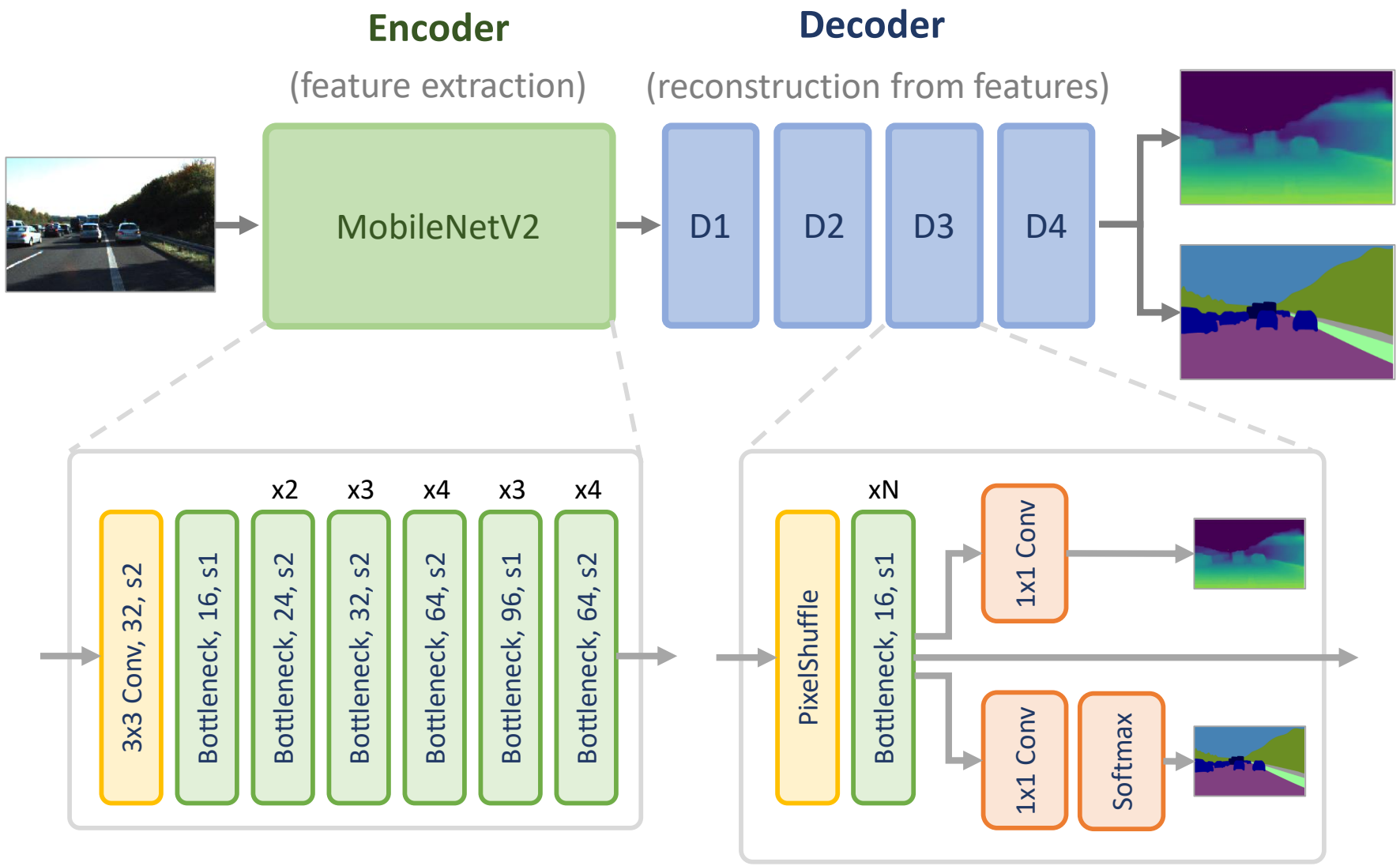


Network Architecture



[13] Shi et al., Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, CVPR 2016

Network Architecture



Loss Function

$$\mathcal{L}_{total} = \sum_{i=1}^{4} \alpha_i \mathcal{L}_{D_i} + \beta_i \mathcal{L}_{S_i}$$

Depth loss

$$\mathcal{L}_{D_i} = \frac{1}{W_i H_i} \sum_{x=1}^{W_i} \sum_{y=1}^{H_i} F\big(G(\tilde{d}_{x,y}) - G(d_{x,y})\big)$$

$$F(x) = \begin{cases} |x| & |x| \le \alpha \\ \dfrac{x^2 + \alpha^2}{2\alpha} & |x| > \alpha. \end{cases}$$

$$\alpha = \frac{1}{5} \max_i \left(\left| G(\tilde{d}_{x,y}) - G(d_{x,y}) \right| \right)$$

$$G(d) = \frac{(\log d - \log m) \times M}{\log M - \log m}; m = 4, M = 80$$

Semantic segmentation loss

$$\mathcal{L}_{S_i} = -\frac{1}{W_i H_i} \sum_{x=1}^{W_i} \sum_{y=1}^{H_i} \sum_{c=1}^{C} s_{x,y}^{c} \log \tilde{s}_{x,y}^{c}$$

(W_i, H_i): resolution of the depth map
$\tilde{d}$: predicted depth map
d: ground truth depth map
$\tilde{s}$: predicted semantic segmentation
s: ground truth semantic segmentation
C: number of classes of semantic segmentation

Ablation Study

Evaluation of models trained with differently pre-processed training data.

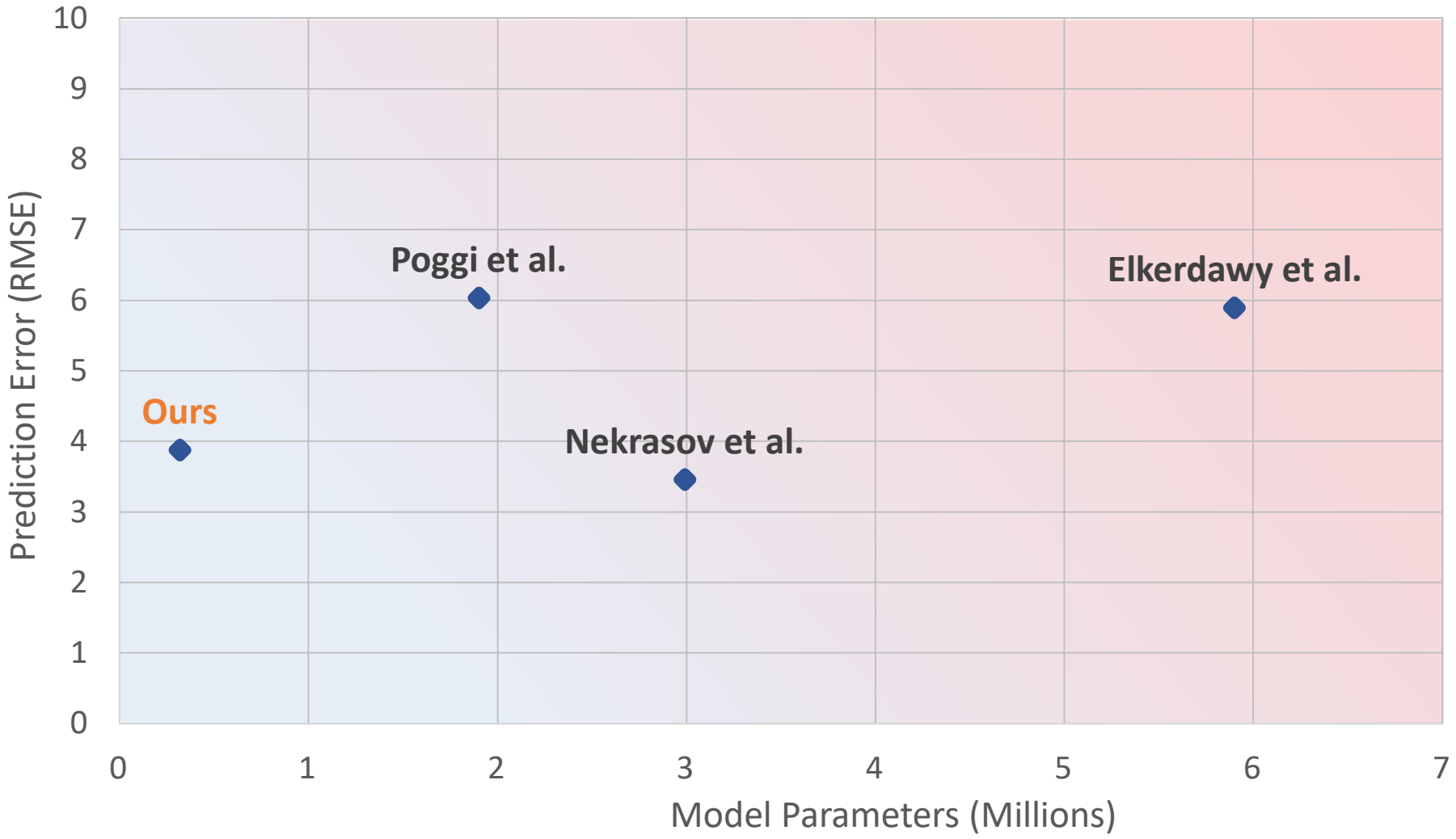
Data type	**Training data**	**RMSE (meter)**
Baseline	Sparse depth map	4.922
Pre-processed	PSMNet (origin)	4.716
	PSMNet (compensated)	3.945

Evaluation of models trained with differently pre-processed training data

Improvement			**RMSE (meter)**
Depth Teacher	*Log Depth*	*Segment Teacher*	
✓			3.945
✓	✓		3.884
✓	✓	✓	3.871

Performance Evaluation (Model Size)

Model Size vs. Accuracy



Performance Evaluation (Computation Speed)

Processing Speed Evaluation

Evaluation of our model on GTX 1060 and Jetson TX2

Model	Output dim.	1060GPU (FPS)		Jetson TX2 (FPS)		RMSE
		Before TRT	After TRT	Before TRT	After TRT	
L	$(240, 160)$	100	121.6	21	33.5	4.315
M	$(120, 80)$	126.5	148.7	31.5	42.8	4.344
S	$(60, 40)$	148.6	174.8	36.4	49.5	4.619
XS	$(30, 20)$	151.2	179.4	45.4	54.1	4.549

TRT: TensorRT

Conclusion

- Design an efficient CNN for depth estimation with only 2.1 GFLOPs computations and 0.3M parameters.
- Propose effective training strategies for such extremely small model:
  - (i) joint-training
  - (ii) data generation by complex teacher model
  - (iii) using a multi-resolution log depth loss
- The detachable structure enables model customization, offering the trade-off between output resolution and computation cost (speed).

Thank you for your attention

Q&A