Dynamic Guided Network for Monocular Depth Estimation

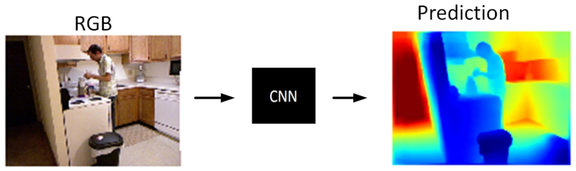
Xiaoxia Xing[1,2] Yinghao Cai [1] Yanqing Wang[1] Tao Lu[1] Yiping Yang [1] Dayong Wen[1]

Institute of Automation, Chinese Academy of Sciences,Beijing, China[1]

University of Chinese Academy of Sciences, Beijing, China[2]

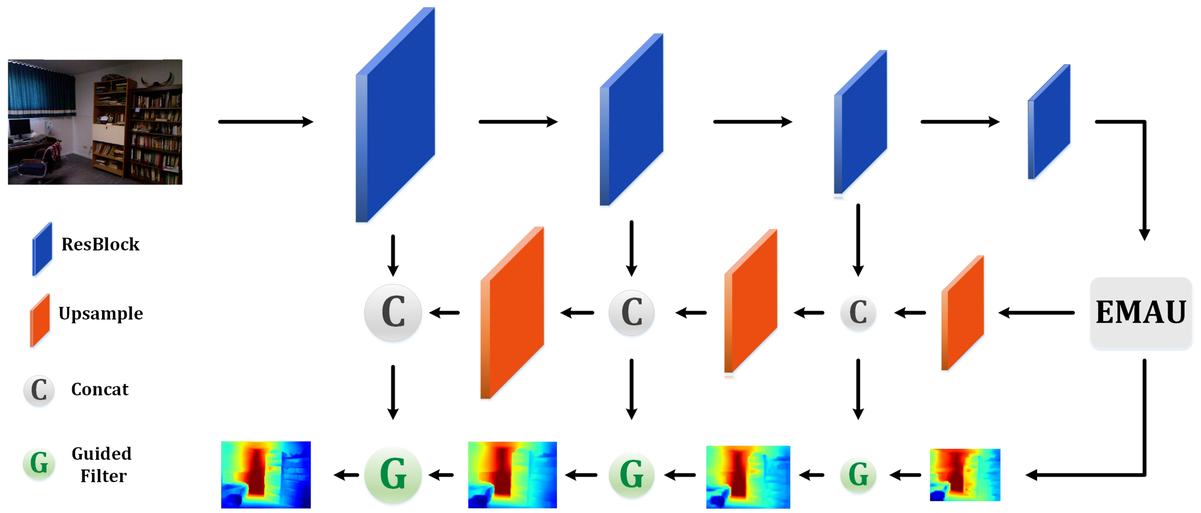
Introduction

Self-attention and encoder-decoder have been widely used in the deep neural network for monocular depth estimation. The self-attention mechanism is capable of capturing long-range dependencies, while the encoder-decoder structure can capture detailed structural information by gradually recovering spatial information. In this work, we combine the advantages of both methods. Specifically, our proposed model, DGNet, extends EMANet [1] by adding an effective decoder module to progressively refine the coarse depth map. In the decoder stage, we design a dynamic guided upsampling module that employs dynamically generated kernel conditioned on low-level features to guide the upsampling of the coarse depth map.



Using single RGB image to generate depth prediction.

Network Architecture



Overview of the proposed DGNet architecture. The network is composed of two sub-networks: Encoder in blue and Decoder in orange. Emau, end of the encoder, is the expectation maximization attention unit. The concatenation of encoder features and decoder features is used to progressively refine a coarse depth map by dynamic guided upsampling. The ResBlock represents the basic residual block structure. The upsampling block represents bilinear interpolation with two sequential 3×3 convolutional layers.

Future Work

We will further explore more effective guided cues such as surface normals or edges to improve the quality of the resultant depth maps.

References

[1] X. Li et al. Expectation-maximization attention networks for semantic segmentation.

[2] X. Wang et al. Non-local neural networks.

[3] M. Ramamonjisoa and V. Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation.

[4] J. Hu et al. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries.

[5] D. Eigen et al. Depth map prediction from a single image using a multi-scale deep network.

[6] I. Laina et al. Deeper depth prediction with fully convolutional residual networks.

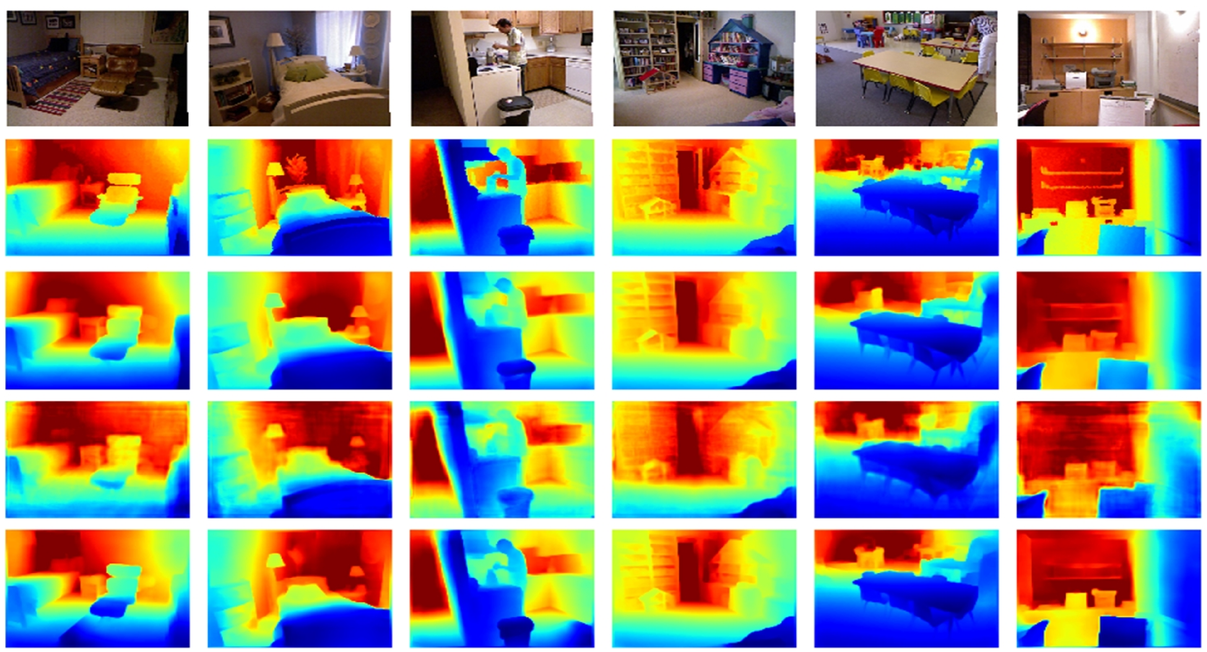
[7] H. Yan et al. Monocular depth estimation with guidance of surface normal map.

[8] B. Li et al. Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference.

[9] Y. Chen et al. Attention-based context aggregation network for monocular depth estimation.

[10] H. Fu et al. Deep ordinal regression network for monocular depth estimation.

Visualization of Results



Example depth results of different methods. From the first to the last row: RGB input, Ground truth, Hu et al. [4], Ramamonjisoa and Lepetit [3], Our result; Note the color range of each image is individually scaled.

Results

Method	RMS↓	REL↓	log_{10} ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Eigen and Fergus [5]	0.907	0.215	-	0.611	0.887	0.971
Laina et al. [6]	0.573	0.127	0.055	0.811	0.953	0.988
Yan et al. [7]	0.502	0.135	0.056	0.813	0.965	0.993
Li et al. [8]	0.505	0.139	0.058	0.820	0.960	0.989
Chen et al. [9]	0.496	0.138	-	0.826	0.964	0.990
Fu et al. [10]	0.509	0.115	0.051	0.828	0.965	0.992
Hu et al. [4]	0.555	0.126	0.054	0.843	0.968	0.991
Ramamonjisoa and Lepetit [3]	0.496	0.148	0.048	0.884	0.980	0.995
Ours	0.525	0.119	0.051	0.863	0.973	0.993

Results on the NYU Depth v2 dataset.

Ablation Study Results

We compare different decoder methods with our dynamic guidance upsampling module. **EMANet** applies several bilinear upsamplings to the size of ground truth. **De-Unet** uses skip connections to combine low-level and high-level features like Unet.

Method	RMS↓	REL↓	log_{10} ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
EMANet	0.5309	0.1204	0.0512	0.859	0.970	0.992
De-Unet	0.5264	0.1196	0.0509	0.861	0.971	0.992
Ours	0.5257	0.1197	0.0510	0.863	0.973	0.993