AUTODEPTH: SINGLE IMAGE DEPTH MAP ESTIMATION VIA RESIDUAL CNN ENCODER-DECODER AND STACKED HOURGLASS

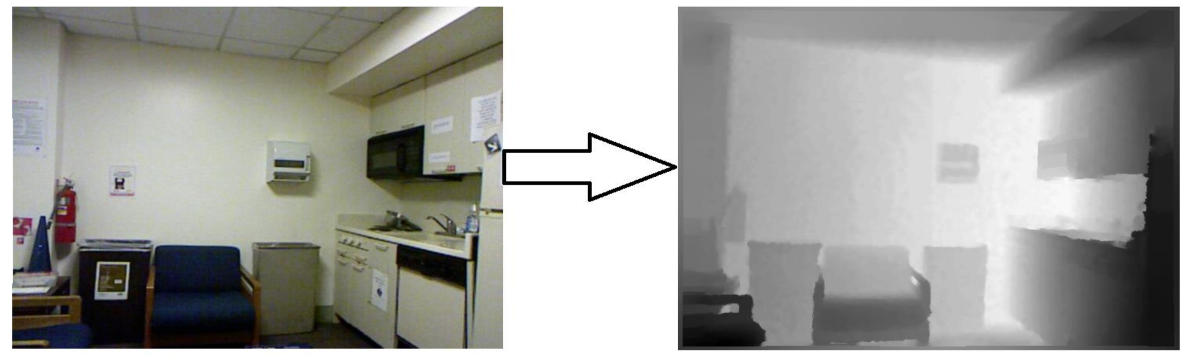
Seema Kumari, Ranjeet Ranjhan Jha, Arnav Bhavsar and Aditya Nigam

SCEE, Indian Institute of Technology Mandi, India

ICIP 2019

OBJECTIVE

- The objective is to estimate depth from a single intensity image.



- Active sensors: Laser depth scanners, time-of-flight cameras, active pattern sensors etc.
- Passive techniques: stereo, structure from motion, depth from defocus etc.
- Depth maps are useful in various 3D based applications such as automatic driving assistance, robotic navigation, 3D television, scene classification, dehazing, object recognitions etc.

RELATED WORK

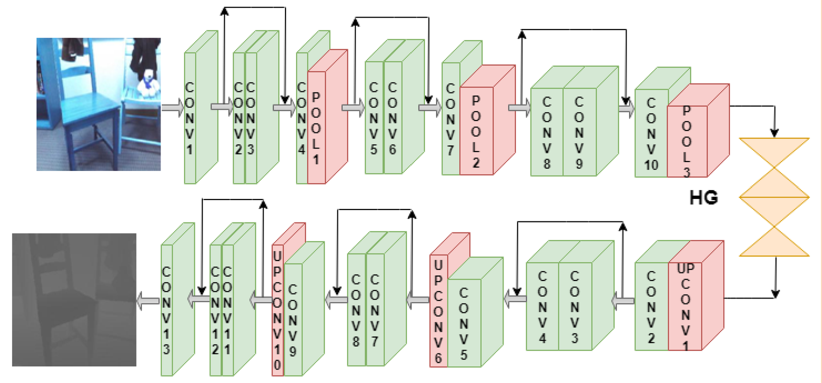
- Multi-scale deep network [1, 2].
- Fully convolutional neural network (FCNN) [3].
- Deep CNN with continuous random fields [4, 5].
- Deeper residual convolutional neural network [6].
- Auto-encoder with skip connection convolutional neural network embedding focal length [7].

CONTRIBUTION

- Proposed stacked hourglass module in the encoder-decoder architecture for estimating the depth map.
- To optimize the network, we have used perceptual loss along with the mean squared error loss.
- Depth estimation in presence of noise in input intensity image

PROPOSED APPROACH

- Block diagram of our network for depth estimation is consisted of multiple stacked layers with hourglass in encoder-decoder.



HOURGLASS

- The hourglass module is used to incorporate features from different scales.
- The residual blocks are labeled as $R1$, $R2$, … $R9$, each of which consists of three convolutional layers.

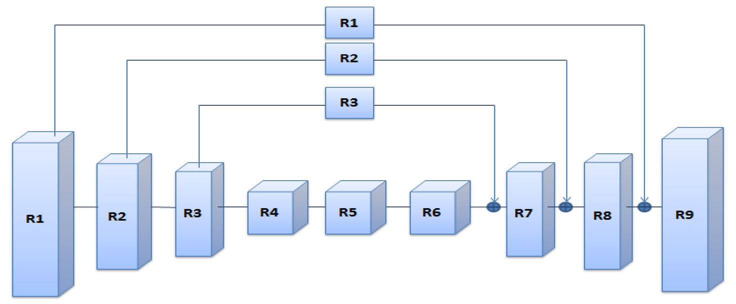


Figure. Block diagram of an hourglass module

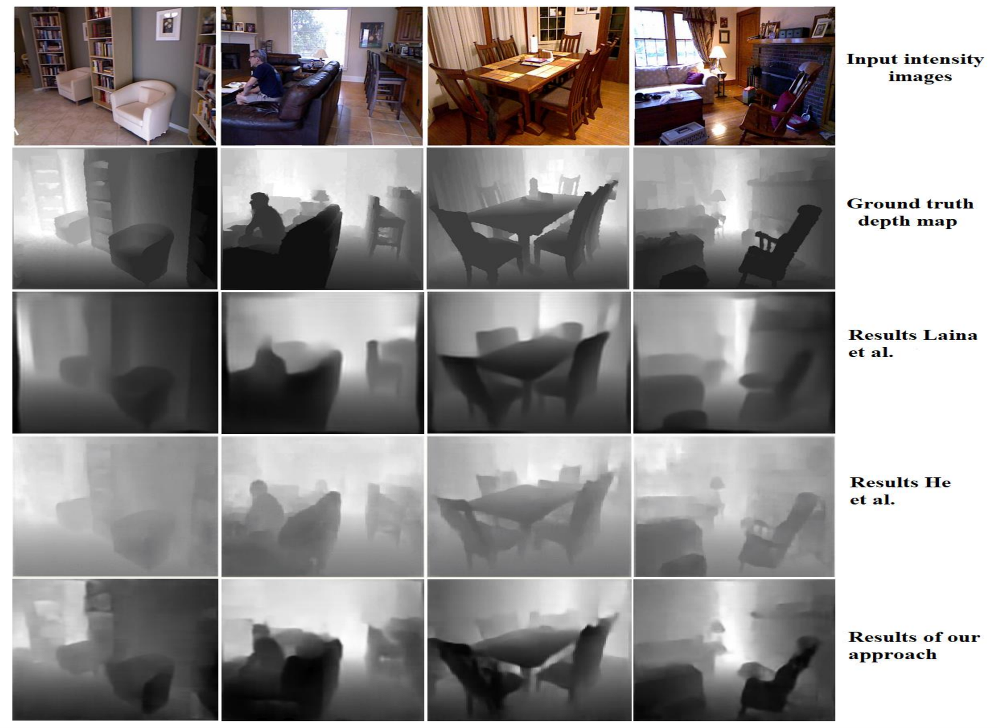
LOSS FUNCTION

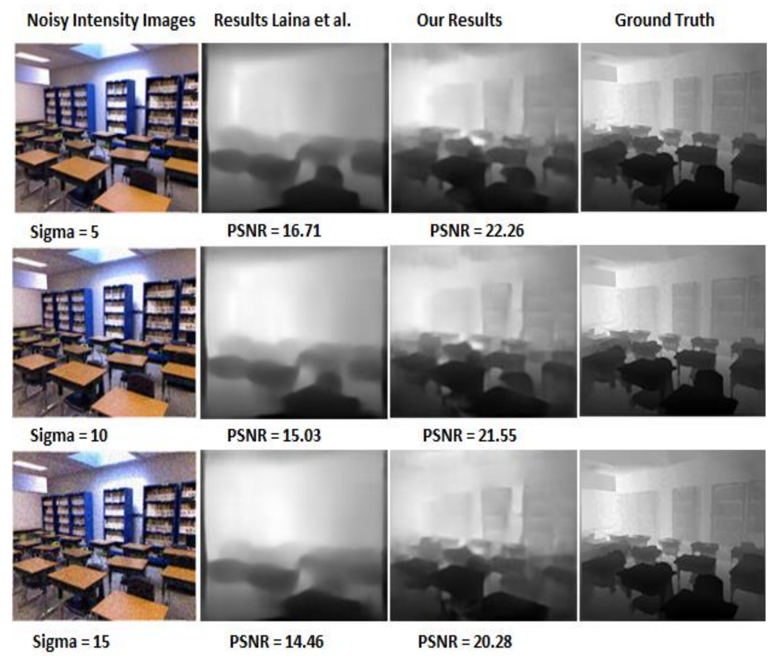
- Our loss function can be represented the following loss function:

$$L(\hat{x},\ x) = D^{j}_{feat}(\hat{x}, x) + \frac{1}{2} MSE(\hat{x}, x)$$

- This combination of perceptual loss as well as MSE loss increases accuracy and improved the perceptual quality of the predicted depth map.

EXPERIMENTATION





QUALITATIVE RESULTS

Table: Quantitative comparison of results on the Ikea chair dataset

Method	Rel	rms	log_{10}	$a_1 < 1.25$	$a_2 < 1.25^2$	$a_3 < 1.25^3$
FCNN [3]	0.413	1.128	0.165	0.370	0.647	0.828
Ours	0.194	0.625	0.065	0.762	0.893	0.942

Table: Quantitative comparison of results on the NYU V2 dataset

Method	Rel	rms	log_{10}	$a_1 < 1.25$	$a_2 < 1.25^2$	$a_3 < 1.25^3$
Eigen et al. [1]	0.215	0.907	-	0.611	0.887	0.887
Roy. et al. [5]	0.187	0.744	0.078	-	-	-
E. & F. [2]	0.158	0.641	-	0.769	**0.950**	**0.988**
Laina et al. [6]	0.194	0.790	0.083	-	-	-
He et al. [7]	0.151	0.572	**0.064**	**0.789**	0.948	0.986
Ours	**0.104**	**0.324**	0.065	0.787	0.946	0.987

ABLATION STUDY

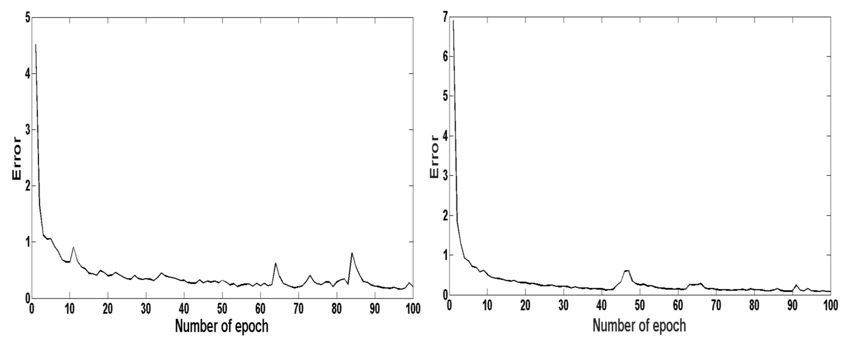


Figure: Convergence plot with respect to epoch, in left without perceptual loss and right with perceptual loss.

Figure: Visual results on NYU V2 dataset: First left - intensity image, Second left - result without perceptual loss, Third left - result with perceptual loss, Fourth column - ground truth.

Figure: Visual results on NYU V2 dataset: First left - intensity image, Second left - result without using hourglass model, Third left - result with hourglass, Fourth column - ground truth

REFERENCE

[1] David Eigen, Christian Puhrsch, and Rob Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in NIPS*, pp. 2366–2374, 2014.
[2] David Eigen and Rob Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *ICCV*, pp. 2650–2658, 2015.
[3] A. J. Afifi et al., "Object depth estimation from a single image using fully convolutional neural network," in *DICTA*, pp.1–7, 2016.
[4] Fayao Liu et al., "Deep convolutional neural fields for depth estimation from a single image," in *CVPR*, pp. 5162–5170, 2015.
[5] A. Roy et al., "Monocular depth estimation using neural regression forest," in *CVPR*, pp. 5506–5514, 2016.
[6] Iro Laina et al., "Deeper depth prediction with fully convolutional residual networks, in *3D Vision (3DV)*, pp. 239–248, 2016.
[7] Lei He et al., "Learning depth from single images with deep neural network embedding focal length," in *IEEE Transactions on Image Processing*, vol. 27, pp. 4676–4689, 2018.