ROBUST FULL-FOV DEPTH ESTIMATION IN TELE-WIDE CAMERA SYSTEM

Kai Guo, Seongwook Song, Soonkeun Chang, Tae-ui Kim, Seungmin Han and Irina Kim

Corresponding author: Kai Guo, visionkai@gmail.com

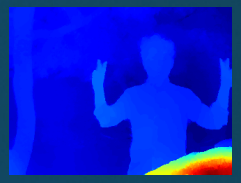
Samsung electronics, South Korea

ICASSP 2020,
May 4-8, 2020

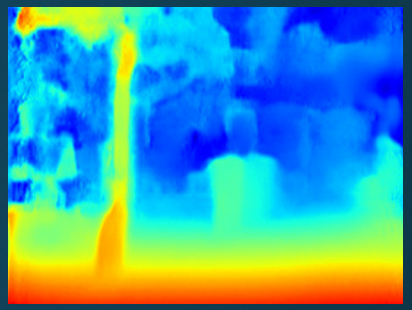
Depth estimation in tele-wide camera system

- Tele-wide camera system becomes popular in current mobile devices

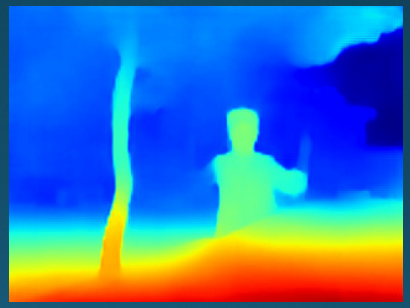
- Various depth estimation



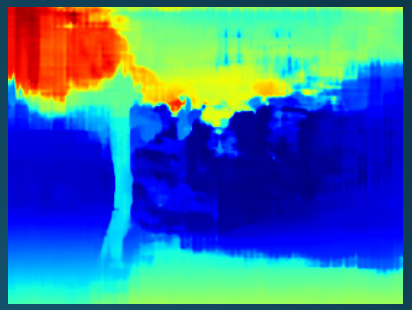
Traditional stereo matching [3]+[7]



DNN based single image depth [11]



DNN based single image depth [13]



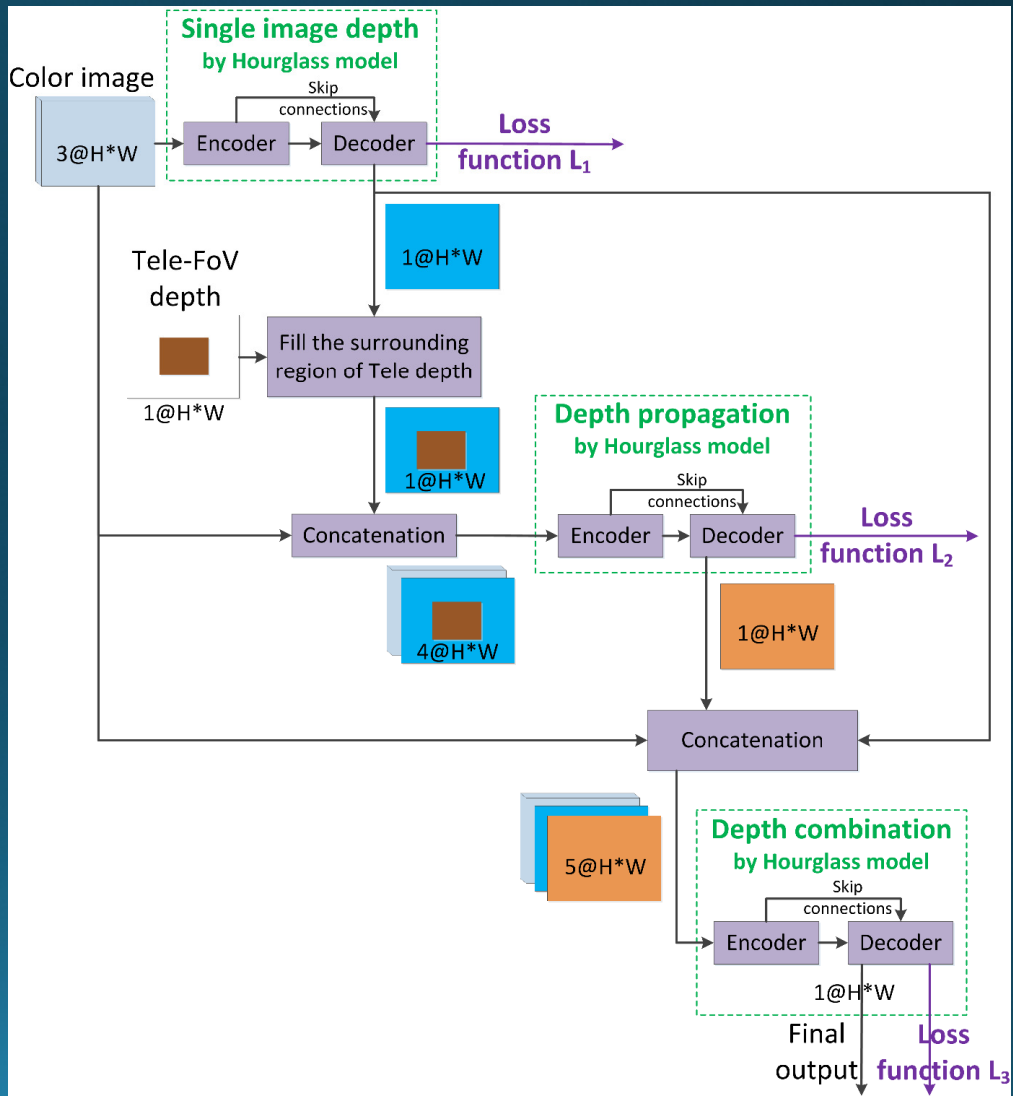
DNN based tele-wide depth [14]

- Usually it is difficult to obtain full-FoV depth based on traditional stereo-matching methods.
- Pure Deep Neural Network (DNN) based depth estimation methods can obtain full-FoV depth, but have low robustness for scenarios which are not covered by training dataset.

Architecture of the proposed Hierarchical Hourglass network

- A hierarchical hourglass network in tele-wide camera system
  - Combines the robustness of traditional stereo-matching methods with the accuracy of DNN.

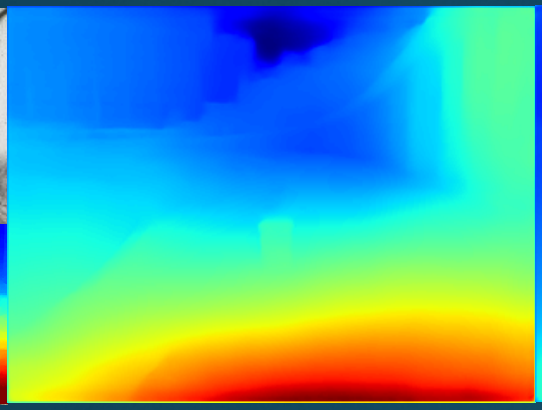
- The proposed network comprises three major modules:
  - Single image depth prediction module infers initial depth from input color image,
  - Depth propagation module propagates traditional stereo-matching tele-FoV depth to surrounding regions,
  - Depth combination module fuses the initial depth with the propagated depth to generate final output.
  - Each of these modules employs an hourglass model [17], which is a kind of encoder-decoder structure with skip connections.



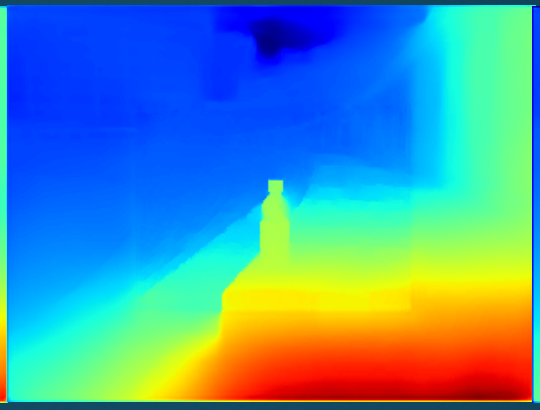
Architecture of the proposed Hierarchical Hourglass network

- Effectiveness of each module

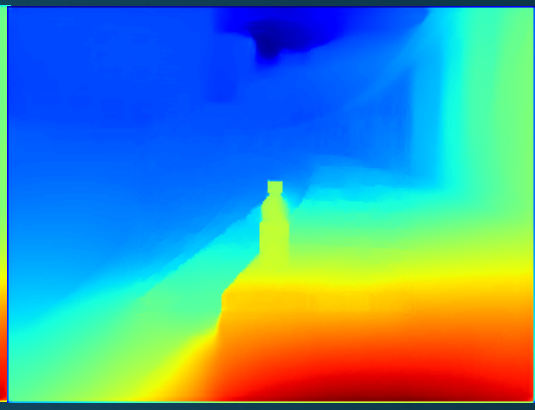
(a) Input wide image and stereo depth [3] + [7]



(b) Intermediate result of single image depth module



(c) Intermediate result of depth propagation module



(d) Final depth after depth combination module

- The single image depth module can predict the global structure but lack of details (b), especially for uncommon objects which are not covered by training dataset.
- The depth propagation module would refine the stereo depth at tele-FoV region, at the same time propagate it to surrounding regions, but has slight discontinuity artifact at tele-FoV boundary (c).
- The depth combination module will fuse the initial depth with propagated depth to generate better result, and smooth out the aforementioned discontinuity artifact (d).

Loss function of the proposed Hierarchical Hourglass network

- L1-norm scale-invariant loss function

The loss function is weighted sum of these modules loss functions

$$L = w_1 L_1 + w_2 L_2 + w_3 L_3$$

where L is the final loss, L_1, L_2 and L_3 are the loss function of single image depth, depth propagation and depth combination modules, respectively. w_1, w_2 and w_3 are the weights, they are set as 0.5, 0.5 and 1, respectively

For loss function L_k, k = 1, 2, 3, we propose a L1- norm scale-invariant loss function, which regulates predicted log depth to have similar between-points relationships with ground truth. Compared with widely-used L2 norm, L1 norm is robust and less sensitive to outliers [19]. It is written as:

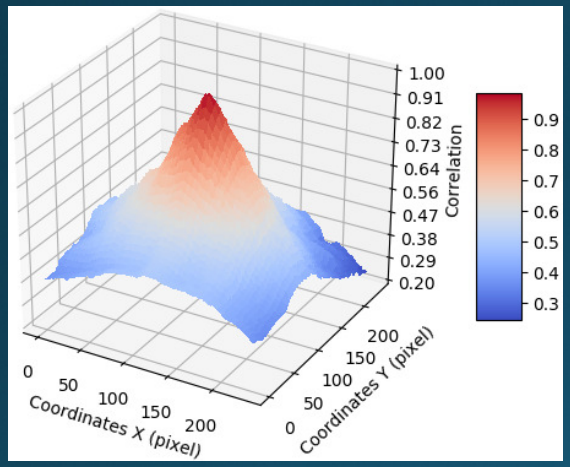
$$\begin{aligned} L_k &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\left(P_k^i - P_k^j\right) - \left(T^i - T^j\right)\right| \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\left(P_k^i - T^i\right) - \left(P_k^j - T^j\right)\right| \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\left(D_k^i - D_k^j\right)\right| \end{aligned}$$

where P^i_k and P^j_k are predicted log depth of module k at pixel position i and j, respectively. T_i and T_j are ground-truth log depth at pixel position i and j, respectively. N is the total number of pixels. D_k is the deviation between prediction P_k and ground truth T, e.g. $D_k = P_k - T$

Loss function of the proposed Hierarchical Hourglass network

- Accelerated calculation of L1-norm scale-invariant loss function

Direct calculating the absolute difference of deviations $|D^i_k - D^j_k|$ on all possible pixels pairs is quite time consuming. To accelerate calculation, we compute absolute difference of deviations only between each pixel and its neighboring pixels, because of low correlations between a pixel and other spatially distant pixels of depth map (as shown in the figure below).



Correlation between the center pixel and other pixels of depth map. This correlation is calculated based on the center 240 × 240 region of resized 320 × 240 depth maps from NYU depth V2 dataset [18].

Then the L1-norm scale-invariant loss function can be rewritten

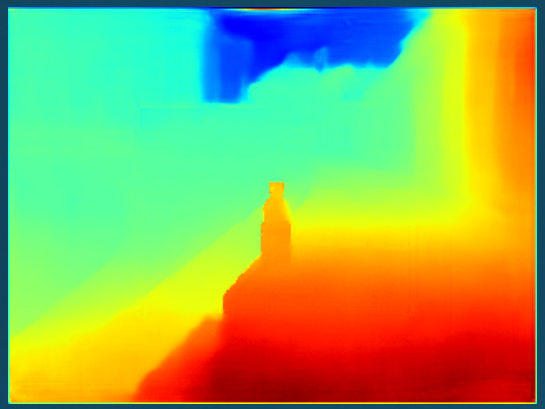
$$L_k = \frac{1}{N \times M} \sum_{i=1}^{N} \sum_{m=1}^{M} \left| D_k^i - D_k^{im} \right|$$

where m is the neighboring pixels index of the pixel i, and M is the number of neighboring pixels. Absolutely the larger neighborhood would produce better results. Considering time complexity, we set neighborhood as a 17 × 17 window.

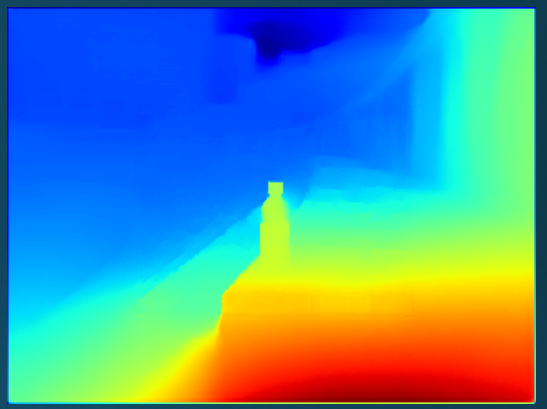
Loss function of the proposed Hierarchical Hourglass network

- Effectiveness of L1-norm scale-invariant loss function

(a) Wide image



(b) Our network with widely-used L2-norm scale-invariant loss



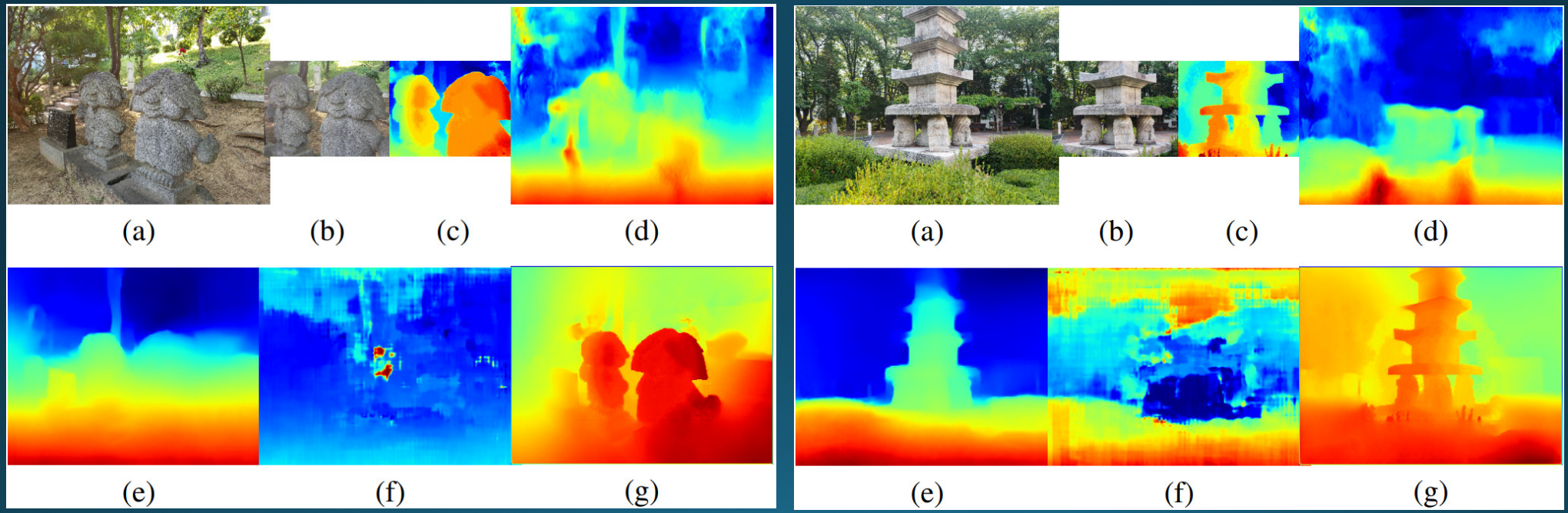
(c) Our network with proposed L1-norm scale-invariant loss

To justify the effectiveness of the proposed L1-norm scale-invariant loss function, we compare it with the result of our network which employs widely-used L2-norm scale invariant loss function [8], as shown in (b) (c) in above figure. It can be observed that the L1-norm loss function can generate better global structure, especially for background.

Experiments

- Wild test images

We capture several test images at various scenarios by tele-wide camera of Galaxy S9 plus. MegaDepth [13] is employed as training dataset. All of the images with their depth are scaled to 240 × 320, and the center 120 × 160 region is set as tele FoV. During test, the tele-FoV stereo-matching depth is obtained by the traditional stereo matching [3] + post processing [7].

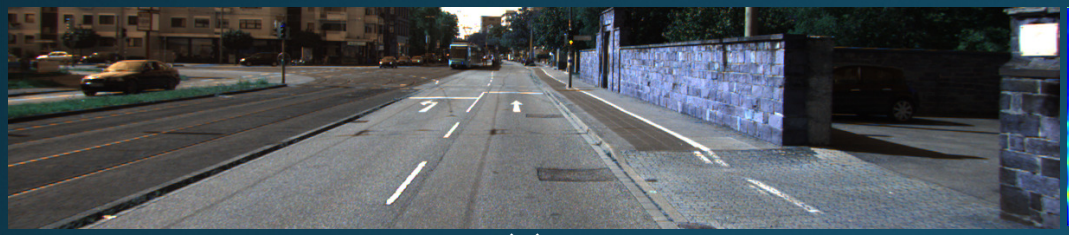


Depth comparisons. (a) Wide image; (b) Tele image; (c) Traditional tele-FoV stereo-matching method: matching cost calculation/optimization [3] + post processing [7]; (d) DNN-based single image depth method in [11]; (e) DNN-based single image depth method in [13]; (f) Pure DNN-based tele-wide stereo matching method in [14]; (g) Our result.

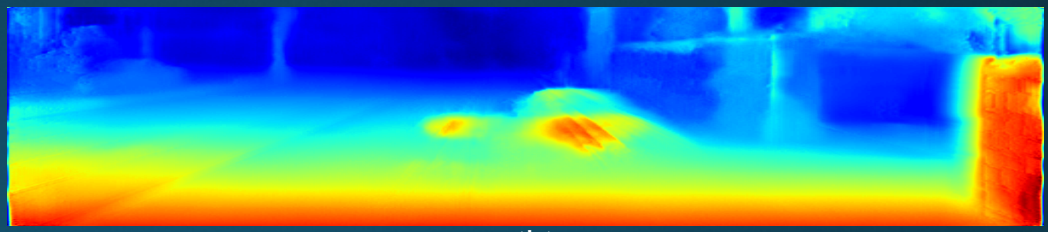
Experiments

- KITTI dataset

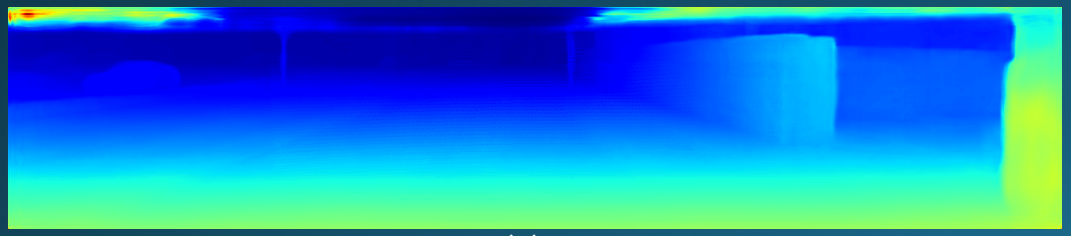
For KITTI dataset [22], we use the 22600 training images from 28 scenes and 697 test images from another 29 scenes based on Eigen split [8]. A 256 × 1216 region is horizontally-random and vertically-bottom cropped from each image for training and testing, wherein the center 128 × 608 region of ground truth is used as tele-FoV depth for training. To clearly know the maximum benefit of our method regardless of stereo depth quality, we use the center 128 × 608 region of ground-truth depth as tele-FoV depth for test. Because all of the training and testing images are captured by the same device, we can combine the proposed L1-norm scale-invariant loss function with the common L1 norm loss function to get better results.



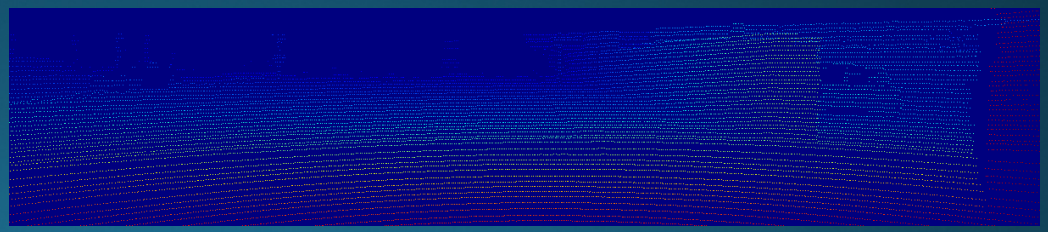
(a)



(b)



(c)



(d)

Depth comparisons on KITTI test images. (a) Reference wide image; (b) DNN method in [11]; (c) Our result; (d) Ground truth.

Experiments

- KITTI dataset

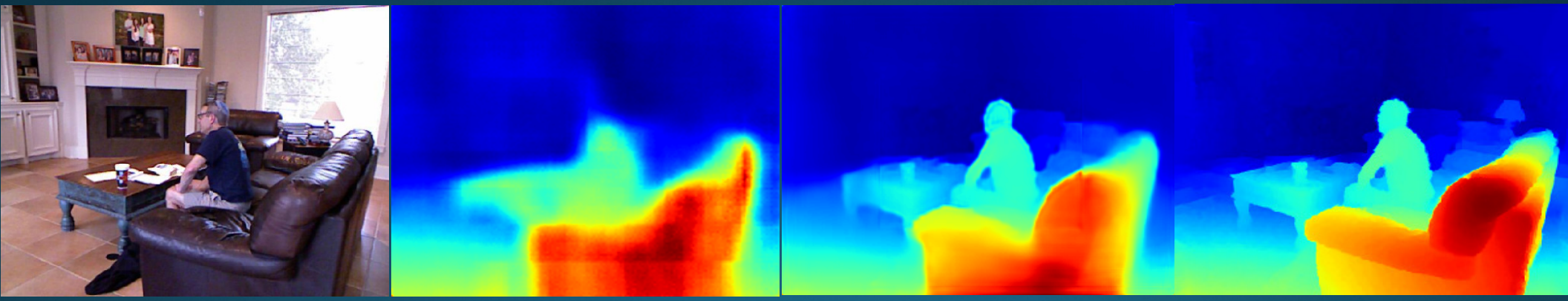
Performance comparison on KITTI dataset. RMSE: root mean squared error; REL: mean absolute relative error; δ_i : percentage of predicted pixels where the relative error is within a threshold 1.25^i [20]

	Lower is better		Higher is better		
Method	RMSE	REL	δ_1	δ_2	δ_3
Mancini [21]	7.508	-	31.8	61.7	81.3
Eigen et al. [8]	7.156	0.190	69.2	89.9	96.7
Ma et al. [20]	6.266	0.208	59.1	90.0	96.2
Godard et al. [11]	5.927	0.148	80.3	92.2	96.4
Our method	**2.440**	**0.05**	**95.2**	**98.3**	**99.3**

Experiments

- NYU dataset

For the NYU Depth V2 dataset [18], we use 249 scenes for training, and 654 images [8, 10] of the rest 215 scenes for testing based on official split. All of the images with their depth maps are scaled to 224×304, and the center 112×152 region of depth is used as tele-FoV depth for training. We use ground-truth tele-Fov depth together with color image as input for test, not only because the NYU dataset has only single image without stereo pairs, but also because we want to know the maximum benefit of our model regardless of stereo depth quality. Because all of the training and testing images of NYU dataset are captured by the same device, we can use the combined loss function from the proposed L1-norm scale-invariant loss and the common L1 norm loss (same as KITTI dataset).



(a) (b) (c) (d)

Depth comparisons on NYU test images. (a) Reference image; (b) DNN method in [20]: RGB image as input; (c) Our result: RGB image + tele-FoV depth as input; (d) Ground truth.

Experiments

- NYU dataset

Performance comparison on NYU Depth V2 dataset. RMSE: root mean squared error; REL: mean absolute relative error; δ_i : percentage of predicted pixels where the relative error is within a threshold 1.25^i [20]

	Lower is better		Higher is better		
Method	RMSE	REL	δ_1	δ_2	δ_3
Roy et al. [23]	0.744	0.187	-	-	-
Eigen et al. [8]	0.641	0.158	76.9	95.0	98.8
Laina et al. [10]	0.573	0.127	81.1	95.3	98.8
Ma et al. [20]	0.514	0.143	81.0	95.9	98.9
Our method	**0.334**	**0.087**	**92.4**	**98.0**	**99.4**

Conclusion

We introduced a hierarchical hourglass network for robust full-FoV depth estimation in tele-wide camera system, which combines the robustness of traditional stereo-matching methods with the accuracy of DNN methods. Experiments demonstrate its robustness and better quality in both subjective and objective evaluations. We believe this new method opens up a door for research on combining robustness of traditional signal processing into deep learning for depth estimation. In the future, we will investigate new network structure and extend our framework into other computer vision problems.

Thanks