

Deep Modular Network Architecture for Depth Estimation from Single Indoor Images

Seiya Ito*, Naoshi Kaneko*, Yuma Shinohara, and Kazuhiko Sumi

Aoyama Gakuin University, Kanagawa, Japan
{ito.seiya, shinohara.yuma}@vss.it.aoyama.ac.jp,
{kaneko, sumi}@it.aoyama.ac.jp

Abstract. We propose a novel deep modular network architecture for indoor scene depth estimation from single RGB images. The proposed architecture consists of a main depth estimation network and two auxiliary semantic segmentation networks. Our insight is that semantic and geometrical structures in a scene are strongly correlated, thus we utilize global (i.e. room layout) and mid-level (i.e. objects in a room) semantic structures to enhance depth estimation. The first auxiliary network, or *layout network*, is responsible for room layout estimation to infer the positions of walls, floor, and ceiling of a room. The second auxiliary network, or *object network*, estimates per-pixel class labels of the objects in a scene, such as furniture, to give mid-level semantic cues. Estimated semantic structures are effectively fed into the depth estimation network using newly proposed *discriminator networks*, which discern the reliability of the estimated structures. The evaluation result shows that our architecture achieves significant performance improvements over previous approaches on the standard NYU Depth v2 indoor scene dataset.

Keywords: Depth estimation · Convolutional neural network

1 Introduction

Depth estimation is one of the fundamental problems of 3D scene structure analysis in computer vision. Traditional approaches including structured lights [27], time-of-flight [8], multi-view stereo [28], and structure from motion [3] have been extensively studied for decades. Most of these approaches are built upon stereo geometry and rely on reliable correspondences between multiple observations.

In contrast, depth estimation from a single RGB image is a relatively new task and has been actively studied for the last decade. Without prior knowledge or geometrical assumption, the problem is known as ill-posed, since numerous real-world spaces may produce the same image measurement. Early studies tackled this problem using Markov Random Fields (MRF) to infer the depth values of image patches [25] or the planar parameters of superpixels [26]. Later, approaches based on Conditional Random Fields (CRF) emerged [22, 36]. More

* The authors assert equal contribution and joint first authorship.

recently, the task has enjoyed rapid progress [1, 4, 5, 13, 20, 21, 23, 33, 37] thanks to the recent advances of deep architectures [16, 31] and large-scale datasets [9, 24, 30]. Without a doubt, the deep architectures, especially Convolutional Neural Networks (CNN), greatly contribute to the performance boost.

In this paper, we propose a novel deep modular network architecture for monocular depth estimation of indoor scenes. Our insight is that semantic and geometrical structures in a scene are strongly correlated, therefore, we use the semantic structures to enhance depth prediction. Interestingly, while the insight itself is not new [12], there are relatively few works [13, 33] that use both deep architectures and semantic structure analysis. We will show that the proposed architecture, which effectively merges the semantic structure into the depth prediction, clearly outperforms previous approaches on the standard NYU Depth v2 benchmark dataset.

The proposed architecture is composed of a main depth estimation network and two auxiliary semantic segmentation networks. The first auxiliary network, or *layout network*, gives us the global (i.e. room layout) semantic structure of a scene by inferring the positions of walls, floor, and ceiling. The second auxiliary network, or *object network*, provides the mid-level (i.e. objects in a room) cues by estimating per-pixel class labels of the objects. To effectively merge the estimated structures into the depth estimation, we also introduce *discriminator networks*, which discern the reliability of the estimated structures. Each semantic structure is weighted by the respective reliability score and this process reduces the adverse effect on the depth estimation when the estimation quality of semantic segmentation is insufficient.

To summarize, we present:

- A novel deep modular network architecture which considers global and mid-level semantic structures.
- Discriminator networks to effectively merge the semantic structures into the depth prediction.
- Significant performance improvements over previous methods on the standard indoor depth estimation benchmark dataset.

2 Related Work

One of the first studies of single image depth estimation was done by Saxena et al. [25]. This method used hand-crafted convolutional filters to extract a set of texture features from an input image and solved the depth estimation problem using Markov Random Fields (MRF). The authors later proposed Make3D [26] to estimate 3D scene structure from a single image by inferring a set of planar parameters for superpixels using MRF. This approach depends on the horizontal consistency of the image and suffers from lack of versatility.

Instead of directly estimating depth, Hoiem et al. [12] assembled a simple 3D model of a scene by classifying the regions in an image as a geometrical structure such as sky or ground, and indirectly estimating the depths of the image. Liu et al. [19] used predicted semantic labels of a scene to guide depth estimation and

solved a simpler problem with MRF. Ladicky et al. [18] proposed a pixel-wise classification model to jointly estimate the semantic labels and the depth of a scene, and showed that semantic classification and depth estimation can benefit each other.

Liu et al. [22] proposed a discrete-continuous Conditional Random Fields (CRF) model to consider relationships between neighbouring superpixels. Zhuo et al. [36] extended the CRF model to a hierarchical representation to model local depth jointly with mid-level and global scene structures. These methods lack generality in that they rely on nearest neighbour initialization from a training set. Besides, all of the above techniques used hand-crafted features.

In recent years, methods based on deep neural networks have become successful. Eigen et al. [5] proposed a robust depth estimation method using multi-scale Convolutional Neural Networks (CNN), and later extended it to a network structure that can also estimate the semantic labels and the surface orientation of a scene [4]. Thanks to the learning capability of multi-scale CNN and the availability of large-scale datasets, their latter work showed prominent performance. There are several works that combine CNN with CRF based regularization [20, 21, 33]. Liu et al. [20, 21] tackled the problem with Deep Convolutional Neural Fields which jointly learn CNN and continuous CRF. Wang et al. [33] used two separate CNN to obtain regional and global potentials of a scene and fed them into hierarchical CRF to jointly estimate the depth and the semantic labels. Roy and Todorovic [23] showed that random forests can also be used as a regularizer. While the majority of existing approaches trained the estimator with supervised learning using metric depth, there are some works that used relative depth [1, 37] or semi/unsupervised learning [10, 17].

In the literature, the method closest to our approach was proposed by Jafari et al. [13]. They first performed depth prediction and semantic segmentation using existing methods [4, 29] and then merged them through a Joint Refinement Network (JRN). Compared to [13], our architecture differs in two aspects. First, in addition to the mid-level object semantics, we consider the room layout of a scene as a global semantic structure for more consistent depth estimation. Second, we propose simple yet effective discriminator networks, which discern the reliability of the estimated structures, to further improve the performance.

3 Modular Network Architecture

This section presents the details of the proposed indoor scene depth estimation architecture. Figure 1 shows an overview of the proposed method. Taking a single RGB image as an input, we first estimate the global (i.e. room layout) and the mid-level (i.e. objects in a room) semantic structures of a scene using two separate semantic segmentation networks. The layout labels are estimated by the layout network N_L , and the object labels are estimated by the object network N_O . To treat the different number of classes in the same domain, we convert the estimated labels to 3-channel label images using a predefined colour palette. Then, the layout label image and the object label image are respectively fed into

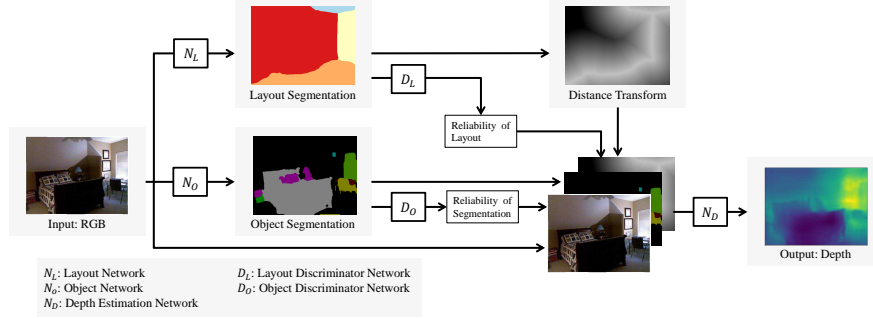


Fig. 1: Proposed depth estimation architecture. First, an input RGB image is fed into two separate semantic segmentation networks N_L and N_O to estimate room layout and object class labels, respectively. Then, discriminator networks D_L and D_O discern the reliability of the estimated layout and object labels. Lastly, the depth estimation network N_D takes the input image, two label images, and two reliability scores as inputs to infer the final depth values.

the discriminator networks D_O and D_L which estimate the reliability scores of these images; i.e. how *real* the estimated label images are. Before feeding the layout label image into the depth estimation network, we apply edge extraction and distance transform [7] to it. Lastly, the depth estimation network N_D takes the input image, two label images, and two reliability scores as inputs to infer the final depth values. Each label image does not directly flow into N_D , but is weighted by the respective reliability score. This weighting process reduces the adverse effect on the depth estimation score when the estimation quality of semantic segmentation is insufficient.

3.1 Semantic Segmentation

In the proposed method, two types of semantic segmentation are performed: (1) room layout segmentation to estimate the positions of walls, floor, and ceiling of a room and (2) object segmentation to recognize the items in a room such as furniture (hereinafter referred to as layout estimation and object recognition, respectively). We utilize Pyramid Scene Parsing Network (PSPNet) [35] for both segmentations.

In the layout estimation, we train the layout network N_L with five room layout classes: Ceiling, Floor, Right Wall, Left Wall, and Front Wall. The trained network infers dense labelling of the layout classes as depicted in Figure 1.

Object recognition gives us a more detailed, mid-level semantic structure of a scene. 11 object classes including Bed, Chair, Table, etc., are used to train the object network N_O . The estimated mid-level cues support the depth estimation network to make object depths consistent. Figure 1 shows the estimated object labels.

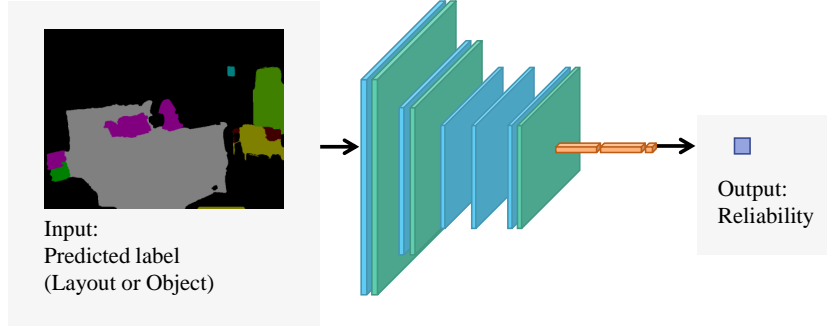


Fig. 2: Structure of the discriminator network. It takes the estimated label image and provides the reliability of the estimation in the interval $[0, 1]$.

To treat different numbers of classes in the same domain, we convert the estimated labels to 3-channel label images using a predefined colour palette.

3.2 Reliability Estimation

Since most CNN architectures assume that all of the input information is equally reliable, the input signals are not weighted. Rather, their reliability, or the amount of influence, is *implicitly* learned in the network. However, in the proposed modular architecture, where the layout estimation and the object recognition results are received as inputs, their quality may affect the final depth estimation result. Therefore, instead of implicit learning, we perform *explicit* weighting to reduce the influence of erroneous results.

We propose a reliability estimation network, which takes the estimated label image and provides the reliability of the estimation. The proposed reliability estimation network is inspired by the discriminator of Generative Adversarial Networks (GAN) [11] which discerns a given instance as being fake or real. Thus, we refer to this network as *discriminator network*. Figure 2 shows the network structure. We built it upon the AlexNet [16] with some modifications. We reduce the dimensions of the first two fully connected layers to 2,048 and set the output dimension to 1. The output reliability is activated by a sigmoid function.

The discriminator network is trained to output a value 1 for the ground truth label image and 0 for the estimated label image. We denote a training example as $\{l, \hat{l}\}$, where l denotes the estimated label image and \hat{l} denotes the corresponding ground truth label image. We define the loss function L_{dis} for the discriminator network as follows:

$$L_{dis} = -\frac{1}{m} \sum_{i=1}^m \left\{ \log(D(l_i)) + \log(1 - D(\hat{l}_i)) \right\} \quad (1)$$

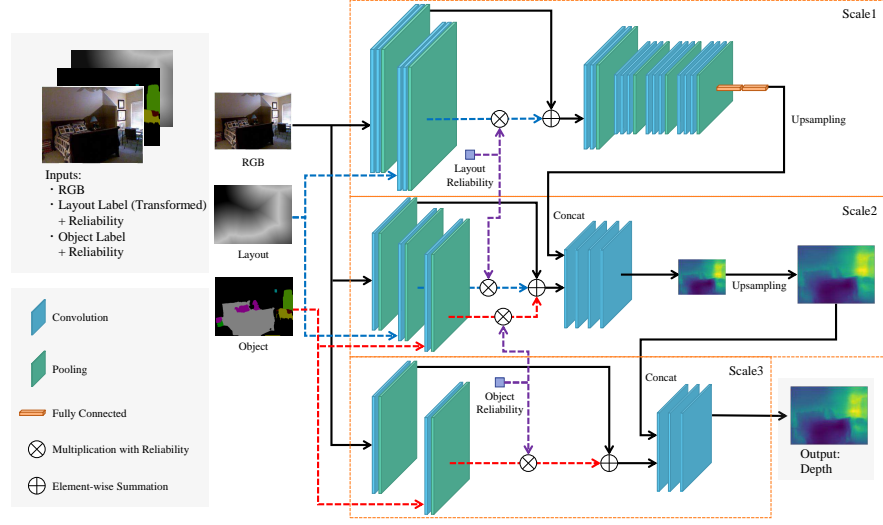


Fig. 3: Structure of the depth estimation network. Each scale takes different semantic information as the inputs. The layout label image is fed into Scale 1 and 2, and the object recognition label image is given to Scale 2 and 3. The layout label image produces the global semantic structure of a scene, while the object recognition label image gives the mid-level semantic structure.

where m is the mini-batch size, i is the index of each label image in the mini-batch, and $D(\cdot)$ is the reliability in the interval $[0, 1]$ estimated by the discriminator network. Note that the two discriminator networks D_L and D_O in the proposed architecture are individually trained with the results of the layout estimation and the object recognition, respectively.

3.3 Extension of Depth Estimation Network

Taking the original image, two label images, and their reliabilities as inputs, our depth estimation network N_D infers the detailed depth of a cluttered indoor scene. Figure 3 shows a conceptual diagram of the network structure. Our network is based on the multi-scale CNN [4] and is extended to consider semantic structures. Through preliminary experiments, we found that preprocessing the layout estimation label image before feeding it to the depth network yields better performance. Specifically, we apply edge extraction and distance transform [7] to the label image. For ease of notation, we refer to the transformed layout estimation label image and the object recognition label image as global semantics and mid-level semantics, respectively.

Scale 1. The first scale provides a coarse global set of features over the entire image region by processing low-level image features with convolutional layers followed by two fully connected layers which introduce global relations. To enhance

the global consistency of the depth prediction, we feed both the global semantics and the RGB image to this scale. The two input signals are separately mapped to feature spaces by the dedicated input convolutional layers. Then, the global semantics is weighted by multiplication with the estimated reliability score. The two feature maps are fused by element-wise summation and are processed by 11 convolutional layers and two fully connected layers. The output feature vector of the last layer is reshaped to $1/16$ of the spatial size of the inputs, then bilinearly upsampled to $1/4$ the scale.

Scale 2. As the scale increases, the network captures a more detailed scene structure. In the second scale, the network produces a depth prediction at mid-level resolution by combining feature maps from a narrower view of the image along with the global features provided by Scale 1. This scale takes as inputs both the global and the mid-level semantics in addition to the RGB image, and acts as the bridge between the global and the mid-level structures. The last convolutional layer outputs a coarse depth prediction of spatial resolution 74×55 . The predicted depth is upsampled to 148×110 and fed to the later stage of Scale 3.

Scale 3. The third scale further refines the prediction to a higher resolution. To recover the detailed structure of a scene, we feed both the mid-level semantics and the RGB image to this final scale. After merging the two input signals, we concatenate the output from Scale 2 with the feature maps to incorporate multi-scale information. The final output is a depth prediction of size 148×110 .

We train the network using the loss function motivated by [4]. We denote training examples as $\{Y, \hat{Y}\}$, where Y denotes the predicted depth map and \hat{Y} denotes the ground truth depth map. Putting $d = Y - \hat{Y}$ to be their difference, the loss function L_{depth} is defined as:

$$L_{depth}(\hat{D}, D^*) = \frac{1}{n} \sum_i d_i^2 - \frac{1}{2n^2} \left(\sum_i d_i \right)^2 + \frac{1}{n} \sum_i [(\nabla_x d_i)^2 + (\nabla_y d_i)^2] \quad (2)$$

where n represents the total number of valid pixels in the image (we mask out the pixels where the ground truth is missing), i represents the pixel index, and $\nabla_x d_i$ and $\nabla_y d_i$ are the image gradients of the difference in the horizontal and vertical directions. We convolve a simple 1×3 filter to calculate $\nabla_x d_i$ and use its transposed version to calculate $\nabla_y d_i$.

4 Experiments

We evaluate our depth estimation architecture on the standard NYU Depth v2 indoor scene depth estimation dataset [30] which contains 654 test images. We compare our architecture with the published results of recent methods [4, 5, 13, 20, 23, 33]. For quantitative evaluation, we report the following commonly used metrics:

- Absolute relative difference (abs rel): $\frac{1}{n} \sum_i \frac{|y_i - \hat{y}_i|}{y_i}$
- Squared relative difference (sqr rel): $\frac{1}{n} \sum_i \frac{\|y_i - \hat{y}_i\|^2}{y_i}$
- RMSE (rms(linear)): $\sqrt{\frac{1}{n} \sum_i \|y_i - \hat{y}_i\|^2}$
- RMSE in log space (rms(log)): $\sqrt{\frac{1}{n} \sum_i \|\log(y_i) - \log(\hat{y}_i)\|^2}$
- Average log₁₀ error (log10): $\frac{1}{n} \sum_i |\log_{10}(y_i) - \log_{10}(\hat{y}_i)|$
- Threshold: % of y_i s.t. $\max\left(\frac{y_i}{\hat{y}_i}, \frac{\hat{y}_i}{y_i}\right) = \delta < thr$, where $thr \in \{1.25, 1.25^2, 1.25^3\}$

where y_i is the predicted depth value of a pixel i , \hat{y}_i is the ground truth depth, and n is the total number of pixels. The next subsection describes training procedures and datasets used to train the network modules.

4.1 Implementation Details

Layout Estimation. We train the layout estimation network N_L with the LSUN layout estimation dataset [34], which contains 4,000 indoor images. Following the procedure of [2], we assign dense semantic layout labels to the images and train the network as a standard semantic segmentation task. Since the dataset contains images of various sizes, we apply bicubic interpolation to resize the images to 321×321 pixels. We utilize the PSPNet model [35] and initialize its parameters using pre-trained weights (trained with the Pascal VOC2012 [6]) which is publicly available¹. We set the base learning rate for the SGD solver to 0.0001 and apply polynomial decay of power 0.9 to this rate at each iteration during the whole training. Momentum and weight decay rate are set to 0.9 and 0.0001, respectively. Due to physical memory limitations on our graphics card, we set the mini-batch size to 2.

Object Recognition. We use almost the same set-up as for the layout estimation, except for training dataset. We train the object recognition network N_O with 5,285 training images from SUN RGB-D semantic segmentation dataset [32]. We apply no resizing to this dataset and feed the original 640×480 images to the network. We modify the standard 37 object categories by mapping to 13 categories [4] and removing duplicated layout classes (i.e. Wall, Floor, Ceiling). In addition, we add a ‘background’ class and assign the above removed categories to this class. This results in 11 classes. To improve the segmentation quality, we use a fully-connected CRF [15] for post-processing.

Reliability Estimation. We train two discriminator networks D_L and D_O using the LSUN dataset [34] and the SUN RGB-D dataset [32], respectively. We use the same training procedure for D_L and D_O . For ease of notation, we omit the subscripts in the following explanation. First, we acquire estimation results from a training set of the dataset using the trained semantic segmentation network N . The estimated labels and the corresponding ground truth labels are coloured using a predefined colour palette. Then, we assign a label 0 for the estimated label image and 1 for the ground truth label image. Finally, we train

¹ <https://github.com/hszhao/PSPNet>

Table 1: Quantitative comparison against previous approaches on the NYU Depth v2 dataset.

Method	Error (lower is better)					Accuracy (higher is better)		
	abs rel	sqr rel	rms(linear)	rms(log)	log10	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen et al. [5]	0.215	0.212	0.907	0.285	-	0.611	0.887	0.971
Joint HCRF [33]	0.220	0.210	0.745	0.262	0.094	0.605	0.890	0.970
Liu et al. [20]	0.230	-	0.824	-	0.095	0.614	0.883	0.971
Eigen and Fergus [4]	0.158	0.121	0.641	0.214	-	0.769	0.950	0.988
Roy and Todorovic [23]	0.187	-	0.744	-	0.078	-	-	-
Jafari et al. [13]	0.157	0.123	0.673	0.216	0.068	0.762	0.948	0.988
Ours	0.151	0.107	0.601	0.203	0.061	0.801	0.969	0.992
Ours w/o Disc.	0.163	0.127	0.631	0.207	0.064	0.794	0.963	0.991
Objects + Disc.	0.155	0.119	0.619	0.202	0.065	0.780	0.961	0.992
Layout + Disc.	0.151	0.112	0.643	0.202	0.071	0.778	0.959	0.992
Distance trans. + Disc.	0.161	0.121	0.626	0.206	0.063	0.774	0.959	0.991

the discriminator D using the loss function defined in Equation 1. We use Adam optimizer [14] and set the learning rate to 10^{-10} .

Depth Estimation. Following [4, 5], we train our depth estimation network N_D using the raw distribution of the NYU Depth v2 dataset [30] which contains many additional images. We extract 16K synchronized RGB-depth image pairs using the toolbox provided by the authors². We downsample the RGB images from 640×480 to 320×240 pixels. The ground truth depth maps are converted into log space and resized to the network output size 148×110 . We train the network using the SGD solver with mini-batches of size 8. Learning rate and Momentum are set to 10^{-6} and 0.9, respectively. Note that, our training is done by end-to-end learning instead of the incremental learning in [4].

4.2 Results on the NYU Depth v2

Table 1 shows the quantitative comparison of the proposed architecture against previous approaches on the NYU Depth v2 dataset [30]. The proposed architecture shows the best performance in most metrics. Comparing to the baseline [4], which our architecture is built upon, we achieve consistent improvements in all metrics. To evaluate the proposed architecture in detail, we conduct experiments with several settings. As shown in Table 1, using the object recognition (Objects + Disc.) has positive impacts on all metrics. Interestingly, in the individual case, using the layout estimation without distance transform (Layout + Disc.) performs better than the distance transformed layout (Distance trans. + Disc.). However, we found that the distance transformed layout provides better results when it is integrated into the whole architecture. More importantly, one can see the discriminator networks play an important role in our architecture (Ours w/o Disc.). These results validate the effectiveness of our modular architecture.

Figure 4 presents the qualitative comparison between the proposed architecture and the prediction of [4]. For detailed comparison, we also visualize the

² https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html

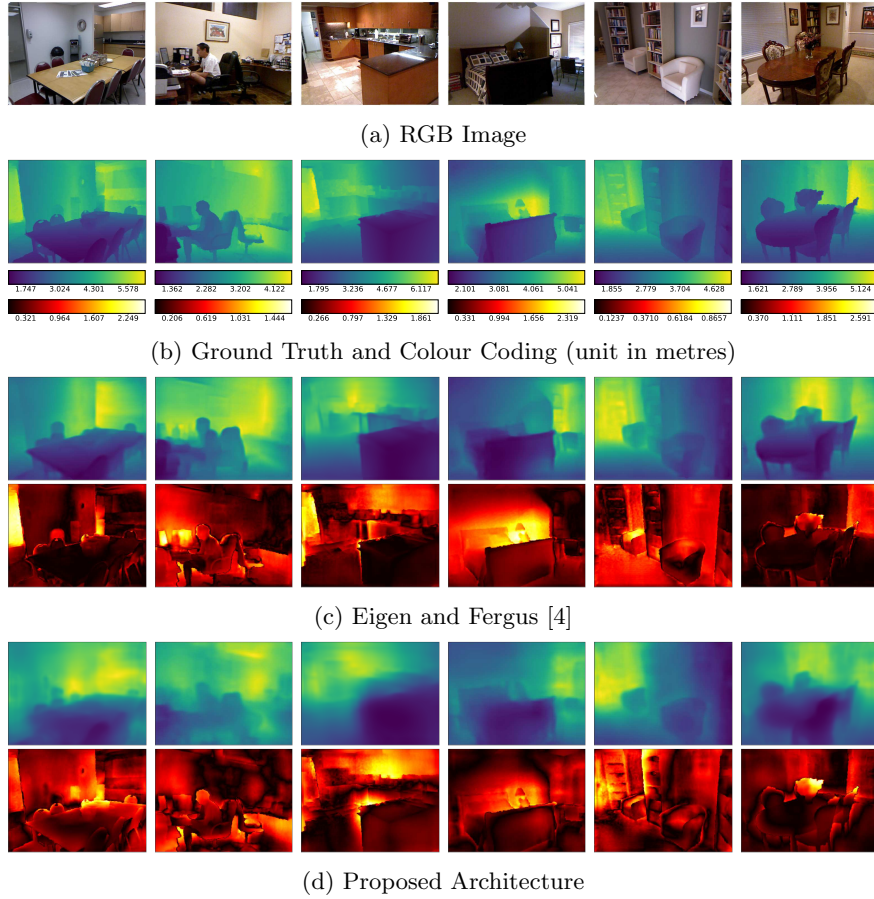


Fig. 4: Depth estimation results for the NYU Depth v2 dataset. We show the depth maps in upper rows and the corresponding error maps in lower rows.

errors in the depth maps. In addition to the quantitative performance improvements, we found that the proposed architecture is more robust to the appearance changes inside objects. In the second scene from the left, the sticky notes pasted on a computer change its appearance and [4] produces large estimation error. In contrast, our architecture consistently estimates the depth inside the object. A similar effect appears in the centre of the fourth scene from the left, where a shadow changes the appearance of a wall.

One drawback of our architecture is “blur effect” in object boundaries. The visualized results show that feeding the object recognition label image into the depth estimation contributes to the accuracy improvement. Nevertheless, the object boundaries become unclear due to imperfect segmentation results. The

layout estimation has similar effects. Although it improves the global consistency of the prediction, it smooths out the local object boundaries.

5 Conclusions

We have proposed a novel deep modular network architecture for monocular depth estimation of indoor scenes. Two auxiliary semantic segmentation networks give us the global (i.e. room layout) and the mid-level (i.e. objects in a room) semantic structure to enhance depth prediction. Inspired by GAN, we have introduced discriminator networks, which discern the reliability of the estimated semantic structures. Each semantic structure is weighted by the respective reliability score, and this process reduces the adverse effect on the depth estimation when the estimation quality of semantic segmentation is insufficient. We evaluated the proposed architecture on the NYUD Depth v2 benchmark dataset and showed significant performance improvements over previous approaches.

Acknowledgments

This work was partially supported by Aoyama Gakuin University-Supported Program “Early Eagle Program”. The authors are grateful to Prof. M.J. Dürst from Aoyama Gakuin University for his careful proofreading and kind advices on the manuscript.

References

1. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: NIPS. pp. 730–738 (2016)
2. Dasgupta, S., Fang, K., Chen, K., Savarese, S.: Delay: Robust spatial layout estimation for cluttered indoor scenes. In: CVPR. pp. 616–624 (2016)
3. Dellaert, F., Seitz, S.M., Thorpe, C.E., Thrun, S.: Structure from motion without correspondence. In: CVPR. pp. 557–564 (2000)
4. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV. pp. 2650–2658 (2015)
5. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NIPS. pp. 2366–2374 (2014)
6. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>
7. Felzenszwalb, P., Huttenlocher, D.: Distance transforms of sampled functions. Tech. rep., Cornell University (2004)
8. Foix, S., Alenya, G., Torras, C.: Lock-in time-of-flight (ToF) cameras: A survey. IEEE Sensors Journal **11**(9), 1917–1926 (2011)
9. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
10. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR. pp. 270–279 (2017)

11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)
12. Hoiem, D., Efros, A.A., Hebert, M.: Automatic photo pop-up. *ACM Transactions on Graphics* **24**(3), 577–584 (2005)
13. Jafari, O.H., Groth, O., Kirillov, A., Yang, M.Y., Rother, C.: Analyzing modular CNN architectures for joint depth prediction and semantic segmentation. In: ICRA. pp. 4620–4627 (2017)
14. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
15. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS. pp. 109–117 (2011)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1097–1105 (2012)
17. Kuznetsov, Y., Stücker, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. In: CVPR. pp. 6647–6655 (2017)
18. Ladicky, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: CVPR. pp. 89–96 (2014)
19. Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: CVPR. pp. 1253–1260 (2010)
20. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: CVPR. pp. 5162–5170 (2015)
21. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(10), 2024–2039 (2016)
22. Liu, M., Salzmann, M., He, X.: Discrete-continuous depth estimation from a single image. In: CVPR. pp. 716–723 (2014)
23. Roy, A., Todorovic, S.: Monocular depth estimation using neural regression forest. In: CVPR. pp. 5506–5514 (2016)
24. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* **115**(3), 211–252 (2015)
25. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: NIPS. pp. 1161–1168 (2005)
26. Saxena, A., Sun, M., Ng, A.Y.: Make3D: Learning 3D scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(5), 824–840 (2009)
27. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: CVPR. pp. 195–202 (2003)
28. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: CVPR. pp. 519–528 (2006)
29. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(4), 640–651 (2017)
30. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: ECCV. pp. 746–760 (2012)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
32. Song, S., Lichtenberg, S.P., Xiao, J.: SUN RGB-D: A RGB-D scene understanding benchmark suite. In: CVPR (2015)

- 33. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L.: Towards unified depth and semantic prediction from a single image. In: CVPR. pp. 2800–2809 (2015)
- 34. Yu, F., Zhang, Y., Song, S., Seff, A., Xiao, J.: LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
- 35. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 2881–2890 (2017)
- 36. Zhuo, W., Salzmann, M., He, X., Liu, M.: Indoor scene structure analysis for single image depth estimation. In: CVPR. pp. 614–622 (2015)
- 37. Zoran, D., Isola, P., Krishnan, D., Freeman, W.T.: Learning ordinal relationships for mid-level vision. In: ICCV. pp. 388–396 (2015)