Multi-task Driven Network for Indoor Layering Prediction with Human-in-loop

Yuhuan Yang
Shanghai JiaoTong University
yangyuhuan@sjtu.edu.cn

Qingyao Xu
Shanghai JiaoTong University
xuqingyao@sjtu.edu.cn

Han Lu
Shanghai JiaoTong University
sjtu_luhan@sjtu.edu.cn

ABSTRACT

Scene understanding has always been a topic for researches. Models like semantic parsing and spatial position have been made to extract information from images. However, few models can present a general but intuitive impression on the input image. To improve this situation, this paper proposes a novel architecture: human-in-loop Multi-task Driven Network(MD-Net), aiming to give the layering of the spatial structure of images. MD-Net consists of an encoder for the input and two decoders for two auxiliary tasks, and finally a convolution layer to generate the output prediction. This model can run with humans selecting good results from new unseen datasets in each loop to enrich the training dataset and improve performance. MD-Net is proved to be effective and can generate new datasets efficiently. The code, model, and dataset are publicly available.

CCS Concepts

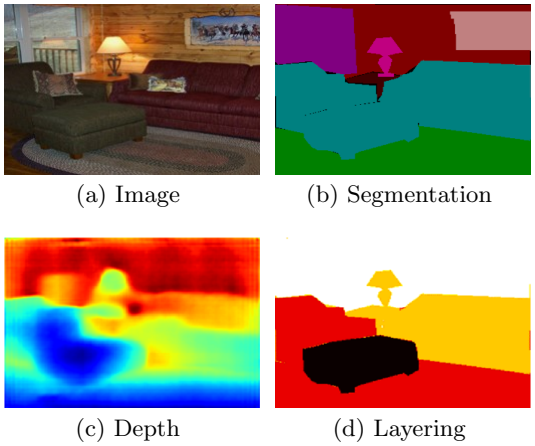
• **Computiong methodologies→ Artificial intelligence →Computer vision**

Keywords

Indoor Layering, human-in-loop, Multi-task Driven Network

1. INTRODUCTION

Indoor scene understanding is an important topic and has wide commercial applications like the domestic robot. By observing the process of human observation,we find that at the first glimpse of an image, people tend to notice objects in the order from near to far and form a sense of spatial hierarchy of the image.People may classify objects that are relatively near to them as the first layer, which are easy to interact with. And objects a bit far as the second layer, followed by the possible third and fourth layers. We define this spatial hierarchy as layering. Obviously, layering plays a significant role in the indoor scene understanding task.



(a) Image (b) Segmentation (c) Depth (d) Layering

Figure 1: Segmentation, depth and layering for an image. We aim at generating layering(d) for an input image. The black, red, orange and white colors in (d) represent the nearest, medium, far, farthest layer in the image respectively.

Existing models such as depth model and segmentation can be transferred to handle the image layering prediction task. However, since they are not specially designed for this task and thus have serious defects in this situation. Depth model like [8] may be unable to capture objects' semantic parsing and thus lead to dividing some furniture half-baked. And segmentation model like [5] is incapable to obtain objects' spatial information. Apart from the deficiency of proper methods, this task also suffers from the lack of dataset.

Our target is generating a hierarchy structure of an image by setting the intact objects in the corresponding spatial layer and creating a suitable indoor dataset for this task. We aim our target on the indoor dataset for two reasons. One is that the indoor scenes usually have a similar layerings with a limited depth span, and are thus more familiar to people. So it is relatively easy to divide images into several layers without much controversy. The other reason is that our method may be the base of tasks like indoor furniture arranging, indoor robot navigation and other indoor scene understanding.

Therefore, We proposed the human-in-loop Multi-task Driven Network(MD-Net), an end-to-end network, for the special task layering prediction. We obtain data from SUN [13] and generate ground truth for our task with specific rules with the help of segmentation and depth information. The human-in-loop mechanism is proposed to improve the data.

Inspired by [7], our MD-Net uses ResNet [6] as the encoder and two decodes for depth and segmentation respectively. The output from two decoders is used in an auxiliary way. Then we design a convolution layer to combine the generated depth and segmentation for the final output. Besides, we propose a human-in-loop approach to generate new annotations for the raw images, which enriches the training set.

To summarize, the main contributions of our work are four-fold:

1. We propose the task of indoor layering prediction to understand the spatial hierarchy.

2. We propose the MD-Net which can extract and explicit both depth and segmentation information to solve the task efficiently.

3. We propose a human-in-loop method to improve the performance of our model.

4. We propose an efficient method to generate a new scalable dataset for some tasks with minor cost in human resources.

2. RELATED WORK

In this section, we briefly review some previous modules that are highly related to our work.

Depth estimation The works on monocular depth estimation can be mainly grouped into four categories. The first one focus on the network. [4][3] mainly rely on the methods of the deep learning and network architecture to get results. Besides, to enhance the resolution of the output, [8] presents a new method which can learn feature map upsampling efficiently. They also introduce the reverse Huber loss for optimization. The second one focus on the property of the depth information itself like [11] [12]. The third one is based on the CRF method like [9][15]. CRF can make the blurred images generated by CNN not fuzzy through conditional probability modeling. The fourth one is based on relative depth like [17] [2].

Sematic segmentation Pyramid Scene Parsing Net [16] has the ability to dig global information through context aggregation of different regions by using its pyramid pooling pattern. It achieves excellent performance on multiple semantic segmentation benchmarks. DeepLabv3+ [1] is an encoder-decoder architecture, where the encoder architecture uses DeepLabv3 and the decoder uses a effective module for recovering target boundary details. It also makes good use of Xception and deep separation convolution on the model to further improve the speed and performance.

Multi-task Prediction [14] proposes Multi-Tasks Guided Prediction-and-Distillation Network (PAD-Net) for synchronous depth estimation and scene parsing. The network generates intermediate auxiliary tasks to provide rich multi-modal data for target tasks. [10] combined object detection and segmentation in natural scene into a single framework. Inspired by this, we proposed our MD-Net.

However, the models listed above are not suitable enough for our task. Therefore, we propose our human-in-loop MD-Net.

3. DATASET

We construct our dataset from Sun-dataset[13].

3.1 Ground Truth Annotation

Our task is generating indoor layering prediction. Since there's no dataset about this task, the ground truth for our task is generated by ourselves using a specific rule with the help of depth and segmentation information.

Based on the idea that both objects' depth and completeness contribute to human's sense of spatial hierarchy, we design a rule that first calculates the average depth for each instance and then layer all instances into four layers according to their average depth in ascending order. Our rule is designed in the following format Alg.1.

Algorithm 1 The rule to generate level

Require:
depth:pixel-level depth information
seg:pixel-level segmentation information
classes: classes list in segmentation
threshold: threshold for depth to classify level
function RULE($depth, seg, classes, threshold$)
$area \leftarrow EmptyDic()$
$layer \leftarrow EmptyDic()$
for i **in** classes **do**
$area[i] \leftarrow$ average depth for class i
end for
$area.value\_sort()$
for j **in** classes **do**
$layer[j] \leftarrow \arg\max_i area[j] > threshold[i]$
end for
end function
return $layer$

Both segmentation and depth model may introduce error to the generated images. The segmentation part may overlap, and the depth information may be distracted by a light source. Our rule is sensitive to this information, and consequently may not be precise enough to describe human sense. To overcome the shortcomings listed above, Ten volunteers with different ages and backgrounds are invited to sift the imprecise data.

3.2 Data Augmentation

To expand the size of our original training dataset, we apply image data augmentation with horizontal flip, salt and pepper noise,Gaussian noise, brightness adjustment. By doing augmentation, variations of the training set images are more likely to be seen by the model. For example, the horizontal flip operation may represent the change of a viewer's horizontal location.

Apart from this, we also enrich the dataset with the help of human-in-loop. We generated layering prediction for new data, and pick those satisfying data manually. Chosen ones are then fed into our network for another round of training. Both generating and picking are time-saving tasks.

3.3 Advantages of Generating Method

Our approach is relatively faster to generate data for a brand new task. In our experiment, we invite 8 volunteers to annotate pixel-level layering prediction for our dataset. Another 2 volunteers are arranged to pick satisfying data

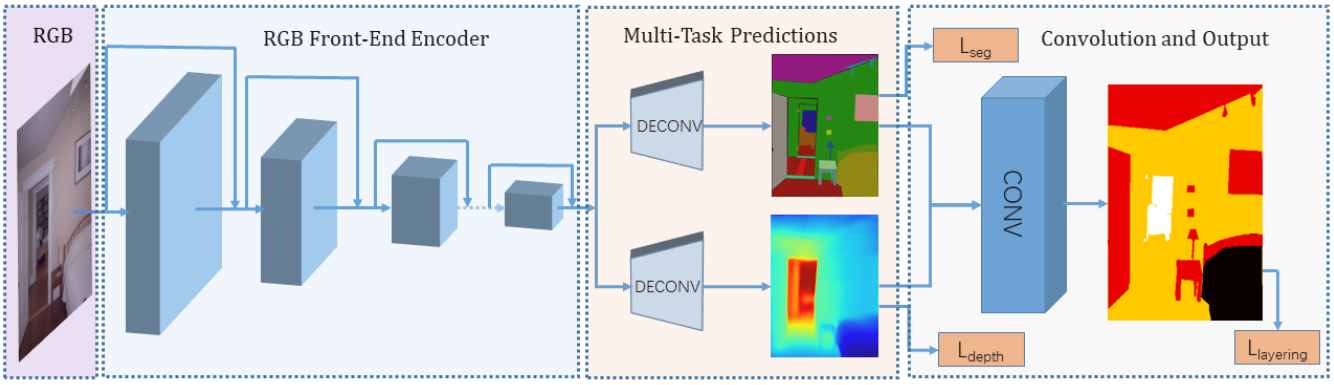


Figure 2: Illustration of the proposed MD-Net for layering prediction. The front end encoder is Resnet. The symbols of L_{seg}, L_{depth} and $L_{layering}$ denote different parts of losses for different tasks. 'DECONV' parts denote the deconvolutional operation for upsampling and generate task-specific output. The cube 'conv' represents a convolution layer for the generation of the final task.

from generated ones. It takes nearly 8 hours for the first group to annotate 100 images. However, the second group spends just 2 hours to pick more than 1000 images successfully.

Therefore, with the human-in-loop mechanism, our network can effectively learn about human preferences while using only a small amount of human resources.

4. METHOD

In this section, we describe our human-in-loop Multi-task Driven Network(MD-Net) for indoor layering prediction. We first present an overview of the proposed MD-Net. And then, we introduce the details of each part. Finally, we illustrate the human-in-loop mechanism for the network.

4.1 Overview

The following figure 2 depicts the framework of the proposed MD-Net. MD-Net consists of three main components. First, a front-end convolutional encoder is used to produce deep features. Second, a multi-task prediction decoder generates intermediate predictions. Third, a convolution layer is used for incorporating useful multi-modal information from the intermediate predictions and generating the final result. The whole network is firstly trained with labeled data, i.e. RGB images with pixel-level segmentation and depth information. Secondly, use the trained network to generate pseudo labels for raw data, which are feed back to the model to improve the performance.

4.2 Encoder-Decoder Structure

Denote our labeled dataset as $\mathcal{X}, \mathcal{Y}$, where $\mathcal{X} = \{x_1, ...x_n\}$ is RGB pictures and $\mathcal{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_3\}$ is the ground truth for segmentation, depth and layering respectively. We implement commonly used ResNet[6] as the front-end encoder and input $\mathcal{X}$ to extract features from the last convolutional layer. The decoder is composed of several transpose-convolution layers and its structure is the same as ResNet[6]. We use two decoders for the task of predicting depth and segmentation separately. The output from two decoders are denoted as $\mathcal{Y}'_1$ and $\mathcal{Y}'_2$ respectively.

4.3 Prediction and Loss

To generate prediction from the output of multiple tasks, we use a pixel-level attention mechanism. The output of this network is a pixel-level prediction of layers. There are three parts of loss in the whole network:

$$\mathcal{L}_{seg}(\mathcal{Y}'_1, \mathcal{Y}_1) = -\log\left(\frac{exp(\mathcal{Y}'_1[\mathcal{Y}'_1]}{\sum_j exp(\mathcal{Y}'_1[j])}\right) \quad (1)$$

$$\mathcal{L}_{depth}(\mathcal{Y}'_2, \mathcal{Y}_2) = \frac{1}{n}\sqrt{\mathcal{Y}'_2 - \mathcal{Y}_2} \quad (2)$$

$$\mathcal{L}_{layerering} = -\log\left(\frac{exp(Conv(\mathcal{Y}'_1 + \mathcal{Y}'_2)[\mathcal{Y}_3]}{\sum_j exp(Conv(\mathcal{Y}'_1 + \mathcal{Y}'_2)[j])}\right) \quad (3)$$

In the formula above, L_{seg} and $L_{layering}$ are the cross-entropy loss for segmentation and layering prediction tasks, and L_{depth} is the MSE loss for depth prediction task.

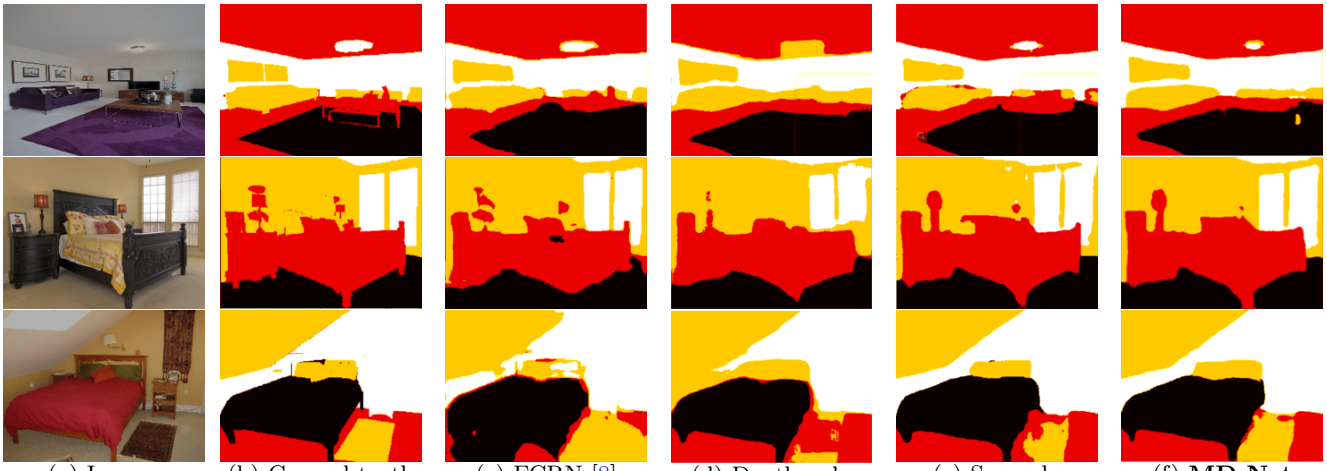
4.4 Training Process

We use [8] to generate the depth for images. After that, we made some rules to generate the layering prediction of the given image using the depth, semantic segmentation and other information. The whole training process is divided into two stages: supervised learning and unsupervised enhancement. In the supervised stage, the network is trained with labeled dataset and minimize the total loss $\mathcal{L}_{layering} + \lambda_1\mathcal{L}_{seg} + \lambda_2\mathcal{L}_{depth}$.

After the model is trained in the supervised part, we use our network to generate layering prediction for new data. Then we hire several volunteers to evaluate the generated layering for new data and pick out unwanted ones, ones that classify some objects in the wrong layer. The sifted dataset is then fed into the network for a new round of training. Different from the supervised stage, the new round of training only minimizes $\mathcal{L}_{layering}$. The whole training process is described in Alg.2.

5. EXPERIMENTS

In this section, we first give the implementation details of our MD-Net and show the performance on our evaluate dataset. Then we conduct some comparative experiments,



(a) Image (b) Ground truth (c) FCRN [8] (d) Depth only (e) Seg only (f) **MD-Net**

Figure 3: Qualitative examples of indoor layering prediction on our dataset.

Algorithm 2 The whole training process

model← MD-Net
for i=0:train_epoch **do**
Train model on labeled dataset for one iteration
end for
for i=0:loop_epoch **do**
output ← model prediction on unlabeled dataset.
Sift output by human
Use the output as ground truth and train model for one iteration
end for

showing that our method is the most adaptive one to handle the indoor layering prediction problem.

5.1 Implementation Details

We propose an end-to-end MD-Net for our task. We use the pre-trained ResNet-18 [6] as our encoder backbone. Both segmentation decoder and depth decoder consist of four decoder blocks correspond to the four encoder blocks. We then concatenate two outputs and use a convolution layer to get the final result.

After data augmentation, we get 12708 images in total. For all experiments, we train for 20 epochs with a batch-size of 8. The training images are resized to 640×465.

We adopt MSE loss $\mathcal{L}depth$ for depth prediction and cross-entropy loss $\mathcal{L}_{seg}$ for the segmentation phase. For the final result, we utilize cross-entropy loss $\mathcal{L}_{layering}$ to measure the difference between our result and the ground truth. The total loss is

$$\mathcal{L} = \lambda_1 \mathcal{L}_{seg} + \lambda_2 \mathcal{L}_{depth} + \mathcal{L}_{layering}. \quad (4)$$

We conduct lots of experiments to find the fittest loss weight as Table1 shows. We finally determine $\lambda_1 = 1, \lambda_2 = 0.5$ in our implementation.

We use Adam optimizer for the back-propagating. The initial learning rate used in all trainings is set to 0.001 initially.

5.2 human-in-loop Enhancement

Table 1: The mIOU of different λ_1 and λ_2

λ_1 \ λ_2	0.25	0.5	1	2
0.25	0.879	0.889	0.895	0.883
0.5	0.885	0.876	0.877	0.869
1	0.908	**0.916**	0.909	0.898
2	0.902	0.902	0.899	0.904

We propose the human-in-loop method to help with our model. After the evaluation, we invite some volunteers to pick satisfying results. Then we combine the chosen images together with our initial training set to continue the training. We apply 3 loops of training, and uses raw data from SUN [13]. We get 532 more images. After doing data augmentation, this part of data is used for the second stage of training. The performance of the model improves and reaches the final score of 0.919.

5.3 Comparative Experiments

To emphasize our idea that both depth and semantic segmentation information contribute to the final result, we conduct some comparative experiments:

Table 2: Comparison between different methods

Method	mIOU	MSE
FCRN [8]	0.3943	1.0680
Depth_only	0.8787	0.1365
Seg_only	0.9121	0.1205
MD-Net	0.9166	0.1152
MD-Net-human-in-loop	**0.9190**	**0.1147**

1. We employ the FCRN model proposed in [8] to generate the layered image. We use the input size 304×228 to consistent with the pre-trained model.
2. We develop a Depth_only model. We discard the segmentation predict decoder and only employ a depth decoder to generate the final result.
3. Similar to 2, we use the Seg_only model that generates

the result with only segmentation decoder.

We show the performance of each methods in Table 2. Several examples are shown in figure 3.

In the table above, MSE is the mean square error between prediction and ground truth. From the first three lines, we can see that the segmentation model plays a more significant role in generating than depth estimation. Our network architecture is more efficient than FCRN [8], but achieves excellent performance on our layering prediction benchmarks. The human-in-loop method proves to be capable of enriching our dataset and it can improve the performance. And we believe that with more additional data, our model will perform better.

6. CONCLUSION

In this paper, we focused on the layering of images and proposed our MD-Net. Our network can generate layering prediction for RGB images effectively. Besides, we created a scalable dataset for this new task. We proposed a human-in-loop architecture to enrich the dataset much more efficiently than annotated by humans. As a future work, we would focus on applying our model to a variety of scenes and adapt the human-in-loop in a larger dataset for better results.

7. REFERENCES

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[2] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. In *Advances in neural information processing systems*, pages 730–738, 2016.

[3] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

[4] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[5] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019.

[8] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[9] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1119–1127, 2015.

[10] J. Vansteenberge, M. Mukunoki, and M. Minoh. Combined object detection and segmentation. *International Journal of Machine Learning and Computing*, 3(1):60, 2013.

[11] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2809, 2015.

[12] P. Wang, X. Shen, B. Russell, S. Cohen, B. Price, and A. L. Yuille. Surge: Surface regularized geometry estimation from a single image. In *Advances in Neural Information Processing Systems*, pages 172–180, 2016.

[13] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.

[14] D. Xu, W. Ouyang, X. Wang, and N. Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018.

[15] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5354–5362, 2017.

[16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[17] D. Zoran, P. Isola, D. Krishnan, and W. T. Freeman. Learning ordinal relationships for mid-level vision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 388–396, 2015.