Depth Map Prediction from a Single Image using a Multi-Scale Deep Network

Akshay Verma
Dept. of Electrical Engineering
IIT Bombay
200070005

Pulkit Adil
Dept. of Electrical Engineering
IIT Bombay
200070062

Sahil Garg
Dept. of Electrical Engineering
IIT Bombay
200070070

***Abstract*—Depth estimation is a crucial task in computer vision, owing to its numerous applications in several fields. Depth estimation using stereo images has been extensively studied, but estimating the depth from a single image still remains an important challenge. This is due to the complexity of mapping 2d images to 3d images, and the lack of ground truth maps available for real-world images. A combination of coarse and fine-scale networks can be used to predict depth maps with high accuracy. The coarse network predicts the depth of the scene globally, while the fine-scale network refines the prediction locally to incorporate finer-scale details. In this report, we present the results obtained using such a combination of networks on the NYUDepth dataset.**

***Index Terms*—Depth map, depth estimation, coarse networks, fine-scale networks, single-image, convolutional neural networks**

I. Introduction

Depth estimation is the task of predicting the distance of objects in a scene from a given viewpoint, typically in the form of a depth map. This is an important problem in computer vision as it has various applications such as 3D reconstruction, robotics, autonomous driving, portrait mode in camera, etc.

While there has been a lot of research done on depth estimation using stereo images (extracting 3d information from various 2d images of the same scene), accurately estimating depth from a single RGB image is still a difficult problem. This is because there are multiple possible 3D scenes that can be mapped to a 2D image, and also because there are limited ground truth depth maps available for real-world images. A combination of coarse and fine-scale networks can be used to predict depth maps with high accuracy. The coarse network predicts the depth of the scene globally, while the fine-scale network refines the prediction locally to incorporate finer-scale details. The approach uses both the original input and the coarse network output as additional first-layer image features, enabling the local network to edit the global prediction.

II. Problem Statement

The paper focuses on the tough challenge of estimating the depth map of a scene using only one image. The problem becomes difficult because of the complexity of mapping a 2D image to a 3D scene. Furthermore, most of the real-world images lack ground truth depth maps, which adds to the complexity of the problem.

III. Method

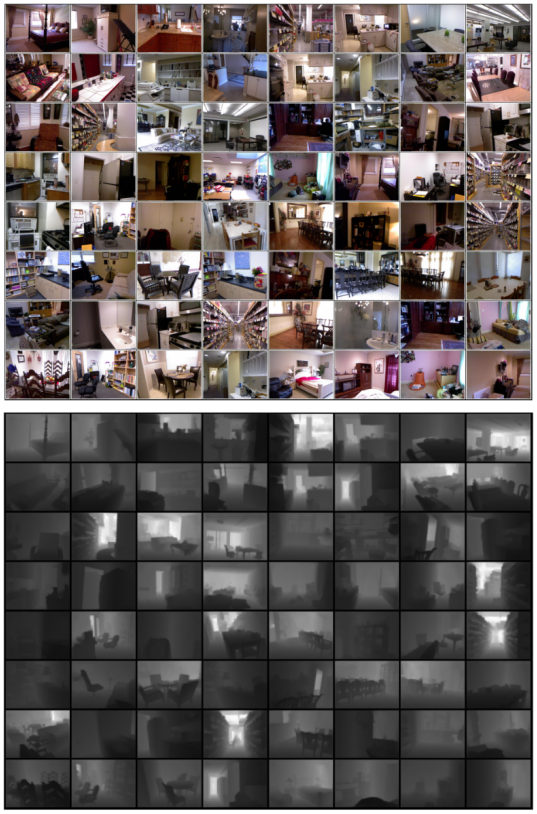

Fig. 1. Image dataset

The network comprises two stacks: a coarse-scale network that initially estimates the depth of the scene globally and a fine-scale network that further refines the estimation within local regions. Both stacks are applied to the original input, and the output of the coarse network is additionally used as first-layer image features for the fine network. This allows the fine network to enhance global prediction with more detailed information.

We are using the NYUDepth dataset and the images, as shown in Fig. 1, are used for training the model. The NYUDepth dataset is composed of 464 indoor scenes, which are captured as video sequences. The RGB samples are downsampled by half, from 640x480 to 320x240.

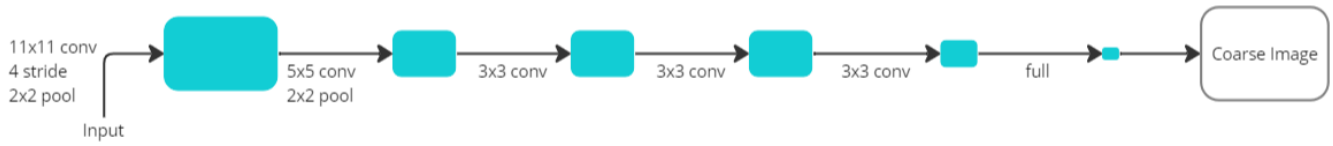

Fig. 2. Global Coarse-Scale Network

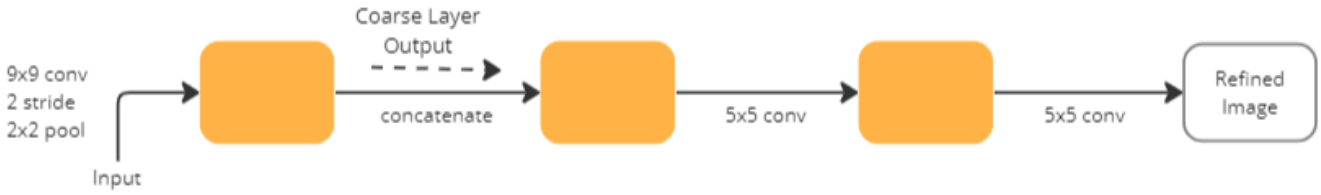

Fig. 3. Local Fine-Scale Network

		Coarse					Fine
Layer	Input	1	1,2,3,4	5	6	7	1,2,3,4
NYU Depth	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Ratio to input	1:1	1:8	1:16	1:32	-	1:4	1:4

TABLE I
SIZE OF OUTPUTS OF VARIOUS LAYERS

A. Global Coarse-Scale Network

The coarse-scale network predicts the overall depth map structure by analyzing the entire image and integrating a global understanding of the scene. The network uses fully connected upper layers and max-pooling operations in lower and middle layers to combine information from different parts of the image. This enables the network to effectively utilize cues such as vanishing points, object locations, and room alignment, which are crucial in a single-image scenario. A local view, as used in stereo matching, is insufficient for detecting such features.

As can be seen in Fig. 2, this network contains a total of 7 seven layers – five feature extraction layers of convolution and max-pooling, followed by two fully connected layers. The hidden layers use ReLU as the activation function while the coarse output layer employs linear activation.

Also, it can be noted that the output size of the network is allowed to be larger than the size of the top convolutional feature map. Templates are learned by the top full layer over this larger area, rather than limiting the output to the feature map size and relying on hardcoded upsampling. By allowing the network to learn its own upsampling based on the features, the limitations of a fixed upsampling method can be avoided. Although the templates may be blurry, they provide a better output than upsampling a small feature map (8×6 for NYUDepth).

B. Local Fine-Scale Network

Once the coarse depth map is predicted from a global perspective, a fine-scale network is employed to make local refinements. The purpose of this network is to adjust the coarse prediction by taking into account local details such as object and wall edges. The fine-scale network stack is composed solely of convolutional layers, with only one pooling stage for the edge features in the first layer.

Although the coarse network has a complete view of the scene, an output unit in the fine network only covers a field of view of 45x45 pixels of input. To achieve a relatively high-resolution output at $\frac{1}{4}^{th}$ of the input scale, convolutional layers are applied across feature maps at the target output size.

The output of the coarse network, which predicts the overall structure of the depth map, is used as an additional feature map for the fine network. This coarse prediction has the same spatial size as the output of the first fine-scale layer, and the two are combined by concatenating them. Later layers of the fine network maintain this size using zero-padded convolutions, which helps to incorporate the global understanding of the scene while also refining the predictions locally.

C. Scale Invariant Error

The inherent uncertainty in predicting the depth of a scene stems from the global scale, making it a crucial factor. In fact, the accuracy of predicting the average depth alone can account for a significant portion of the errors observed when employing current elementwise metrics. To handle the issue,

Fig. 4. Original coloured images, the corresponding ground truth depth maps from the NYUDepth dataset and the coarse and fine layer predictions of the depth maps in that order (from top to bottom). The coarse network output shows a global view of the scene which is almost the same for all the images shown. The fine layer output shows depth map predictions with much finer details.

the following scale-invariant mean squared error (in log space) was proposed.

where $\alpha(y_i, y_i^*) = \frac{1}{n}\sum_i(\log y_i - \log y_i^*)$ is the value of α that minimizes the error for a given (y, y^*). For any prediction y, e^α is the scale that makes it best aligned to the ground truth. Clearly, all the scalar multiples of y have the same error, and hence the scale invariance.

We can rearrange (1) to obtain the following equivalent expressions for the scale-invariant error.

$$D(y, y^*) = \frac{1}{n^2}\sum_{i,j}((y_i - y_j) - (y_i^* - y_j^*))^2 \quad (1)$$

$$= \frac{1}{n}\sum_i d_i^2 - \frac{1}{n^2}\sum_{i,j} d_i d_j \quad (2)$$

$$= \frac{1}{n}\sum_i d_i^2 - \frac{1}{n^2}(\sum_i d_i)^2, \quad (3)$$

where $d_i = \log y_i - \log y_i^* \ \forall i \in \{1, \cdots, n\}$. Eqn. (2) expresses the error as a comparison of relationships between pairs of pixels i, j in the output: to have a low total error, each pair of pixels should differ by an amount nearly equal to the difference in the corresponding pixels in the ground truth (depth) image. Eqn. (3) can be used to relate the scale-invariant error to the (log scale) MSE with an additional $\frac{1}{n^2}\sum_{i,j} d_i d_j$ term which penalizes mistakes in the opposite direction and credits them if they are in the same direction. Thus, an imperfect prediction will have a lower error when its mistakes are consistent in direction with each other. Finally, eqn. (4) expresses the error in a linear-time-computable form.

D. Training Loss

In addition to its use as a performance evaluation metric, the scale-invariant error was used as a training loss with the following modification:

$$L(y, y^*) = \frac{1}{n}\sum_i d_i^2 - \frac{\lambda}{n^2}(\sum_i d_i)^2, \quad (4)$$

where $\lambda \in [0, 1]$. In our training, we've used $\lambda = 0.5$ as it was suggested that this produces good absolute-scale predictions while slightly improving qualitative output.

Note that the output of the linear layer is $\log y$, i.e., the neural network predicts the log depth.

IV. Results

Using our model, we obtained a scale-invariant error of 0.14727 and RMSE of 1.46618 on the test dataset. These results have been obtained with the number of epochs used for training being 4. The number of epochs used for training the network is kept low due to GPU (CUDA) and computing power limitations. Due to the high resolution of the images in the training dataset, all attempts at training with a higher number of epochs led to CUDA running out of memory. Despite this constraint, the results obtained are acceptable and not much compromised when compared with the expected output.

Metric	Coarse + Fine Network	Baseline
RMSE (linear)	1.46618	0.871
Scale Invariant RMSE(log scale)	0.14727	0.219

TABLE II
COMAPRISON OF ERRORS OBTAINED WITH BASELINE (RESULTS OBTAINED IN [1]).

The outputs of the coarse and fine networks have been shown in Fig. 4. along with the corresponding original coloured images and ground-truth depth maps.

V. Conclusions

The results obtained from the method deployed above for depth estimation using a single RGB image has shown promising results, which can be further improved by better training (in terms of number of epochs, for example) and datasets. As more dataset becomes available in the future, with ground truth of more real-world images, the accuracy is bound to improve. Due to GPU limitations, we were able to train our

network on a few number of epochs, which can be increased as more GPU and computing power is made available.

VI. Important Links

- Link to the demo video is here
- Link to the github repo is here

References

[1] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," arXiv [cs.CV], 2014.

[2] "Datasets & DataLoaders — PyTorch Tutorials 2.0.0+cu117 documentation," Pytorch.org. [Online]. Available: https://pytorch.org/tutorials/beginner/basics/data_tutorial.html. [Accessed: 29-Apr-2023].