Single-Image Depth Estimation Based on Fourier Domain Analysis (Implementation)

Amine Khelif KHELIF[1] and Dani Bouch[2]

I. INTRODUCTION

Depth estimation plays a pivotal role in modern computer vision, serving as a cornerstone for applications like 3D model reconstruction and human pose estimation. Accurately perceiving depth from a single image, however, introduces a complex set of challenges. Unlike methods that infer depth from multiple images or video sequences, single-image depth estimation lacks vital cues such as binocular disparity and temporal motion, which can significantly aid in gauging the distance of objects.

In recent years, the emergence of deep learning has revolutionized the field, with residual convolutional neural networks (CNNs) leading the charge. These networks leverage vast amounts of data and computational power to learn representations that can predict depth from single images with remarkable accuracy. As researchers continue to innovate, the fusion of classic image processing techniques with the power of neural networks opens new pathways for robust and accurate depth estimation, even from a single viewpoint. And this is what we decided to explore through this paper

II. PROPOSED METHOD

A. CNN Architecture

Drawing from the success of ResNet-152 in image recognition, this project adapts its architecture for depth estimation. ResNet-152, originally featuring 50 blocks with 151 convolutional layers, is renowned for enabling effective feature extraction via deep layering and **shortcut connections**, critical in combating the vanishing gradient problem.

The innovation lies in modifying the last 19 blocks, integrating **additional paths for intermediate feature extraction**. These paths include two convolutional layers with kernels sized 1×3 and 3×1, arranged to extract and emphasize subtle spatial information necessary for depth nuances. This strategic addition, denoted by BC, C' where C' signifies the new channel count, allows the network to capture a comprehensive depth profile.

$$BC, C' : \textbf{Modified Block} \quad BC : \textbf{Original Block}$$

The aggregated feature maps are concatenated to form a rich representation, ultimately processed through a fully connected layer to produce an array of 800 depth estimations for a refined 25×32 depth map. The utilization of ResNet-152 as a starting point, complemented by the modifications for depth-specific feature extraction, exemplifies a targeted approach towards nuanced depth mapping.

B. Depth-Balanced Euclidean Loss

In single-image depth estimation, the traditional Euclidean loss, given by $L_E = \frac{1}{2N}\sum_x (\hat{d}_x - d_x)^2$, is commonly used. However, this loss function tends to disproportionately affect depth estimation of distant objects due to the larger absolute errors for greater true depths d_x. To mitigate this issue, we introduce the Depth-Balanced Euclidean (DBE) loss:

$$L_{DBE} = \frac{1}{2N}\sum_x \left(g(d_x)(\hat{d}_x - d_x)\right)^2$$

where $g(d) = a_1 + \frac{a_2}{2}d^2$ is a quadratic function for balancing errors, particularly to ensure that errors at shallower depths are emphasized, improving the reliability of depth estimation across different regions of the depth map.

The parameters a_1 and a_2 are chosen to balance the impact of depth errors; a_1 is a relatively large number to give more weight to errors in shallower regions, and a_2 is a negative number to counteract the overemphasis of errors at greater depths. This tailored loss function enables the network to focus on accurate depth predictions for both near and far objects, as confirmed by the experimental results.

C. Depth Map Candidate Generation

The paper introduces a technique to generate depth map candidates from a single image by strategically cropping the image. For an original image of size $W \times H$, we create a zoomed-in version by cropping the image to a size of $rW \times rH$ with $0 < r < 1$. Each cropped version provides a partial view of the scene, emphasizing details at a specific scale. To account for the zooming effect inherent in the cropping process, each depth map is scaled by a factor of $\frac{1}{r}$ to align with the original image scale. This step is crucial as objects in a cropped image appear closer and must be adjusted to maintain consistent depth estimation.

The scaled depth maps from different crops are then merged to form a comprehensive depth map. This approach leverages the varying perspectives and scales of the cropped images to improve the robustness of the estimated depth map. The final depth map, denoted as $\hat{D}^r$ for a given ratio r, combines these scaled estimates to achieve an accurate representation of depth across the entire scene.

D. Depth Map Combination in the Frequency Domain

Depth maps with a larger cropping ratio r correspond to the **overall depth distribution** (**low frequencies**), while a smaller r provides more insight into **local details** (**high frequencies**)

To enhance depth estimation, the paper leverages Fourier analysis. Depth map candidates are transformed into the frequency domain via the 2D-DFT, with coefficients rearranged into column vectors for computational efficiency. A **bias term** b_k^m is introduced to address variance in depth estimation error, defined as:

$$b_k^m = \frac{1}{T}\sum_{t=1}^{T}(f_{k,t}^m - f_{k,t})$$

The bias compensates for the average deviation from the ground truth across the training dataset of T images.

The weight vector w_k is then determined to minimize the Mean Squared Error (MSE) between the estimated and true depth coefficients:

$$w_k = T_k^+ t_k$$

where T_k^+ is the pseudo-inverse of T_k, encapsulating the differences between DFT coefficients.

Upon optimizing bias and weights, the DFT vectors are combined and the inverse Fourier transform is applied to obtain the final depth map $\hat{D}$, effectively minimizing MSE and ensuring accurate depth reconstruction.

III. EXPERIMENTAL RESULTS

The efficacy of depth estimation models was rigorously validated using the NYUv2 dataset. Inference was carried out on the standard ResNet-152 model as well as a modified version incorporating Fourier domain analysis. Comparative assessments were made by juxtaposing the predicted depth maps against the ground truth for the input images.

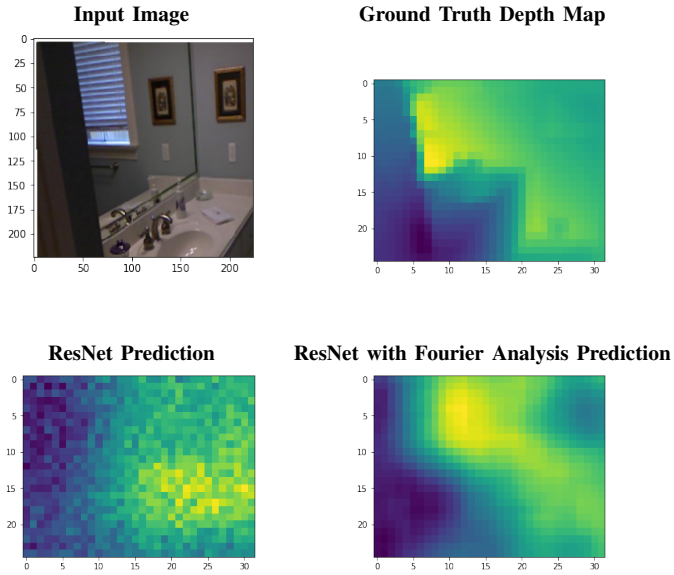

TABLE I
COMPARISON OF IMAGES AND CORRESPONDING DESCRIPTIONS.

The visual outcomes distinctly showcase the improvement in depth prediction accuracy achieved by implementing Fourier domain analysis, evidencing a significant leap in the model's predictive capabilities.

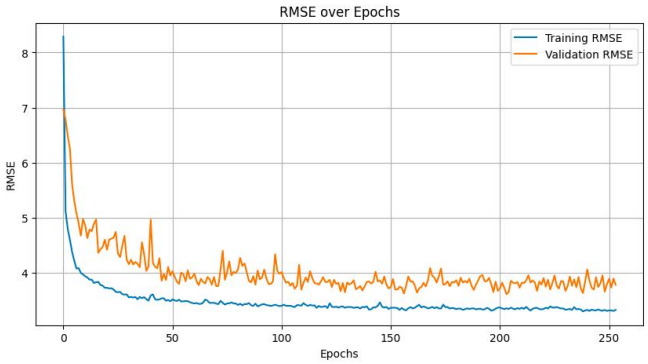

Fig. 1. RMSE

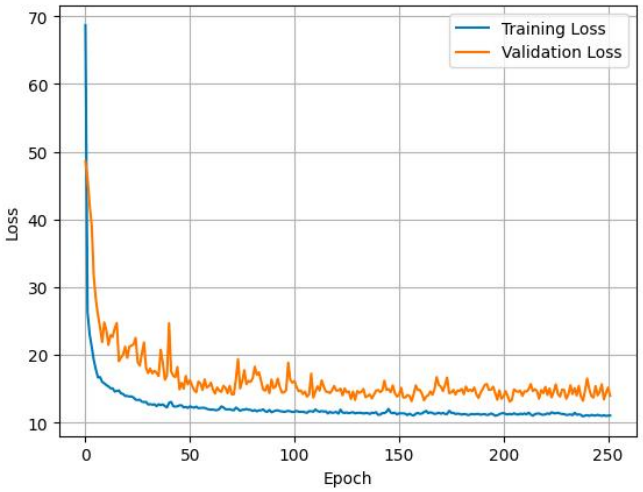

Fig. 2. Loss

Figure 1 shows a Graph depicting the Root Mean Square Error (RMSE) reduction over 250 training epochs for our ML model, illustrating the convergence behavior between the training RMSE (in blue) and the validation RMSE (in orange). Initially, both errors decrease sharply, with training RMSE stabilizing faster than the validation RMSE, which demonstrates a gradual, oscillatory decrease before reaching stability, indicating model learning and generalization

IV. CONCLUSIONS

This project provided our first comprehensive experience in designing a convolutional neural network (CNN) by adapting an existing architecture for depth estimation. We developed and integrated a custom Depth-Balanced Euclidean (DBE) loss function, which significantly enhanced our model's ability to accurately predict depth across various object scales and distances.

Training was conducted on a large dataset, enriched with augmented images, to ensure robust learning and generalization across diverse scenes. The application of Fourier domain analysis was particularly transformative, allowing us to merge multiple depth map predictions into a single, more precise map.

This endeavor not only advanced our technical expertise in neural networks and computer vision but also demonstrated the practical impact of innovative image processing techniques such as Fourier analysis in improving the accuracy of depth estimation.

In sum, this project has been a pivotal step in our journey towards mastering deep learning technologies and sets the stage for future innovations in the field.