CS482/682 Final Project Report Group 12

Monocular Depth Estimation from RGB Videos

Rishi More/rmore2, Mingqi Sheng/msheng4,
Rahul Chemitiganti/rchemit1, Rohan Allen/rallen67

1 Problem Statement

Depth estimation from monocular images is a fundamental task in computer vision with applications in autonomous driving, robotics, and scene understanding. Accurate depth information enables systems to understand the geometric structure of the environment, which is crucial for navigation and interaction. However, obtaining depth information from a single image is inherently challenging due to the loss of the third dimension in the projection from the world to the image plane. The primary objective of this project is to develop a self-supervised learning framework for monocular depth estimation that leverages the temporal coherence of RGB video sequences to generate pseudo-ground truth depth maps with limited labeled data.

2 Methods

Dataset Summary The KITTI dataset was used for this project, which provides raw stereo images and corresponding depth maps captured from a driving car in urban environments. The dataset contains various scenes with diverse lighting and weather conditions. For self-supervised learning, we utilized consecutive frames from the monocular camera. For supervised fine-tuning, we used the corresponding ground truth depth maps provided by the dataset. All images were resized to 384×640 pixels to balance computational efficiency and detail preservation.

3 Related Work

Previous approaches to monocular depth estimation have primarily focused on supervised learning methods that require large datasets with ground truth depth information, such as the works of Eigen et al. [1] and Liu et al. [2], which demonstrated the effectiveness of CNNs for depth prediction. More recent efforts have shifted towards unsupervised and self-supervised methods, including Zhou et al. [3], who introduced view synthesis for training without ground truth depth, and Godard et al. [4], who enhanced this approach with stereo image pairs. Our work builds on these methodologies, focusing exclusively on monocular video sequences.

Setup, Training and Evaluation The depth estimation model, DepthNet, utilizes an encoder-decoder architecture, where the encoder consists of convolutional layers that extract features from the input image while reducing its spatial dimensions, and the decoder employs transpose convolutional layers to upsample these encoded features and reconstruct the depth map.

Self-Supervised Training: The model was initially trained using self-supervised learning by exploiting temporal consistency between consecutive frames. The photometric reconstruction loss was utilized, defined as:

$$L_p = \frac{1}{N} \sum_{i=1}^{N} \left| I_t(p_i) - I_{t+1}(p'_i) \right|, \tag{1}$$

where I_t and I_{t+1} are consecutive frames, p_i represents pixel coordinates, and p_i' are the projected coordinates in the next frame.

Supervised Fine-Tuning: After self-supervised training, the model was fine-tuned using supervised learning with ground truth depth maps. The supervised loss function combines:

1. **L1 Loss** (L_{L1}): Mean absolute error between predicted depth D_p and ground truth D_t.
2. **Scale-Invariant Loss** (L_{SI}): Captures relative errors and scale differences.
3. **Gradient Loss** (L_{grad}): Encourages edge preservation.

The total supervised loss is:

$$L_{total} = L_{L1} + 0.5L_{SI} + 0.5L_{grad}. \tag{2}$$

Training Details:
The training setup employs the Adam optimizer with an initial learning rate of 1×10^{-4} and utilizes a ReduceLROnPlateau scheduler to dynamically adjust the learning rate based on validation loss. A batch size of 64 is maintained for both training and validation. The training process consisted of 10 epochs of self-supervised training, followed by 10 epochs of supervised fine-tuning.

Evaluation: The model's performance was evaluated using the training and validation loss curves. Additionally, qualitative results were assessed by visualizing predicted depth maps against ground truth.

4 Results and Discussions

Training Loss Curves Figures 1 and 2 show the training and validation losses during self-supervised learning and supervised fine-tuning, respectively.

During self-supervised training, the model's loss decreased steadily, indicating successful learning of depth cues from temporal consistency.

In supervised fine-tuning, both training and validation losses decreased, demonstrating improved alignment between predicted and ground truth depth maps.

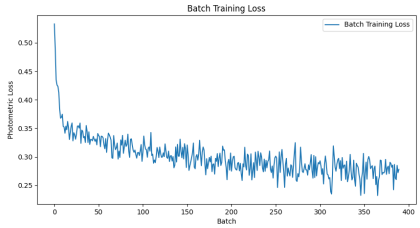

Figure 1: Batchwise Training Loss Curve during Self-Supervised Learning

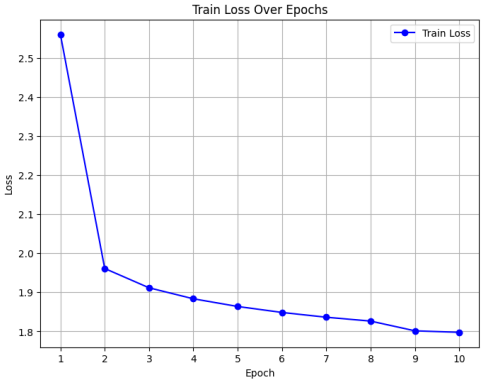

Figure 2: Epochwise Training Loss Curve during Supervised Fine-Tuning

Qualitative Results Figure 3 illustrates a sample input image, the corresponding ground truth depth map, and the predicted depth map after supervised fine-tuning. The predicted depth map closely resembles the ground truth, effectively capturing the spatial geometry and depth discontinuities.

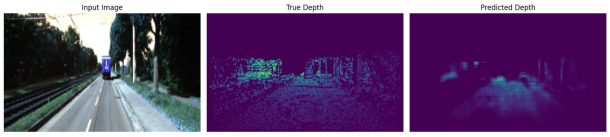

Figure 3: Sample Input Image and Corresponding Depth Maps

References

[1] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," 2014. [Online]. Available: https://arxiv.org/abs/1406.2283

[2] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, 2016. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2015.2505283

[3] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," 2017. [Online]. Available: https://arxiv.org/abs/1704.07813

[4] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," 2019. [Online]. Available: https://arxiv.org/abs/1806.01260