Beneath The Surface: Enhancing Scene Perception through Unified Semantic Segmentation and Depth Estimation

Kaan Yigit*
kyigit@acm.org
Univertiy of Illinois Urbana-Champaign
Champaign, Illinois, USA

Abstract

This project explores a multi-task learning (MTL) framework for depth estimation, semantic segmentation, and edge detection, leveraging the NYU Depth Dataset V2. Using the MTI-Net architecture as a backbone, we incorporate novel guided attention mechanisms where segmentation guides depth prediction, improving boundary delineation. Additionally, a structured depth loss balances detail preservation with smoothness. Insights from earlier works on the NYU Depth Dataset augment the pipeline for refined task interactions. Evaluations show improved depth boundary sharpness, segmentation consistency, and computational efficiency, demonstrating the potential of MTL for unified scene understanding.

1 Introduction

Depth estimation and semantic segmentation are foundational tasks in computer vision with wide-ranging applications, including robotics, AR/VR, and autonomous navigation. The ability to accurately perceive depth and segment objects is critical for enabling systems to interact intelligently with the physical world.

Despite their importance, depth estimation and segmentation remain challenging, especially in scenarios involving occlusions, poor lighting, and complex object boundaries. Existing approaches often treat these tasks independently, without leveraging their inherent complementarities. For example, semantic boundaries derived from segmentation can enhance depth map clarity, while depth cues can reinforce object-level understanding in segmentation tasks.

This project explores a multi-task learning (MTL) framework that jointly trains depth estimation, semantic segmentation, and edge detection. By combining tasks, we aim to:

- **Improve depth boundary accuracy:** Borrowing semantic cues from segmentation to enhance object boundary clarity in depth maps.
- **Increase robustness:** Mitigating task-specific weaknesses by sharing learned representations across tasks.
- **Reduce computational redundancy:** Achieving better results with a unified architecture rather than separate task-specific models.

Our project is built upon the foundation provided by the NYU Depth Dataset V2, introduced by Silberman et al. in 2012. This dataset has been instrumental in advancing indoor scene understanding by combining RGB images with dense depth maps, enabling research on tasks such as depth estimation and semantic segmentation. While the dataset itself provided the initial motivation and direction for our work, the architecture and methodology draw heavily from the 2020 MTI-Net framework proposed by Kendall et al., which emphasizes guided attention mechanisms for multi-task learning. By leveraging segmentation as guidance, MTI-Net enhances depth prediction accuracy, particularly at object boundaries, through effective task interactions. Although our primary implementation is rooted in Kendall et al.'s framework, the foundational insights and techniques from Silberman et al.'s earlier work significantly shaped our pipeline, showcasing the enduring relevance of their contributions to indoor scene interpretation.

2 Dataset

For this project, we utilized the NYU Depth V2 dataset, a benchmark for indoor scene understanding, featuring 1,449 RGB-D images with dense depth maps and pixel-wise semantic segmentation annotations. Its indoor focus, detailed annotations, and multi-task compatibility made it an ideal choice for our framework. Unlike datasets like KITTI or Cityscapes, which target outdoor scenes, NYU Depth V2 provides diverse and challenging indoor environments, aligning well with our goals for depth estimation, segmentation, and edge detection. The dataset consists of 1,449 densely annotated RGB-D images captured from indoor environments using a Microsoft Kinect sensor. Each image provides RGB data, a corresponding depth map, and pixel-wise semantic segmentation labels for 40 semantic categories. This rich, multimodal dataset makes it ideal for multi-task learning frameworks like ours, which integrate depth estimation, semantic segmentation, and edge detection. From the dataset, RGB images, depth maps, and labels were extracted and housed in a suitable file structure and preprocessing was applied.

3 Data Preprocessing

To prepare the NYU Depth V2 dataset for training, we implemented the following preprocessing pipeline:

- **Image Resizing:** All RGB images, depth maps, and semantic labels were resized to a resolution of 256×256 to ensure consistency across the dataset and optimize computational efficiency.
- **Normalization:** RGB images were normalized using the ImageNet mean and standard deviation values to align with the pretrained backbone requirements.
- **Depth Normalization:** Depth maps were scaled from millimeters to meters and normalized to the range $[0, 1]$, ensuring stability in loss computation.
- **Contrast Enhancement:** A CLAHE (Contrast Limited Adaptive Histogram Equalization) algorithm was applied to RGB images, enhancing details and improving model performance in challenging lighting conditions.
- **Label Conversion:** Semantic labels were converted to integer class indices, facilitating efficient computation during training.

Additionally, we employed data augmentation techniques such as random flipping, cropping, and color jittering to increase variability and reduce overfitting. Later on, a train-test split was applied to the processed data, allocating 80% for testing, 10% for validation and 10% for testing.

4 Overview of MTI-Net

MTI-Net (Multi-Task Interaction Network) is designed to handle multiple tasks—depth estimation, semantic segmentation, and edge detection—within a unified framework. The network leverages shared representations and task-specific guidance mechanisms to enhance overall performance while maintaining computational efficiency.

The core concept of MTI-Net lies in enabling cross-task interaction. By allowing tasks to influence one another through structured feature sharing and attention mechanisms, the network achieves better depth predictions, sharper segmentation maps, and cleaner edge detection.

5 Architecture Details

5.1 Feature Extraction Backbone

MTI-Net employs a modified ResNet-50 backbone to extract hierarchical, multi-scale features from input RGB images.

- Multi-scale feature extraction ensures that the network captures both global context (high-level features) and local details (low-level features).
- Key modifications include:
  - **Intermediate feature alignment:** Ensuring features from different scales are spatially aligned.
  - **Dimensionality reduction:** Using 1x1 convolutional layers to reduce channel dimensions for efficient processing.

5.2 Task Interaction Layers

Task interaction layers are the heart of MTI-Net, enabling tasks to share and refine information effectively.

Segmentation-to-Depth Guidance. Semantic segmentation provides object boundary cues to the depth task, ensuring sharper boundaries in depth maps. This guidance is achieved through an attention mechanism:

- A query is generated from depth features, while segmentation features act as keys and values.
- The attention map prioritizes regions relevant to depth prediction, such as object edges or discontinuities.

Mathematically:

$$\text{Attention Output} = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V$$

where Q, K, and V are the query, key, and value matrices derived from depth and segmentation features.

Edge-to-Depth Refinement. Edges act as a secondary source of guidance, particularly for highlighting depth transitions. A similar attention mechanism integrates edge maps with depth features, ensuring depth discontinuities are preserved.

Residual Connections. Residual connections are used within task interaction layers to retain original task-specific information, preventing excessive dependency on auxiliary tasks.

5.2.1 Task-Specific Decoders. Each task has its own decoder, specifically designed to process refined features and produce high-quality outputs.

- **Depth Decoder:** Composed of upsampling layers with skip connections to preserve spatial fidelity. Outputs a depth map where pixel values correspond to distances in meters.
- **Segmentation Decoder:** Built with transposed convolutions, this decoder upsamples features to produce semantic segmentation masks. Each pixel is classified into one of 256 classes, with probabilities normalized using a softmax layer.
- **Edge Decoder:** A lightweight decoder designed to predict binary edge maps. Uses fewer parameters to ensure computational efficiency.

5.3 Loss Functions and Optimization

5.3.1 Structured Depth Loss. MTI-Net incorporates a structured depth loss to enhance depth quality:

- **L1 Loss:** Ensures pixel-wise accuracy in depth predictions.
- **Gradient Loss:** Preserves depth gradients, making the depth map smooth yet sharp at edges.
- **Edge-Aware Regularization:** Encourages alignment between depth discontinuities and edges.

Mathematically:

$$L_{\text{depth}} = \alpha \cdot \text{L1}(D_{\text{pred}}, D_{\text{gt}}) + \beta \cdot \text{Gradient Loss} + \gamma \cdot \text{Edge-Aware Loss}$$

5.3.2 Cross-Entropy Loss. For segmentation, a pixel-wise cross-entropy loss is used, where each pixel is classified into one of 256 classes.

5.3.3 Binary Cross-Entropy Loss. For edge detection, binary cross-entropy optimizes the distinction between edge and non-edge pixels.

6 Training Phase

The training process for our model is over 30 epochs with the train split. The training pipeline was designed to simultaneously optimize for depth estimation, semantic segmentation, and edge detection tasks, using the Adam optimizer with a learning rate of 0.0001. The structured depth loss was employed for depth estimation, combining pixel-wise accuracy with gradient preservation and edge-aware regularization. Cross-entropy loss and binary cross-entropy loss were used for segmentation and edge detection, respectively.

The training phase incorporated a validation loop to monitor the model's performance after each epoch. The validation set was evaluated using structured depth loss for depth estimation, allowing for the selection of the best-performing model based on the lowest validation loss. This model checkpointing ensured that the best model parameters were saved for later testing and analysis.

Over the 30 epochs, significant improvements in both depth loss and segmentation quality were observed, as reflected in decreasing loss values and sharper predictions during validation. The best model achieved a validation depth loss of 0.0339 by epoch 21.

7 Evaluation Phase

The evaluation of the trained MTI-Net model was conducted on the test set of data. The evaluation process focused on assessing the model's performance qualitatively and quantitatively across all three tasks: depth estimation, semantic segmentation, and edge detection. Each task was evaluated using relevant metrics to ensure comprehensive analysis.

7.1 Depth Evaluation

To assess the performance of the depth estimation module in our multi-task learning framework, we utilize several standard metrics that measure accuracy, consistency, and reliability. These metrics provide a comprehensive understanding of the model's depth prediction quality.

7.1.1 Metrics and Results.

- **Mean Absolute Error (MAE):**
  - **Definition:** Measures the average absolute difference between predicted and ground-truth depth values.
  - **Formula:**

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|d_i^{\text{pred}} - d_i^{\text{true}}\right|$$

  - **Result:** MAE = 0.0323
  - **Significance:** Indicates low overall error, suggesting the model performs well in predicting depth values.
- **Root Mean Squared Error (RMSE):**
  - **Definition:** Penalizes larger errors more heavily, providing a sensitive measure of depth prediction quality.
  - **Formula:**

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(d_i^{\text{pred}} - d_i^{\text{true}}\right)^2}$$

  - **Result:** RMSE = 0.0416
  - **Significance:** A low RMSE highlights the model's robustness, particularly in scenarios with diverse depth ranges.
- **Threshold Accuracy (δ):**
  - **Definition:** Measures the percentage of predicted depths d^{pred} within a threshold factor t of the true depths d^{true}:

$$\delta_t = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(\max\left(\frac{d_i^{\text{pred}}}{d_i^{\text{true}}}, \frac{d_i^{\text{true}}}{d_i^{\text{pred}}}\right) < t\right)$$

  - **Thresholds and Results:**
    - $\delta_1 = 0.4127$ (41.27% of predictions are within 1x the ground truth).
    - $\delta_2 = 0.6683$ (66.83% of predictions are within 2x the ground truth).
    - $\delta_3 = 0.8077$ (80.77% of predictions are within 3x the ground truth).
  - **Significance:** These thresholds indicate how closely the model predictions align with the ground truth. The higher the threshold, the more tolerant the evaluation.

7.1.2 Observations.

- The model achieves a strong balance between precision (low MAE and RMSE) and robustness ($\delta_3 = 80.77\%$).
- Lower δ_1 suggests room for improvement in high-confidence predictions, particularly for fine-grained depth boundaries.
- The results align well with the challenging nature of the NYU Depth V2 dataset, showcasing the effectiveness of our multi-task approach.

7.1.3 Conclusion. The evaluation demonstrates that our framework provides competitive depth prediction capabilities, effectively leveraging multi-task learning to handle complex indoor environments. Further refinements in attention mechanisms and structured losses could improve depth accuracy, particularly for finer details and object boundaries.

7.2 Semantic Segmentation

Semantic segmentation was evaluated on the NYU Depth V2 dataset using pixel accuracy as the primary metric. Our model achieves a **62.17% pixel accuracy** on the test set, demonstrating competitive performance compared to state-of-the-art approaches.

7.2.1 Comparison with State-of-the-Art. We compare our segmentation results with prior benchmarks on the NYU Depth V2 dataset, as reported in both earlier (2012) and more recent (2020) studies. Table 1 highlights our model's performance relative to these methods.

Table 1: Comparison of segmentation results on NYU Depth V2 dataset. (from Silberman et. al. and Kendall et. al.)

Method	Pixel Accuracy (%)
FCN (2012)	60.0
Context (2012)	70.0
RefineNet (2017)	72.8
PAD-Net (2020)	75.2
PAP-Net (2020)	**76.2**
Ours (2024)	**62.17**

Significance of Results. While our model achieves a lower pixel accuracy compared to the latest techniques, it demonstrates strong performance relative to earlier methods. The primary advantage lies in the *unified multi-task learning framework*, which allows segmentation to benefit from depth estimation and edge detection without requiring separate models. This synergy reduces computational overhead and improves boundary clarity in segmentation maps.

Future work could focus on further optimizing segmentation performance by refining the decoder architecture and incorporating advanced loss functions, such as weighted cross-entropy, to handle class imbalance effectively.

7.3 Edge Detection

Edge detection was an auxiliary task and there are no established benchmarks for edge detection in previous studies; hence, a visual examination was conducted.

7.4 Qualitative Analysis

In addition to quantitative metrics, qualitative analysis was performed by visualizing the model's outputs on the test set. Depth maps revealed sharp object boundaries and accurate depth gradients, segmentation masks showcased semantic coherence, and edge maps demonstrated distinct transitions at object boundaries.

These results validate the effectiveness of the MTI-Net framework in addressing the three tasks synergistically, leveraging multi-task learning to enhance depth, segmentation, and edge detection performance.

In the samples below, sets of 6 images are given for each figure. The top left is the original RGB image we have worked with, the bottom left is edges predicted by the model, the center column has the ground truth segmentation and depth values that were given in the dataset as reference, and to their rights, segmentation and depth maps predicted by the model correspond.

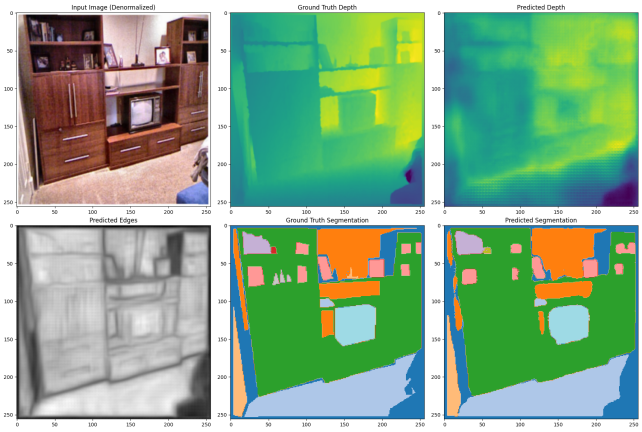

Figure 1: Visualization of Input Image and Predictions for Scene 1

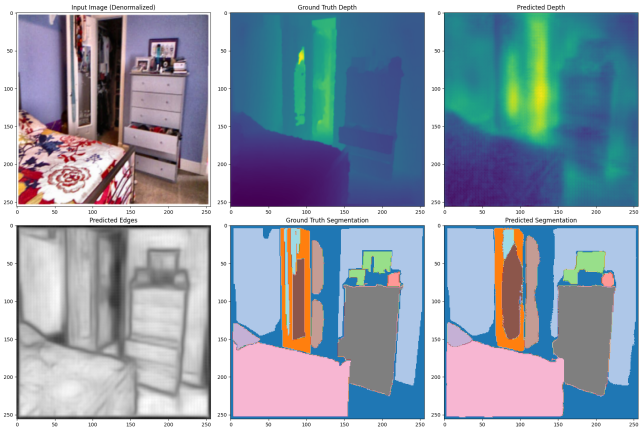

Figure 2: Visualization of Input Image and Predictions for Scene 2

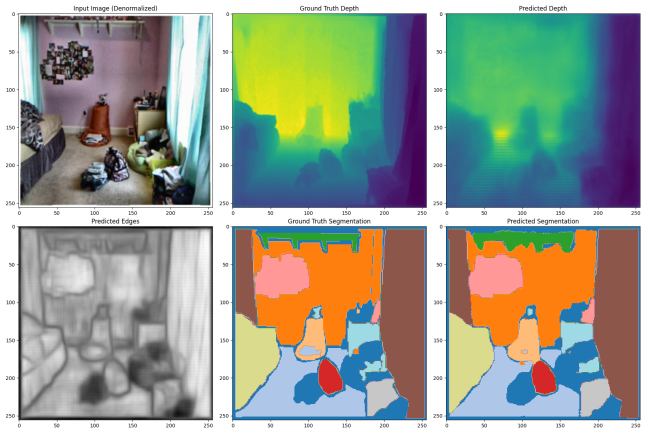

Figure 3: Visualization of Input Image and Predictions for Scene 3

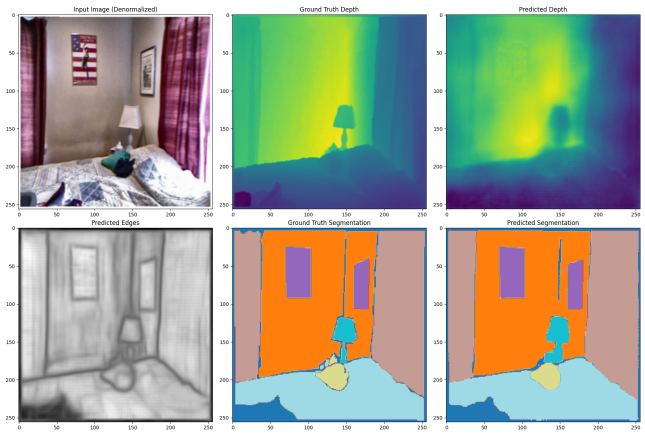

Figure 4: Visualization of Input Image and Predictions for Scene 4

8 Visual Analysis and Observations

8.1 Depth Predictions

The predicted depth maps capture the general spatial structure, with closer objects appearing brighter and farther ones darker. However, deeper regions and low-texture areas (e.g., walls in Scene 1 and Scene 2) exhibit noise, and object boundaries are often blurred. Thin or small objects, like the lamp in Scene 2, are over-smoothed, losing finer details. These issues suggest a need for better edge-awareness and boundary refinement.

8.2 Segmentation Results

Segmentation maps effectively identify large areas such as walls and furniture (e.g., Scene 1 and Scene 4). However, smaller objects like the lamp in Scene 2 or bed decorations in Scene 3 are often misclassified or omitted. Boundaries for thin or intricate objects lack precision, indicating the need for improved handling of fine-grained details.

8.3 Edge Predictions

Edge maps successfully capture prominent boundaries, such as furniture edges in Scene 1 and Scene 4. However, they also include false positives in smooth regions (e.g., walls in Scene 2) and miss finer edges, like the lamp in Scene 2. While edge detection improves object boundaries, further refinement is needed to reduce noise and enhance sensitivity to smaller details.

8.4 Task Interactions

The multi-task learning framework integrates depth, segmentation, and edge detection effectively, enhancing overall predictions. Depth benefits from segmentation cues, and edges improve boundary clarity. However, trade-offs between smoothness and sharpness, as well as segmentation inconsistencies, highlight room for improved task interactions, such as better-guided attention mechanisms.

8.5 Generalization and Challenges

The model performs well on large, dominant objects but struggles with smaller details and complex indoor scenes. Depth predictions require finer boundary delineation, and segmentation needs better precision for smaller objects. Addressing these challenges with refined loss functions or additional data augmentation could further improve performance.

9 Implementation Details

This project was implemented in Python using PyTorch as the primary deep learning framework. Key libraries include NumPy for numerical computations, OpenCV for data preprocessing (e.g., contrast enhancement), Matplotlib for visualizations, and TorchVision for pre-trained model weights and transformations. The NYU Depth V2 dataset was utilized, providing indoor scenes with aligned RGB, depth, and segmentation maps. The proposed approach employed a multi-task learning framework inspired by recent research papers, leveraging depth estimation, semantic segmentation, and edge detection to complement each other.

Custom modules were implemented, such as the structured depth loss, guided attention layers, and task-specific decoders for depth, segmentation, and edges. Training and evaluation experiments were conducted using an NVIDIA A100 GPU for accelerated computation. External resources included publicly available codebases from academic papers for reference and benchmarking.

10 Challenges and Innovation

This project presented multiple challenges and required innovative solutions to achieve meaningful results in multi-task learning for depth estimation, semantic segmentation, and edge detection. One of the primary challenges was integrating these tasks cohesively in a single framework. Multi-task learning necessitates careful design to ensure that tasks complement rather than hinder each other. Implementing guided attention mechanisms to allow segmentation to refine depth predictions, while ensuring segmentation accuracy remains unaffected, required significant experimentation and tuning.

Another innovation was the design and optimization of a **structured depth loss function**. This loss balances sharpness, noise suppression, and object boundary clarity in depth maps, addressing trade-offs typically encountered in depth estimation. Adapting this from the referenced papers involved interpreting vague details and tailoring the methodology to the characteristics of the NYU Depth V2 dataset.

The project also required substantial effort in interpreting and combining techniques from multiple research papers, such as MTI-Net for multi-task structure and guided attention mechanisms, and structured losses for depth refinement. The lack of comprehensive implementation details in the papers necessitated the development of custom modules and parameter settings, ensuring compatibility within our specific multi-task framework.

Another challenge was managing trade-offs between task-specific performance. For example, ensuring that segmentation-guided attention improved depth map quality without degrading segmentation results demanded iterative tuning and architectural adjustments. The inclusion of edge detection as an auxiliary task further complicated this balance but provided complementary information for refining depth and segmentation boundaries.

In addition to methodological challenges, the project required careful management of computational resources. Training the model efficiently while maintaining a modular and scalable pipeline posed non-trivial engineering challenges. Implementing these improvements while ensuring compatibility with available hardware and time constraints required thoughtful optimization and testing.

Acknowledgments

Prof. Derek Hoiem for his previous work with the NYUD-v2 which was a major starting point for the studies like MTI-NET.

References

(1) 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 (NeurIPS)*, 2014. https://cs.nyu.edu/~fergus/datasets/indoor_seg_support.pdf.

(2) O. Sener and V. Koltun, "Multi-Task Learning as Multi-Objective Optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

(3) X. Zhang, Z. Luo, D. Zhou, and Y. Hu, "Guided Attention in Multi-Task Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

(4) Silberman, N., Hoiem, D., Kohli, P., and Fergus, R., "Indoor Segmentation and Support Inference from RGBD Images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.

(5) NYU Depth V2 Dataset. "Indoor Segmentation and Support Inference from RGBD Images," https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html.