Monocular Depth Estimation and 3D Reconstruction Approaches For Endoscopic Videos

Shivank Gupta
University of Pennsylvania
shivankg@seas.upenn.edu

Abraham Paroya
University of Pennsylvania
aparoya@seas.upenn.edu

***Abstract*—Accurate depth estimation in endoscopic imagery is a critical challenge for enhancing surgical navigation and 3D scene understanding in minimally invasive procedures. This paper investigates monocular depth estimation using the EndoSLAM dataset by comparing a lightweight convolutional neural network (CNN) architecture with Depth Anything, a transformer-based foundation model. We train and evaluate models using pixel-wise error metrics (MSE, MAE, RMSE) and perceptual similarity (SSIM), and explore architectural improvements such as ResNet-based encoders and perceptual loss functions. Beyond quantitative benchmarks, we implement a pipeline for 3D reconstruction from predicted depth maps and camera poses, generating qualitative point cloud and mesh visualizations of gastrointestinal anatomy. This work highlights key trade-offs between model complexity, generalization, and spatial reconstruction quality in medical vision tasks.**

I. Introduction

Minimally invasive surgical (MIS) procedures have become increasingly common and crucial to understanding simultaneous localization and mapping (SLAM) within the human body. The lack of a kinematic view enforces that depth estimation and reconstruction be performed on a single endoscopic view. Surgeons primarily rely on monocular cues, such as shading, texture gradients, motion parallax, and their experience to infer depth and understand the 3D structure. However, this monocular view naturally lacks metric depth information, which is challenging in complex and confined spaces.

Solving the problem of accurate 3D perception in endoscopy is important for several reasons. Providing surgeons with a clear and accurate 3D understanding of the surgical field allows for better orientation and navigation within complex anatomy, reducing the risk of disorientation and accidental tissue damage. Accurate depth information allows for more precise manipulation of surgical instruments relative to tissue structures. Real-time guidance information could enable improved planning and execution, leading to a decrease in post-operative failures. Lastly, realistic 3D reconstructions can be used to create immersive and accurate surgical training simulations, providing objective assessment of their performance.

The EndoSLAM dataset provides standard and capsule endoscopy data from ex-vivo gastrointestinal tract organs with time-synchronized, high-precision 6D ground truth pose, as well as a video sequence recorded by a clinically in-use colonoscope from a fully representative silicon colon phantom with CT scan ground truth. It also includes synthetically generated data with pixel-wise depth and 6D pose ground truth, which facilitates the simulation to real domain adaptation algorithms. The dataset is divided into 35 sub-datasets. Specifically, 18, 5, and 12 sub-datasets exist for the colon, small intestine, and stomach respectively.

II. Contributions

This project explores monocular depth estimation for endoscopic imagery using both custom convolutional networks and pretrained models. The main contributions of this work include:

- Implementing a baseline 3-layer CNN model achieving an MSE of 0.0343 and SSIM of 0.6140.
- Improving accuracy by replacing the baseline encoder with a ResNet-18 backbone, reducing MSE to 0.0125 and increasing SSIM to 0.8633.
- Evaluating Depth Anything, a foundation model, on EndoSLAM data in a zero-shot setting.
- Performing a quantitative comparison across models using four metrics: MSE, MAE, RMSE, and SSIM.
- Generated qualitative heatmap visualizations and 3D point cloud reconstructions from predicted depth maps to assess spatial consistency.

III. Related Work

Recent advancements in monocular depth estimation and SLAM for medical imaging have been driven by the development of domain-specific datasets and general-purpose vision foundation models.

Srivastava et al. [1] introduced the EndoSLAM dataset, which provides monocular endoscopic video sequences paired with depth ground truth obtained from a CT-registered phantom. This dataset enables benchmarking of SLAM and depth estimation algorithms in challenging, real-world gastrointestinal environments. The associated GitHub repository [2] provides tools and sequences to support reproducible research in visual SLAM for endoscopy.

Meanwhile, foundation models such as Depth Anything [3] represent a shift toward generalizable depth estimators trained on diverse Internet-scale data. The model leverages vision transformers to produce high-quality depth predictions across a variety of input domains, including medical imagery, without requiring task-specific fine-tuning.

These resources form the basis for evaluating traditional CNN-based architectures against pretrained transformer-based models in this work. By building on EndoSLAM's data and Depth Anything's pretrained capabilities, our study bridges domain-specific and general-purpose depth estimation in the context of minimally invasive surgery.

IV. Approach

A. CNN Architecture

For this project, a convolutional neural network (CNN) was developed to estimate depth from monocular endoscopic images. This architecture aims to balance accuracy and efficiency, making it suitable for systems with limited computational resources while remaining scalable for larger datasets.

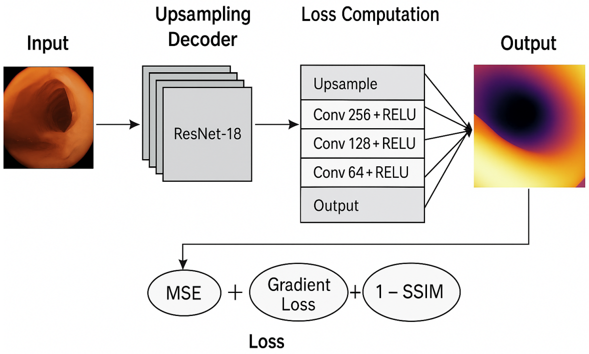


Fig. 1. CNN Architecture Pipeline

The data set is made up of RGB endoscopic frames and their corresponding grayscale depth maps. Before training, both input and ground truths undergo preprocessing through a transformation pipeline that resizes them to 128×128 and converts them to normalized tensors. In later iterations, the images were scaled to 256×256 resolution to create higher-resolution maps. Depth normalization plays a crucial role, as raw values vary significantly; scaling by the maximum value constrains predictions within a [0, 1] range. The model uses an encoder-decoder design; the encoder includes three convolutional layers with progressively increasing channels (32, 64, and 128), interspersed with max-pooling to reduce spatial dimensions and extract abstract features. This setup captures essential structures without excessive computational demand. The decoder follows a symmetric layout, employing three transposed convolution layers to reconstruct the spatial resolution from the latent representation. Each is followed by a ReLU activation, except for the final layer, which uses a sigmoid function to output a normalized depth map in the [0, 1] interval.

Through iterations of training, several enhancements were made to improve the smoothness and realism of the predictions. A key improvement was the introduction of a gradient-based smoothness loss, which penalizes high-frequency noise by aligning spatial derivatives of predicted and ground-truth maps. This helped produce smoother, more realistic outputs. In addition, a Structural Similarity Index Measure term was added to encourage perceptual similarity and preserve structural details. Later in development, the architecture was further upgraded to incorporate a pre-trained ResNet18 backbone as the encoder. This significantly improved feature extraction by leveraging the weights learned on ImageNet. The decoder was redesigned to use bilinear up-sampling followed by convolution instead of transpose convolutions, which helped reduce checkerboard artifacts (see Figure 2) and further smoothed the predictions.

Training utilizes a composite loss that balances pixel accuracy with perceptual quality. The first component is Mean Squared Error, minimizing absolute deviations from the ground truth. Gradient-based loss compares spatial derivatives of predicted and true depth, while SSIM encourages fidelity and contrast. Together, these objectives ensure the network produces depth maps that are both numerically precise and visually coherent. The model is trained using the Adam optimizer with a learning rate of 1e-4. Training runs in batches, on GPU when available. Debug logs regularly output loss components and prediction ranges, helping to monitor performance and convergence. Both CNN models were trained using a NVIDIA GeForce GTX 1660 Ti GPU.

The loss function of the depth estimator is comprised of three terms; MSE, Gradient Loss and SSIM:

1) Mean Squared Error (MSE) Loss:

$$L_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left(D_i^{\text{pred}} - D_i^{\text{gt}}\right)^2$$

2) Gradient Loss (Smoothness Term):

$$L_{\text{grad}} = \frac{1}{N} \sum_{i=1}^{N} \left(\left| \nabla_x D_i^{\text{pred}} - \nabla_x D_i^{\text{gt}} \right| + \left| \nabla_y D_i^{\text{pred}} - \nabla_y D_i^{\text{gt}} \right| \right)$$

3) Structural Similarity Index (SSIM) Loss:

$$L_{\text{SSIM}} = 1 - \text{SSIM}(D^{\text{pred}}, D^{\text{gt}})$$

4) Combined Loss Function:

$$L_{\text{total}} = L_{\text{MSE}} + \lambda_1 L_{\text{SSIM}} + \lambda_2 L_{\text{grad}}$$

The loss weights $\lambda_1 = 0.05$ for SSIM and $\lambda_2 = 0.2$ for gradient loss were chosen to balance accuracy and perceptual quality. MSE serves as the primary objective for pixel-wise correctness, while the SSIM term encourages structural fidelity without dominating the loss. The higher weight on gradient loss promotes smooth and coherent depth maps by penalizing sharp, unnatural transitions.

B. Depth Anything

To benchmark our CNN-based depth estimation models, we incorporated Depth, a transformer-based model, for comparative evaluation. Depth Anything is a pretrained, zero-shot monocular depth estimator. Depth Anything requires no additional fine-tuning or task-specific training, enabling direct inference on unseen endoscopic frames. The motivation for including Depth Anything was to assess how well a general-purpose depth estimation model, trained on large-scale natural image datasets, performs on domain-specific medical imagery like endoscopy. Given that Depth Anything has been trained on diverse scenes, its performance provides a strong reference point for evaluating the generalization ability and limitations of lightweight CNN models trained on smaller, domain-specific datasets.

During evaluation, we passed the same set of RGB endoscopic frames used for CNN inference through the Depth Anything pipeline. The output-predicted depth maps—were post-processed for normalization and visualization. We saved the raw predictions as .npy files for quantitative analysis and converted them into heatmaps for visual comparison. Since Depth Anything operates in a zero-shot manner, the quality of its predictions directly reflects the model's pretraining, without any domain adaptation.

Quantitative comparisons were conducted using standard depth estimation metrics, including Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, and Structural Similarity Index Measure. By contrasting these metrics against our CNN models trained specifically on the endoscopy dataset, we were able to evaluate how much performance can be gained through domain adaptation versus large-scale generalization.

The inclusion of Depth Anything added a valuable dimension to our analysis, highlighting the trade-offs between custom lightweight architectures and foundation models trained at scale. Despite being zero-shot, Depth Anything often yielded competitive visual quality and smoothness in the predicted depth maps, making it a strong baseline for future explorations in endoscopic depth estimation.

C. 3D Reconstruction

Once the training and evaluation proved successful, we had all the data needed to perform 3D reconstruction. RGB images were stored in `eval_frame/` and depth predictions (in `.npy` format) were located in `output_depth/npy/`. Each camera from the EndoSLAM dataset had a corresponding intrinsic camera matrix as well as the pose for each frame.

The depth maps predicted by our model were converted to the Open3D-compatible format. Since the depth predictions were already in meters, we used a depth scale of 1.0 and truncated depth values exceeding 1.0m to remove distant artifacts. Camera intrinsics were loaded from a 3×3 matrix:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

This was parsed and converted to an Open3D `PinholeCameraIntrinsic` object. The camera poses were parsed from the CSV file containing translation (t_x, t_y, t_z) and quaternion rotation (q_x, q_y, q_z, q_w) parameters.

Each quaternion was normalized and converted into a 3×3 rotation matrix R using:

$$R = \begin{bmatrix} 1 - 2q_y^2 - 2q_z^2 & 2q_xq_y - 2q_zq_w & 2q_xq_z + 2q_yq_w \\ 2q_xq_y + 2q_zq_w & 1 - 2q_x^2 - 2q_z^2 & 2q_yq_z - 2q_xq_w \\ 2q_xq_z - 2q_yq_w & 2q_yq_z + 2q_xq_w & 1 - 2q_x^2 - 2q_y^2 \end{bmatrix}$$

The full 4×4 transformation matrix T was constructed as:

$$T = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}$$

We converted each frame to an RGBD image using Open3D's `create_from_color_and_depth()` function. Using nearest-neighbor interpolation, we resized the depth and RGB images to match dimensions.

We then created a point cloud P_i for each frame by projecting depth pixels into 3D using:

$$P_i = T_i \cdot K^{-1} \cdot D(x, y)$$

where T_i is the camera-to-world transformation matrix and $D(x, y)$ is the depth value at pixel (x, y). The resulting point clouds were accumulated and transformed into the global coordinate frame.

Using voxel-based filtering with voxel size $\delta = 0.001$, we downsampled the point clouds. To remove any noisy points, we applied statistical outlier removal with $k = 20$ neighbors and standard deviation ratio $\sigma = 2.0$. Implementing these methods, we successfully reconstructed 3D-point clouds with our predicted depth maps and began qualitatively assessing performance for tuning.

V. Results

A. Depth Estimation

We evaluated three depth estimation models — (1) a baseline 3-layer CNN, (2) an enhanced CNN with ResNet18 encoder and upsampling decoder, and (3) Depth Anything,. The predicted depth maps were qualitatively and quantitatively compared to ground truth depth images using standard metrics including MSE, MAE, RMSE, and SSIM.

Figure 2 illustrates the performance of the baseline CNN. While it successfully captures coarse depth structure, it struggles with finer anatomical details and exhibits visible blocky artifacts, likely due to limited receptive field and lower model capacity. Figure 3 compares outputs from the ResNet18-based architecture, which significantly improves smoothness and alignment with ground truth. The decoder's bilinear upsampling and inclusion of gradient and SSIM losses appear to help preserve geometry and reduce artifacts.

In Figure 4, predictions from Depth Anything demonstrate superior quality. The output depth maps are smooth, structurally consistent, and visually plausible even in challenging regions with lighting variation or low texture. Notably, the perceptual quality of the predicted depth maps aligns closely with the ground truth and exhibits strong robustness across frames.

Quantitative evaluation confirmed these observations. The ResNet-based model outperformed the baseline CNN in all metrics, and Depth Anything achieved the best SSIM, indicating high structural fidelity. A sample comparison of heatmaps and corresponding raw images is provided in Figure 4, further supporting the perceptual gains from pretraining and transformer-based modeling.

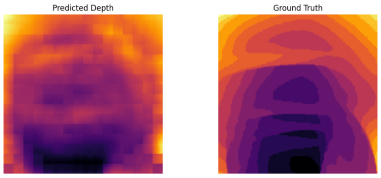
,


Fig. 2. Base CNN Model Heatmap

To quantitatively assess the impact of architectural enhancements, we compared the 3-layer CNN with a ResNet-18 encoder-based model and the Depth Anything transformer-based architecture. The results are summarized in Table 1. The ResNet-18 model achieved the best performance across all evaluation metrics. Specifically, Mean Squared Error dropped from 0.0343 (3-layer CNN) to 0.0125, and the Mean Absolute Error improved from 0.1543 to 0.0888. The Root Mean Squared Error decreased by over 40%, and the Structural Similarity Index increased from 0.6140 to 0.8633, indicating superior structural preservation. Depth Anything, while not outperforming the ResNet-18 model, achieved a higher SSIM (0.7146) than the baseline 3-layer CNN, suggesting improved perceptual quality despite a higher MSE (0.0989) and RMSE (0.2969). These results demonstrate that while transformer-based models like Depth Anything can generalize well, they may benefit from further fine-tuning on domain-specific data. Overall, the improvements underscore the value of leveraging deeper or pretrained architectures for depth estimation in endoscopic imagery.

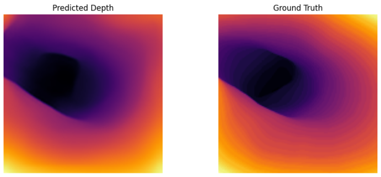


Fig. 3. ResNet18 and Improved Loss Function Model Heatmap

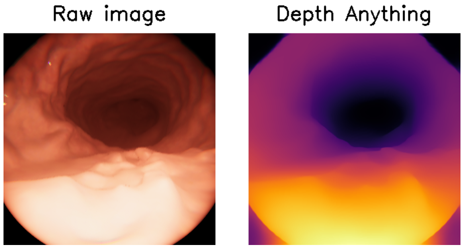


Fig. 4. DepthAnything Heatmap

Model	MSE	MAE	RMSE	SSIM
3-Layer CNN	0.0343	0.1543	0.1848	0.6140
ResNet-18	0.0125	0.0888	0.1075	0.8633
DepthAnything	0.0989	0.2550	0.2969	0.7146

TABLE I
Comparison of depth estimation performance across different models.

B. 3D Reconstruction

As seen in Figure 5, the initial reconstruction was limited in depth, and appeared as layers of depth maps rather than construct a continuous model. This used a depth estimation model trained on 500 images and performed the reconstruction with only 50 frames. The reconstruction noticeably required more tuning, and needed a model

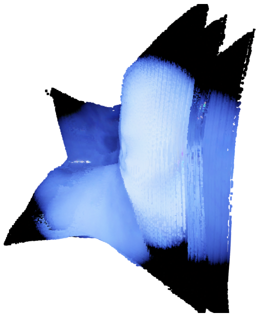


Fig. 5. Initial Stomach 3D Point Cloud Across 50 Frames

that was trained the entirety of the frames to allow for more accurate reconstruction. Specifically, aspects such as diverse lighting, motion, and geometry exposure were minimal.

Upon tuning and referencing the final depth estimation model, trained on 21,887 images, the reconstruction was performed across 500 frames. This final model produced an accurate and important representation of the stomach as seen in Figure 6. The left side shows the viewpoint from within the point cloud model and the right side shows the full reconstruction of the stomach across the entirety of the frames. The model trained, trained on more variation and scenes, allowed for better generalization and ability to construct prior scenes to create the tube-like structure we desired. We also noticed far more surface smoothness and continuity.

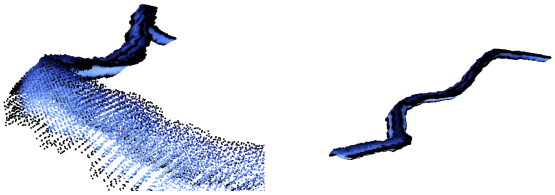


Fig. 6. Final 3D Reconstruction of Stomach Across 500 Frames

VI. Discussion and Future Work

The results demonstrate that the ResNet-18-based encoder-decoder architecture significantly outperforms the simpler 3-layer CNN across all evaluation metrics. However, both CNN-based approaches produce depth maps that appear overly smoothed in certain regions, which may obscure fine anatomical details in endoscopic imagery. Future work could explore adjusting the weight of the gradient loss term in the loss function to reduce excessive smoothing, thereby improving spatial sharpness and better preserving depth discontinuities.

In terms of 3D reconstruction, while the current pipeline effectively generates point clouds from predicted depth maps, the reconstructions remain limited in resolution and accuracy. A promising avenue for improvement lies in leveraging more advanced photogrammetry tools such as COLMAP, which can produce dense point clouds using both structure-from-motion and multi-view stereo techniques. Furthermore, integrating neural radiance field methods—specifically through frameworks like Nerfstudio—could enable more realistic and complete volumetric reconstructions from monocular endoscopic sequences.

Another direction for enhancement is the incorporation of RGB textures into the reconstruction pipeline. Currently, only the depth modality is used for 3D point generation. Mapping the original RGB pixel intensities onto the point cloud would not only improve visualization quality but also support downstream applications such as surgical scene understanding and augmented reality overlays.

Overall, combining depth estimation with advanced multi-view geometry and texture mapping pipelines has the potential to yield more photorealistic and clinically useful reconstructions from endoscopic video.

Acknowledgment

We would like to thank Professor Pratik Chaudhari and the ESE 650 team for their guidance and support.

References

[1] A. Srivastava, A. Srivastava, and P. Kannojia, “EndoSLAM Dataset for Monocular Visual SLAM in Endoscopic Videos,” Mendeley Data, V1, 2021. [Online]. Available: https://data.mendeley.com/datasets/cd2rtzm23r/1

[2] CapsuleEndoscope, ”EndoSLAM: Dataset and benchmark for visual SLAM in endoscopy,” GitHub repository, 2021. [Online]. Available: https://github.com/CapsuleEndoscope/EndoSLAM

[3] L. Young, Y. Wu, and A. A. Efros, ”Depth Anything,” 2024. [Online]. Available: https://depth-anything.github.io/