# Accurate Surgical Scene Reconstruction from Multi-view FoundationStereo

Yujia Gong [1, *]

[1] Basis independent McLean, Virginia, USA

* Corresponding Author Email: gggong1110@gmail.com

**Abstract.** Accurate 3D reconstruction of surgical scenes is a critical enabling technology for advancements in intraoperative navigation, surgical training, and robotic automation. While learning-based stereo depth estimation methods have demonstrated high in-domain accuracy, their performance often degrades significantly under domain shifts. This limitation is particularly acute in surgical applications, where large-scale, annotated datasets are scarce. In this work, we investigate the application of FoundationStereo, a recently proposed vision foundation model for stereo matching, to the task of surgical scene reconstruction. We leverage its zero-shot, single-frame depth estimation capabilities within a multi-view fusion framework based on the Truncated Signed Distance Function (TSDF) to achieve comprehensive scene reconstruction. Our experiments, conducted on the public SCARED dataset captured with a da Vinci Xi surgical robot, demonstrate that FoundationStereo achieves state-of-the-art zero-shot accuracy. We report a sub-millimeter mean error for single-frame depth estimation and an error under 2 mm for fused multi-view reconstructions, significantly outperforming the zero-shot STTR baseline. These results highlight the substantial potential of FoundationStereo for enabling accurate, high-fidelity surgical scene reconstruction without domain-specific training. We also discuss current limitations, including the reliance on known camera pose information and challenges in dynamic scenes, and outline future research directions to enhance robustness for clinical applications.

**Keywords:** Computer Vision, 3D Reconstruction, Truncated Signed Distance Function (TSDF), Foundation Stereo, Multiview Stereo.

## 1. Introduction

Minimally invasive surgery (MIS) has become a cornerstone of modern medical practice, offering well-recognized benefits such as reduced postoperative pain, lower infection risk, and faster patient recovery. However, these procedures are conducted through small incisions using endoscopic cameras, which inherently limit the surgeon's field of view and constrain camera mobility. This restricted observability poses significant challenges to both the accuracy and efficiency of surgical procedures, complicating spatial awareness and hand-eye coordination.

Recent advances in 3D reconstruction techniques [1], [2], [3], [4] offer promising solutions to overcome these limitations. By recovering detailed surgical scene geometry, these methods can significantly improve spatial awareness for the surgical team. When integrated with downstream applications in augmented reality (AR) and virtual reality (VR) [5], such reconstructions can substantially enhance intraoperative navigation, surgical training, and clinical decision-making. Furthermore, in the context of robotic surgery, precise 3D geometric information can serve as reliable spatial supervision, facilitating the training and deployment of deep learning-based models for surgical automation [6].

A variety of paradigms for 3D reconstruction have recently been explored. Neural Radiance Field (NeRF)–based methods [3] and 3D Gaussian Splatting (3D GS)–based methods [4] leverage multi-view observations to generate high-quality novel views but typically require extensive, per-scene iterative optimization, rendering them unsuitable for real-time geometry recovery from live surgical video streams. Monocular depth estimation (MDE) [7] offers real-time inference of relative depth maps but inherently lacks of absolute scale information, which is critical for metrically accurate surgical guidance. In contrast, sstereo depth estimation methods reconstruct depth from epipolar correspondences, recovering accurate scale from known camera parameters. This makes stereo

approaches particularly well-suited for robotic surgery, where stereo endoscopes are standard equipment.

Despite their advantages, state-of-the-art learning-based stereo models often fail to generalize to new domains; their performance is highly dependent on the test data closely matching their training distribution. This limitation is particularly pronounced in surgery, where data annotation is a significant bottleneck. Addressing this challenge, recent vision foundation models [8], [9], [10], [11], [12] have demonstrated remarkable cross-domain generalization by training on internet-scale or large, diverse synthetic datasets. Specifically, FoundationStereo [12], a model pre-trained on a massive and varied corpus, shows strong potential for real-time, zero-shot surgical scene reconstruction without requiring any domain-specific fine-tuning.

In this study, we leverage the powerful zero-shot capability of the FoundationStereo model for surgical scene reconstruction. We extend its single-frame depth predictions to a comprehensive multi-view reconstruction by integrating them into a widely adopted Truncated Signed Distance Function (TSDF)[13] fusion pipeline. Our experiments, conducted on the public SCARED dataset [14], validate the effectiveness of this approach. Without any task-specific training, FoundationStereo achieves a mean frame-wise depth error below 1 mm, and the fused multi-view reconstruction maintains an accuracy of sub-2 mm. These results underscore the model's robustness and potential for enabling high-fidelity 3D reconstruction in surgical settings.

In summary, the key contributions of this work are: (1) We apply the FoundationStereo model to surgical scene reconstruction and integrate it with a TSDF-based multi-view fusion pipeline to generate comprehensive 3D models from frame-wise depth maps; (2) We demonstrate sub-millimeter mean error in single-frame depth estimation and sub-2 mm accuracy in the final fused reconstruction, validating the model's potential for high-fidelity surgical applications; and (3) We identify key challenges, particularly in handling dynamic surgical scenes, and propose directions for future research to improve the adaptation and robustness of foundation models in clinical environments.

## 2. Related Work

Three-dimensional (3D) scene understanding from 2D images is a long-standing challenge in computer vision. Research in this area has primarily evolved along three interconnected paths: monocular depth estimation, multi-view stereo, and the recent emergence of vision foundation models.

### 2.1. Monocular Depth Estimation

Monocular depth estimation (MDE) methods [1], [15], [16], [17], [18], [19] aim to infer a dense depth map from a single image by relying on monocular cues such as texture, shading, and object size. Early approaches often employed probabilistic graphical models like Markov Random Fields to enforce spatial smoothness. More recently, the field has been dominated by deep learning architectures, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), which have significantly improved performance.

These supervised models typically formulate depth estimation as a regression problem, learning a direct mapping from image pixels to continuous depth values using large-scale datasets with ground-truth labels. An alternative paradigm treats depth estimation as a classification task, discretizing the depth range into a set of bins [20], [21]. This approach can enhance robustness compared to direct regression. While these methods achieve high accuracy on in-domain test sets, they are notoriously brittle and struggle to generalize to unseen environments due to domain shift. This limitation has motivated recent efforts in zero-shot depth estimation [22], which seek to generate plausible depth maps for novel scenes without task-specific training.

### 2.2. Multi-View Stereo

In contrast to MDE, Multi-View Stereo (MVS) methods reconstruct 3D geometry by leveraging geometric constraints from multiple viewpoints. The core principle of MVS is triangulation, which

maps corresponding points between two or more images to recover their 3D positions. By aggregating information across multiple views, MVS algorithms can produce highly accurate and metrically consistent reconstructions, effectively mitigating issues like occlusion and ambiguity that plague monocular methods.

A wide array of MVS algorithms exists, spanning from purely geometric techniques to sophisticated deep learning-based pipelines and scene-specific optimization models like NeRF or 3D Gaussian Splatting [23], [24], [25], [26], [27], [28], [29], [30], [31], [32]. A key prerequisite for most MVS systems is the availability of calibrated camera parameters, including both intrinsic (e.g., focal length) and extrinsic (i.e., pose) information. However, recent work such as DUST3R [33] has begun to explore uncalibrated reconstruction, relaxing the strict requirement for prior camera knowledge.

### 2.3. Vision Foundation Models for Zero-Shot Depth Estimation

The limitations of traditional supervised models have catalyzed the development of vision foundation models capable of zero-shot depth estimation. These models are pre-trained on massive, diverse datasets, enabling them to generalize across a wide range of domains without task-specific fine-tuning.

Early efforts [34] focused on improving generalization by simply increasing the diversity of training data. A significant milestone was MiDaS [35], which combined multiple heterogeneous datasets and introduced a scale-and-shift invariant loss to ensure consistent relative depth predictions. Subsequent works [36], [37] have further improved performance by incorporating temporal consistency from video data, integrating powerful transformer backbones [38], and leveraging self-supervision [11], [39], [40] to train on unlabeled data. Diffusion-based generative models [41], [42] have also recently been explored for this task.

While these models produce robust relative depth maps, they cannot recover the true metric scale of a scene. This has led to a new frontier: zero-shot metric depth estimation. The goal is to predict absolute depth values without domain-specific training. Early approaches [21], [43] improved global depth calibration, and ZoeDepth [44] introduced indoor/outdoor fine-tuning on top of MiDaS. ScaleDepth [45] eliminated fine-tuning by using a scale estimator after relative depth prediction, while Metric Depth [46], [47] and UniDepth [48] predicted scale directly from input images, with UniDepth also estimating camera intrinsics. Most recently, DepthPro [10] has pushed the boundary further by producing sharp, detailed metric depth maps using a multiscale Vision Transformer that operates without any explicit camera intrinsic parameters.

## 3. Method

In this study, we propose a two-stage reconstruction pipeline, as illustrated in Fig.1. In the first stage, raw input stereo image pairs are rectified using known camera calibration parameters. The FoundationStereo model then processes these images to infer a dense disparity map, which is then converted into a metric depth map using the known stereo baseline. In the second stage, all depth maps are aligned to a common keyframe coordinate system using the provided relative camera poses. These registered depth maps are then integrated into a single volumetric representation using the TSDF [13] fusion algorithm. This step averages multi-view information to produce a complete and noise-resilient 3D reconstruction of the surgical scene.

The following subsections will detail the FoundationStereo model and the TSDF fusion process.

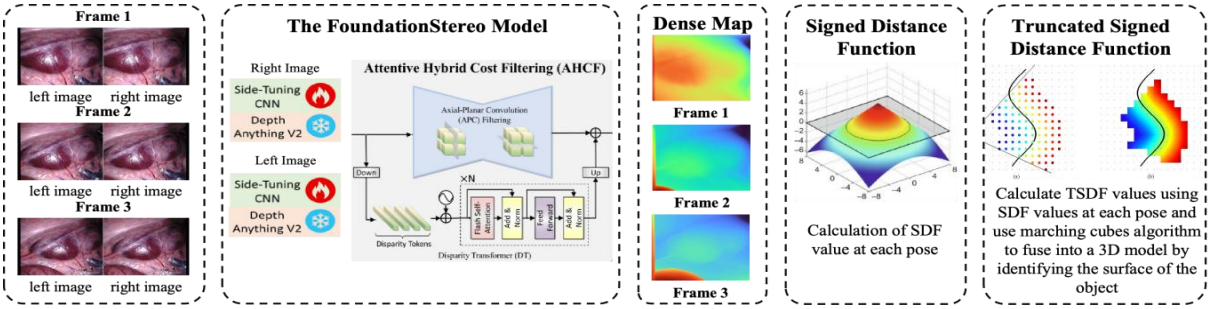



**Figure 1.** The overall structure of the model described in this paper

### 3.1. FoundationStereo

The recently proposed FoundationStereo [12] is currently the state-of-the-art in producing accurate depth maps and point clouds from stereo images, demonstrating strong generalization across datasets from different domains. The model architecture processes stereo image pairs through both the Depth Anything backbone and a side-tuning CNN with shared weights. The extracted features are concatenated and passed through an Attentive Hybrid Cost Filtering (AHCF) module, followed by refinement with GRU blocks. FoundationStereo incorporates three key innovations: (1) a training pipeline that selectively improves or removes samples to enhance cross-domain generalization, (2) a Side-Tuning Adapter (STA) for efficient feature adaptation, and (3) an AHCF module that integrates 3D Axial Planar Convolution (APC) and a Disparity Transformer (DT) for robust cost aggregation.

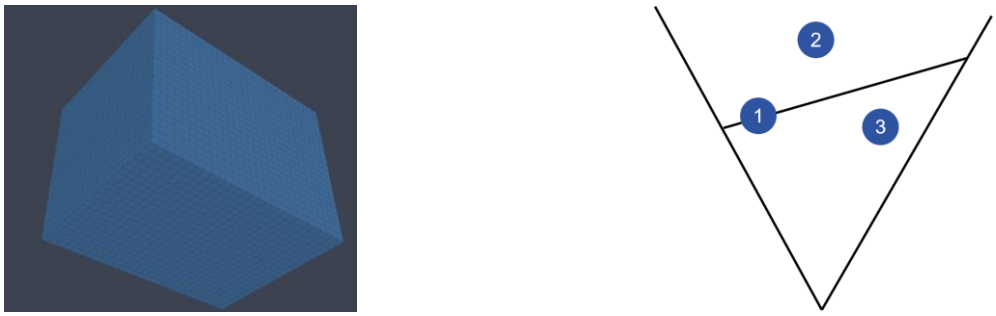
Due to the monotonic nature of many scenes, the similarity of intrinsic parameters, and the limited realism of existing synthetic datasets, FoundationStereo does not restrict its training to publicly available synthetic data. Instead, it employs a novel data generation pipeline to enhance diversity, incorporating a wide range of scene types—both chaotic and structured—with randomized lighting, baseline, and focal length. This variability improves the model's ability to generalize to novel scenarios. However, because most data are generated procedurally, some scenes may contain unrealistic object configurations, which can hinder effective learning. To address this, FoundationStereo applies iterative self-curation to remove ambiguous samples, defined as those in which at least 60% of pixels have a disparity error greater than two pixels. This curation–evaluation process is performed twice to ensure the final training set is both diverse and of high quality, thereby improving generalization capability. The enhanced generalization enabled by this data generation pipeline makes FoundationStereo a strong candidate for surgical scene reconstruction, where domain-specific training data are scarce.

The Side-Tuning Adapter (STA) in FoundationStereo is designed to address the domain gap between simulation and real-world data. In initial experiments, the model was tested using only features from the Depth Anything V2 backbone and, separately, using an adapter to combine features from both a Vision Transformer (ViT) and a CNN. The most effective configuration was achieved by first downscaling features with a $4 \times 4$ convolution before feeding them into the Depth Anything model, then concatenating them with CNN features of the same spatial resolution. This approach successfully leverages the complementary strengths of ViTs and CNNs for richer feature representation. While Depth Anything V2 [49] suffers from ambiguity in the metric scale of its predicted depth maps, its extracted features still encode abundant geometric information that is highly beneficial for depth estimation.

The 3D Axial Planar Convolution (APC) decomposes a standard 3D convolution into separate operations along the spatial and disparity dimensions, enabling more efficient feature processing. The Disparity Transformer (DT) applies attention solely along the disparity dimension, reducing computational overhead while enhancing cost volume filtering performance. In the overall architecture, a GRU-based refinement module further optimizes the disparity predictions through three sequential stages, each building on the disparity and feature representations from the previous stage. Context features extracted by the STA module are incorporated into the refinement process, providing useful priors that guide the optimization toward more accurate depth estimates.

### 3.2. TSDF

The Truncated Signed Distance Function (TSDF) [13] represents a 3D scene as a volumetric grid, where each voxel stores the signed distance to the nearest surface, truncated to a fixed range and normalized by a constant factor. As depth images from multiple viewpoints are integrated, each voxel is projected into the corresponding camera frames to compute the difference between its depth and the observed surface depth along the viewing ray as shown in Figure 2. This value is then fused with previous estimates using a weighted average. A positive distance indicates that a voxel lies in front of the surface, and a negative distance indicates that it's behind, and a near-to-zero value corresponds to a point that's either on or very close to the surface. By accumulating measurements from different views, TSDF produces a smooth and consistent 3D surface reconstruction.




(a) (b)

**Figure 2.** Figure (a) illustrates the view of an endoscope and a rectangular box of voxels. Figure (b) illustrates the sign of the SDF value. Voxel 1 represents an SDF value of near 0, voxel 2 shows an SDF value of smaller than 0, and voxel 3 has an SDF value bigger than 0.

The first step in TSDF-based reconstruction is to discretize the scene into voxels. Using the known camera intrinsics, we determine the scale and orientation of the axes for the bounding box that encompasses the entire scene. This box is then uniformly divided into cubic regions (voxels) of equal size. The voxel resolution is bounded below by the desired mesh accuracy and above by the computational resources available, as higher resolutions require significantly more memory and processing to compute SDF and TSDF values in real time. Because a single viewpoint cannot fully capture the geometry of a scene—due to occlusions and perspective limitations—information from multiple viewpoints must be aggregated to obtain a complete and detailed reconstruction.

The second step is to assign each voxel a value based on the Truncated Signed Distance Function (TSDF). Once the bounding volume is divided into uniform voxels, we first establish a transformation from voxel coordinates to world coordinates, enabling each voxel's position to be expressed in the global reference frame. This transformation is defined in Equation (1).

$$P_x = \left(x_0 + voxel_x * v_x, y_0 + voxel_y * v_y, z_0 + voxel_z * v_z\right) \tag{1}$$

The $(x_0, y_0, z_0)$ in this equation represents the position of the voxel $(0,0,0)$ in world coordinates, and $voxel_x, voxel_y, voxel_z$ represent the size of the different dimensions for a single voxel. $v_x, v_y, v_z$ are the coordinates of the voxel in the voxel system. Next, we transform the voxel coordinates from the world coordinate system into each camera's coordinate frame, as defined in Equation (2). In this frame, the z-value corresponds to the depth of the voxel relative to that camera view.

$$p_x = P_x R + T \tag{2}$$

In this equation, the $R$ and $T$ are the extrinsics for the endoscope, and $p_x$ is the coordinate of the voxel from the camera's perspective. Finally, we would be able to project each voxel to the image plane through the known camera intrinsics by Equation (3).

$$dep_y(x) pos_x = K p_x \tag{3}$$

The $dep_y(x)$ describes the depth of the voxel $x$ under the view of the $y$-th view of the endoscope, $pos_x$ describes the pixel coordinates of the voxel under the projection of the $y$-th view, and $K$ is the intrinsic matrix of the endoscope.

The signed distance function (SDF) value of a voxel is computed as the difference between the voxel's depth and the depth of its corresponding projected pixel in the image. This value represents the distance from the voxel center to the surface intersection along the ray from the camera origin through the voxel center. The truncated signed distance function (TSDF) is then obtained by normalizing the SDF value with a fixed constant and clamping it to the range [-1, 1]. This truncation not only limits the value range but also mitigates inconsistencies in relative scale caused by varying camera positions.

The third step is to fuse the TSDF values computed from multiple viewpoints into a single unified value for each voxel. This is achieved using a weighted average of the individual TSDF measurements. The weight for each view is defined as the cosine of the angle between the ray from the camera origin to the voxel center and the surface normal vector at the corresponding point in the camera frame. Incorporating these view-dependent weights in the fusion process yields a more accurate TSDF estimate, indicating whether the voxel lies on the object surface and enabling consistent 3D reconstruction. This process is formalized in Equation (4).

$$TSDF(x) = \frac{tsdf_1(x)w_1(x)+tsdf_2(x)w_2(x)+\ldots+tsdf_y(x)w_y(x)}{w_1(x)+w_2(x)+\ldots+w_y(x)} \tag{4}$$

where $x$ is the voxel, $TSDF$ is the final value of one voxel considering all perspectives, $tsdf$ and $w$ are the TSDF values and weights under each camera's perspective.

The final step is to extract and represent the 3D surface from the fused TSDF volume. This is accomplished by identifying surface regions where the TSDF values are close to zero, indicating the zero-crossing between free space and occupied space. RGB information from multiple camera views is also projected and integrated to assign consistent color values to each voxel. Surface extraction is performed after using the Marching Cubes Algorithm, which generates a mesh by contouring the surface within each voxel cell, resulting in a detailed 3D reconstruction of the scene.

The Marching Cubes algorithm focuses on identifying how a surface intersects a voxel without requiring an explicit analytical representation of the surface. The procedure begins by selecting a voxel and evaluating the scalar field values at its eight corner vertices to determine whether each lies inside or outside the surface. By enumerating all possible combinations of these binary states, the algorithm classifies the voxel into one of 256 distinct topological configurations. To facilitate efficient lookup and interpolation, these vertex states are encoded as an 8-bit binary index, where each bit represents the state (inside or outside) of a corresponding vertex. This binary index maps to a precomputed case in a lookup table that defines which voxel edges are intersected by the surface. Finally, linear interpolation is applied along these intersected edges to estimate the precise locations of the surface–voxel boundary, enabling smooth and consistent mesh generation across the volume.

## 4. Experiments

In this section, we present the primary experiment designed to evaluate the effectiveness of FoundationStereo and the comprehensive reconstruction obtained via TSDF fusion. The following subsections describe the SCARED dataset, implementation details, and the evaluation protocol. We then report the results of the main experiment, followed by a discussion of the strengths and limitations of FoundationStereo in the context of surgical scene reconstruction.

### 4.1. Dataset

We evaluate our approach on the SCARED dataset [14], which is released for the 2019 Stereo Correspondence and Reconstruction of Endoscopic Scenes challenge. The dataset contains stereo video sequences acquired with a da Vinci Xi surgical robot, closely replicating real surgical

conditions. It provides high-resolution imagery, accurate ground-truth depth maps, and full camera calibration, including intrinsic and extrinsic parameters for both left and right endoscopes, as well as per-frame camera poses. Each sequence is organized into key frames, with associated stereo videos in which only the endoscope moves while the scene remains static. All poses are defined relative to the key frame, enabling consistent multi-view alignment and simplifying TSDF-based fusion. The high accuracy of calibration and ground-truth depth maps make SCARED an ideal benchmark for evaluating both frame-wise depth estimation and comprehensive 3D reconstruction. We take three sequences as the test set in our study.

### 4.2. Implementation Details

We adapted the publicly available FoundationStereo codebase, making minor modifications to account for the specific camera intrinsics and the baseline distance between the stereo endoscopes. Images containing an alpha channel were preprocessed to remove the channel, as it is irrelevant for depth estimation. Image rectification was performed using the official SCARED toolkit, which also provides ground-truth depth maps after rectification for evaluation.

For TSDF fusion, we used the rectified left-view images, the corresponding depth maps predicted by FoundationStereo, and the associated camera poses. To ensure accurate and well-aligned 3D meshes, the scale of both the predicted depth maps and camera poses was adjusted for consistency. In TSDF integration, the voxel size was set to 0.22mm, as smaller voxel sizes would exceed our available computational resources.

### 4.3. Evaluation

#### 4.3.1. Protocol

To evaluate the single-frame performance of FoundationStereo, we compare its predicted depth maps directly to those rendered from the ground-truth point cloud. Error computation is restricted to valid pixels, excluding locations where either the prediction or ground truth is missing. For assessing the accuracy of the FoundationStereo-based comprehensive reconstruction, we render depth maps from the fused TSDF representation for each view and compare them to the corresponding rendered ground-truth depth maps.

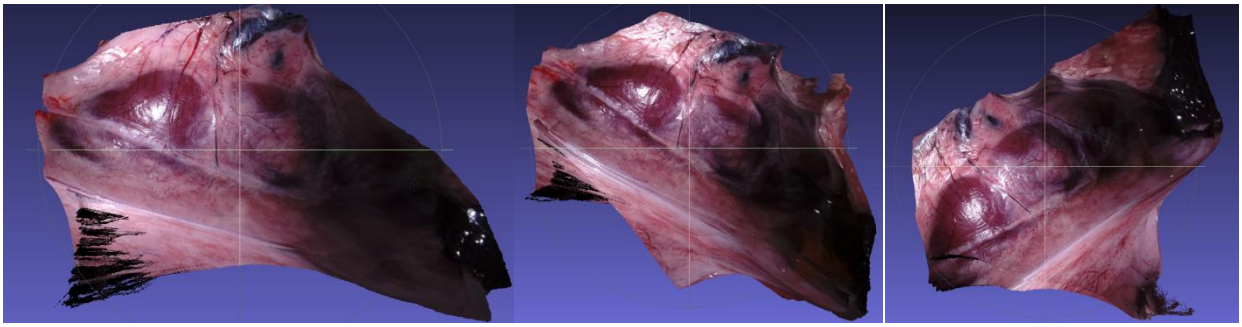
#### 4.3.2. Metric

We evaluate our model's performance by computing the absolute $L_1$ error between the predicted and ground-truth depth maps. The $L_1$ error is defined as the sum of the absolute depth differences across all pixels:

$$L1 = \sum_x |dep_x - real_x|, \tag{5}$$

where $x$ indices each pixel, $dep_x$ is the estimated depth from the endoscope origin, and $real_x$ is the corresponding ground-truth depth.

### 4.4. Main results



(a) (b) (c)

**Figure 3.** The 3 figures show the results of the final mesh with 150 frames being fused in each case

The main experimental results are presented in Table 1. For single-frame, FoundationStereo achieves an error of less than 1mm across all three videos, outperforming the zero-shot STTR [2], which reports an average error of 3 mm as shown in Figure 3. This demonstrates the state-of-the-art zero-shot performance on the SCARED dataset, validating the effectiveness of FoundationStereo for frame-level surgical scene reconstruction. When aggregating multi-view predictions into a comprehensive reconstruction, the proposed pipeline can still achieve a sub-2mm error when fused with 70 frames. Although the comprehensive reconstructions also show very high accuracy, they show an increased error compared to the separate frame-wise depth prediction.

**Table 1.** Main Results. fs means fused with the depth map predicted from the FoundationStereo under a 0.22mm voxel size. gt means fused with ground truth depthmap under a 0.35mm voxel size due to memory limit.

| keyframe# | average L1 error Single-frame | average L1 error Comprehensive 10 frames | average L1 error Comprehensive 30 frames | average L1 error Comprehensive 70 frames | average L1 error Comprehensive 150 frames |
|---|---|---|---|---|---|
| fs_1 | 0.7851mm | 1.1421mm | 1.2584mm | 1.4258mm | 1.7280mm |
| fs_2 | 1.1459mm | 1.6028mm | 1.7269mm | 2.0974mm | 2.3920mm |
| fs_3 | 1.0868mm | 1.8811mm | 1.8975mm | 1.9368mm | 2.1800mm |
| gt_1 | - | 1.6543mm | 2.0690mm | 2.3602mm | 2.6178mm |
| gt-2 | - | 1.6809mm | 2.0366mm | 2.4964mm | 3.0487mm |
| gt-3 | - | 1.5683mm | 1.6947mm | 1.7760mm | 2.4209mm |

### 4.5. Ablation Study

We conduct ablation experiments to investigate the sources of error in multi-view fusion. First, we fuse the ground-truth depth maps into a comprehensive reconstruction and calculate the corresponding frame-wise error. As shown in Table 1, non-negligible errors remain even when using perfect single-frame depth maps, suggesting that voxelization and inevitable inaccuracies in camera pose are primary contributors to error accumulation. Next, we examine the effect of voxelization by varying voxel sizes. The results in Table 2 demonstrate that voxel size has a significant impact on reconstruction accuracy, confirming voxelization as a major error source in multi-view fusion. Then, we explored the influence of depth prediction error. Frames are sorted by their frame-wise L_1 error, and 10 representative frames from different error ranges are selected for fusion. As shown in Table 3, the accuracy of the comprehensive reconstruction is only marginally affected by differences in frame-wise error. We attribute this to the relatively small variation in single-frame accuracy and the tendency of multi-view fusion to cancel out frame-level errors. Finally, we explore the influence of camera pose variance. Frames are sorted by their distance to the reference frame. We sample 10 representative frames with different step sizes to fuse frames with different pose variance. As shown in Table 4, the accuracy of the comprehensive reconstruction is not affected by differences in the variance of the camera pose. This indicates that the noise in the camera pose doesn't contribute significantly to the additional error introduced by the fusion.

**Table 2.** Results with varied voxel size.

| keyframe # | Voxel size | average L1 error Comprehensive 10 frames | average L1 error Comprehensive 30 frames | average L1 error Comprehensive 70 frames | average L1 error Comprehensive 150 frames |
|---|---|---|---|---|---|
| 1 | 0.20mm | 1.1421mm | 1.2584mm | 1.4258mm | 1.7280mm |
| | 0.35mm | 1.5006mm | 1.6260mm | 1.8342mm | 2.1615mm |
| | 0.50mm | 1.8779mm | 2.0192mm | 2.2346mm | 2.5892mm |
| 2 | 0.20mm | 1.6028mm | 1.7269mm | 2.0974mm | 2.3920mm |
| | 0.35mm | 1.9561mm | 2.1240mm | 2.5202mm | 2.7782mm |
| | 0.50mm | 2.3625mm | 2.5550mm | 2.9622mm | 3.1952mm |
| 3 | 0.20mm | 1.8811mm | 1.8975mm | 1.9368mm | 2.1800mm |
| | 0.35mm | 2.2900mm | 2.3194mm | 2.3559mm | 2.6250mm |
| | 0.50mm | 2.7201mm | 2.7484mm | 2.7941mm | 3.0517mm |

**Table 3.** Results of comprehensive reconstruction in different error levels.

| keyframe # | 0-9 | 20-29 | 40-49 | 60-69 | 140-149 |
|---|---|---|---|---|---|
| 1 | 1.3685mm | 1.4063mm | 1.4994mm | 1.4781mm | 1.1231mm |
| 2 | 1.6464mm | 1.7641mm | 1.9146mm | 1.8757mm | 2.2714mm |
| 3 | 1.4342mm | 1.3679mm | 1.3862mm | 1.3664mm | 1.4886mm |

**Table 4.** Results of comprehensive reconstruction in different camera variance.

| step size# | 1 | 3 | 6 | 10 | 15 |
|---|---|---|---|---|---|
| 1 | 1.1421mm | 1.1783mm | 1.1651mm | 1.1673mm | 1.2617mm |
| 2 | 1.6028mm | 1.6513mm | 1.8910mm | 1.9894mm | 2.0247mm |
| 3 | 1.8811mm | 1.8798mm | 1.9044mm | 1.9008mm | 2.0042mm |

In summary, the ablation studies indicate that the error sources are mainly from the less fine-grained voxelization due to limited memory resources. Thus, with better computational resources, the comprehensive reconstruction can be significantly improved with the current pipeline.

## 5. Discussion

The primary strength revealed by our experiments is the exceptional zero-shot generalization capability of FoundationStereo. At the single-frame level, the model achieved a sub-millimeter mean error across the SCARED dataset without any domain-specific fine-tuning. This performance not only surpasses that of the zero-shot STTR baseline [2] but also confirms that large-scale, diverse pre-training can effectively overcome the domain shift inherent in specialized medical imaging. The ability to maintain reconstruction errors below 2 mm after multi-view fusion further underscores the consistency and reliability of the model's depth predictions. These results strongly suggest that foundation models are a viable pathway to developing accurate perception systems in data-scarce surgical environments, reducing the dependency on costly and difficult-to-acquire annotated data.

For downstream applications such as surgical navigation and autonomous surgery, scene complexity introduces additional challenges. Surgical environments often present specular reflections, texture sparsity, and occlusions, which can generate localized depth errors that disproportionately affect volumetric reconstructions. Moreover, our study was conducted under static-scene assumptions, whereas real surgical settings are inherently dynamic. Organ motion—such as the rhythmic deformation of the heart or subtle shifts of surrounding tissue—can introduce temporal inconsistencies that degrade multi-view fusion quality. Another limitation lies in the dependency of the TSDF-based fusion pipeline on precise endoscope pose information, available in the SCARED dataset but not always accessible in practice. Without reliable extrinsic calibration, reconstructions risk becoming misaligned or incomplete, limiting applicability in workflows lacking integrated pose tracking.

Overcoming these challenges will require confidence-aware fusion strategies, such as uncertainty estimation or adaptive TSDF weighting, to mitigate outlier effects. Incorporating pose estimation directly into the pipeline or exploring calibration-independent fusion methods could further improve robustness. Finally, the integration of temporal and deformable tissue modeling will be essential for extending this approach to real-time surgical applications, representing an important direction for future research.

## 6. Conclusion

In summary, we applied FoundationStereo for surgical scene reconstruction, combining zero-shot depth estimation with TSDF-based multi-view fusion. Our experimental evaluation on the SCARED dataset validates the efficacy of this approach. Without any domain-specific fine-tuning, FoundationStereo achieved state-of-the-art zero-shot performance, significantly outperforming prior

baselines: 1) a mean error of less than 1 mm for single-frame depth estimation and 2) a cumulative error of under 2 mm for the final, fused multi-view reconstruction.

Despite the promising results, several limitations must be addressed to ensure clinical viability. The accuracy of our pipeline currently depends on precise camera pose information from the robotic system and is susceptible to voxelization artifacts inherent in the TSDF fusion process. Future research will be directed at overcoming these challenges. We plan to investigate uncertainty-aware fusion techniques to enhance robustness against noisy depth predictions and pose inaccuracies. Furthermore, integrating methods for deformable and dynamic scene modeling will be essential for adapting this framework to the complexities of live intraoperative guidance. Addressing these areas will be a critical step toward the robust deployment of next-generation 3D perception systems in surgical robotics.

## References


[1] 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, pp. 2024–2039, Oct. 2016, doi: 10.1109/tpami.2015.2505283.

[2] Z. Li *et al.*, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6197–6206.

[3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis." 2020. Available: https://arxiv.org/abs/2003.08934

[4] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D gaussian splatting for real-time radiance field rendering." 2023. Available: https://arxiv.org/abs/2308.04079

[5] B. D. Killeen *et al.*, "Stand in surgeon's shoes: Virtual reality cross-training to enhance teamwork in surgery," *International journal of computer assisted radiology and surgery*, vol. 19, no. 6, pp. 1213–1222, 2024.

[6] H. Ding *et al.*, "Towards robust automation of surgical systems via digital twin-based scene representations from foundation models," *arXiv preprint arXiv:2409.13107*, 2024.

[7] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[8] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[9] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht, "Cotracker: It is better to track together," in *European conference on computer vision*, Springer, 2024, pp. 18–35.

[10] A. Bochkovskii *et al.*, "Depth pro: Sharp monocular metric depth in less than a second." 2025. Available: https://arxiv.org/abs/2410.02073

[11] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data." 2024. Available: https://arxiv.org/abs/2401.10891

[12] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "FoundationStereo: Zero-shot stereo matching." 2025. Available: https://arxiv.org/abs/2501.09898

[13] M. Grinvald, F. Tombari, R. Siegwart, and J. Nieto, "TSDF++: A multi-object formulation for dynamic object tracking and reconstruction." 2021. Available: https://arxiv.org/abs/2105.07468

[14] M. Allan *et al.*, "Stereo correspondence and reconstruction of endoscopic data challenge." 2021. Available: https://arxiv.org/abs/2101.01133

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

[16] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks." 2016. Available: https://arxiv.org/abs/1606.00373

[17] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision transformers for dense prediction.” 2021. Available: https://arxiv.org/abs/2103.13413

[18] A. Saxena, S. Chung, and A. Ng, “Learning depth from single monocular images,” in *Advances in neural information processing systems*, Y. Weiss, B. Schölkopf, and J. Platt, Eds., MIT Press, 2005. Available: https://proceedings.neurips.cc/paper_files/paper/2005/file/17d8da815fa21c57af9829fb0a869602-Paper.pdf

[19] A. Saxena, M. Sun, and A. Y. Ng, “Make3D: Learning 3D scene structure from a single still image,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, 2009, doi: 10.1109/TPAMI.2008.132.

[20] S. F. Bhat, I. Alhashim, and P. Wonka, “LocalBins: Improving depth estimation by learning local distributions.” 2022. Available: https://arxiv.org/abs/2203.15132

[21] S. Farooq Bhat, I. Alhashim, and P. Wonka, “AdaBins: Depth estimation using adaptive bins,” in *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, IEEE, Jun. 2021, pp. 4008–4017. doi: 10.1109/cvpr46437.2021.00400.

[22] Y. Wang, Z. Pan, X. Li, Z. Cao, K. Xian, and J. Zhang, “Less is more: Consistent video depth estimation with masked frames modeling,” in *Proceedings of the 30th ACM international conference on multimedia*, in MM ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 6347–6358. doi: 10.1145/3503161.3547978.

[23] J. Choe, S. Im, F. Rameau, M. Kang, and I. S. Kweon, “VolumeFusion: Deep depth fusion for 3D scene reconstruction.” 2021. Available: https://arxiv.org/abs/2108.08623

[24] Y. Furukawa and C. Hernández, “Multi-view stereo: A tutorial,” *Foundations and Trends® in Computer Graphics and Vision*, vol. 9, no. 1–2, pp. 1–148, 2015, doi: 10.1561/0600000052.

[25] Q. Fu, Q. Xu, Y.-S. Ong, and W. Tao, “Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction.” 2022. Available: https://arxiv.org/abs/2205.15848

[26] M. Oechsle, S. Peng, and A. Geiger, “UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction.” 2021. Available: https://arxiv.org/abs/2104.10078

[27] M. Oquab *et al.*, “DINOv2: Learning robust visual features without supervision.” 2024. Available: https://arxiv.org/abs/2304.07193

[28] Y. Wang, I. Skorokhodov, and P. Wonka, “HF-NeuS: Improved surface reconstruction using high-frequency details.” 2022. Available: https://arxiv.org/abs/2206.07850

[29] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, “NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction.” 2023. Available: https://arxiv.org/abs/2106.10689

[30] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, “MVSNet: Depth inference for unstructured multi-view stereo.” 2018. Available: https://arxiv.org/abs/1804.02505

[31] X. Ye, W. Zhao, T. Liu, Z. Huang, Z. Cao, and X. Li, “Constraining depth map geometry for multi-view stereo: A dual-depth approach with saddle-shaped depth cells.” 2023. Available: https://arxiv.org/abs/2307.09160

[32] C. Zhao, Y. Ge, F. Zhu, R. Zhao, H. Li, and M. Salzmann, “Progressive correspondence pruning by consensus learning.” 2021. Available: https://arxiv.org/abs/2101.00591

[33] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, “DUSt3R: Geometric 3D vision made easy.” 2024. Available: https://arxiv.org/abs/2312.14132

[34] Z. Li and N. Snavely, “MegaDepth: Learning single-view depth prediction from internet photos,” in *2018 IEEE/CVF conference on computer vision and pattern recognition*, 2018, pp. 2041–2050. doi: 10.1109/CVPR.2018.00218.

[35] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[36] S. Li, Y. Luo, Y. Zhu, X. Zhao, Y. Li, and Y. Shan, “Enforcing temporal consistency in video depth estimation,” in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV) workshops*, 2021, pp. 1145–1154.

[37] H. Zhang, C. Shen, Y. Li, Y. Cao, Y. Liu, and Y. Yan, “Exploiting temporal consistency for real-time video depth estimation.” 2019. Available: https://arxiv.org/abs/1908.03706

[38] R. Birkl, D. Wofk, and M. Müller, “MiDaS v3.1 – a model zoo for robust monocular relative depth estimation.” 2023. Available: https://arxiv.org/abs/2307.14460

[39] A. Petrovai and S. Nedevschi, “Exploiting pseudo labels in a self-supervised learning framework for improved monocular depth estimation,” in *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 1568–1578. doi: 10.1109/CVPR52688.2022.00163.

[40] J. Spencer, C. Russell, S. Hadfield, and R. Bowden, “Kick back & relax: Learning to reconstruct the world by watching SlowTV.” 2023. Available: https://arxiv.org/abs/2307.10713

[41] M. Gui *et al.*, “DepthFM: Fast monocular depth estimation with flow matching.” 2024. Available: https://arxiv.org/abs/2403.13788

[42] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, “Repurposing diffusion-based image generators for monocular depth estimation.” 2024. Available: https://arxiv.org/abs/2312.02145

[43] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep ordinal regression network for monocular depth estimation.” 2018. Available: https://arxiv.org/abs/1806.02446

[44] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, “ZoeDepth: Zero-shot transfer by combining relative and metric depth.” 2023. Available: https://arxiv.org/abs/2302.12288

[45] R. Zhu, C. Wang, Z. Song, L. Liu, T. Zhang, and Y. Zhang, “ScaleDepth: Decomposing metric depth estimation into scale prediction and relative depth estimation.” 2024. Available: https://arxiv.org/abs/2407.08187

[46] M. Hu *et al.*, “Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10579–10596, Dec. 2024, doi: 10.1109/tpami.2024.3444912.

[47] W. Yin *et al.*, “Metric3D: Towards zero-shot metric 3D prediction from a single image.” 2023. Available: https://arxiv.org/abs/2307.10984

[48] L. Piccinelli *et al.*, “UniDepth: Universal monocular metric depth estimation.” 2024. Available: https://arxiv.org/abs/2403.18913

[49] L. Yang *et al.*, “Depth anything V2.” 2024. Available: https://arxiv.org/abs/2406.09414