Real2Sim2Real: RetinalDepth for Depth Estimation in Posterior Segment Ophthalmic Surgery

Supplementary Material

1. Details of RetinalDepth

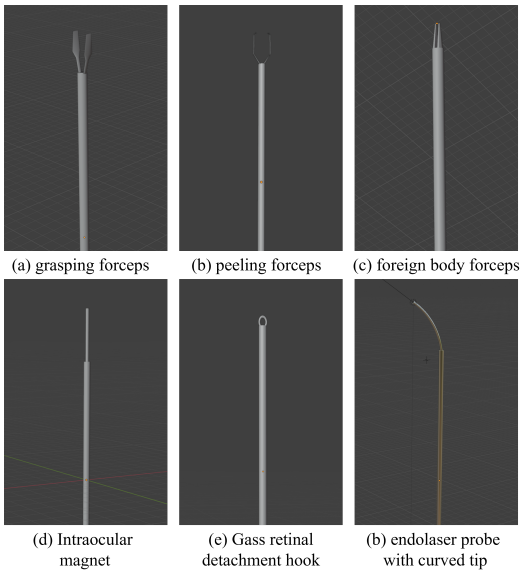


Figure 1. Surgical instruments modeled for RetinalDepth. All tools feature high-reflectivity metallic materials and, where applicable, articulated mechanisms: (a) grasping forceps, (b) peeling forceps with open/close capability, (c) foreign body forceps, (d) intraocular magnet, (e) gas retinal detachment hook, and (f) curved-tip endolaser probe.

1.1. Dataset Diversity and Composition

RetinalDepth is engineered to maximize intra-domain variation while maintaining anatomical and surgical realism. Diversity is introduced at multiple levels during scene construction:

- **Retinal textures**: 10 distinct ultra-wide-field fundus photographs from real patients with varying pathology are mapped onto the inner surface of the spherical retina. Macular and optic disc positions are aligned according to intrinsic retinal texture cues, yielding anatomically plausible topographic variations, including the characteristic foveal depression and optic disc elevation.
- **Aqueous**: Refractive index (from 1.334 to 1.342), absorption color, and surface roughness of the vitreous sphere are randomized per scene to replicate aging, dye usage, or hemorrhage.
- **Surgical instruments**: 6 instrument classes as shown in Fig. 1 appear either singly or in plausible combinations. Articulated tools support continuous opening/closing angles.
- **Instrument trajectories**: The 50-frame sequences simulate surgeon-like micro-movements, including slow translation, tremor, and grasping or peeling actions, all confined to the posterior pole. This ensures that the instruments stay close to or in contact with the retina, without unrealistic penetration.
- **Illumination**: 1–2 endo-illumination probes are placed in varying quadrants such as upper-leftand upper-right with randomized intensity and subtle frame-to-frame flicker. An external point light simulates corneal reflection glare.
- **Camera motion**: Small synchronous sway and shake are applied to the stereo rig to mimic retinal drag induced by instrument contact.

Fig. 2 shows representative frames from 50 different scenes, highlighting the resulting visual heterogeneity. Through systematic combinatorial sampling of the above factors, RetinalDepth comprises 896 independent surgical scenes, each containing a 50-frame stereo video sequence, yielding 44,800 synchronized stereo pairs in total 89,600 RGB images with pixel-perfect ground truth.

This level of controlled yet rich variation ensures that models trained or fine-tuned on RetinalDepth encounter a wide distribution of fundus appearances, lighting conditions, instrument configurations, and dynamic interactions—critical for robust generalization to unseen real posterior segment procedures.

1.2. 3D Visualization of RetinalDepth

To illustrate the construction and utility of RetinalDepth, we present two key visualizations. Fig. 3 captured from the Blender scene showcases the dataset's creation process, featuring components such as the aqueous humor, dual cameras, eyelid, fundus, internal and external light sources including light in, light on and light out, a mask, and a stick representing surgical tools. This holistic view highlights the integration of anatomical and environmental elements in the synthetic environment.

Fig. 4 displays a point cloud derived from the depth data, overlaid with the corresponding left and right camera views and their RGB images, demonstrating the dataset's 3D reconstruction capability. Together, these visualizations underscore the dataset's effectiveness in supporting advanced depth estimation and surgical simulation tasks.

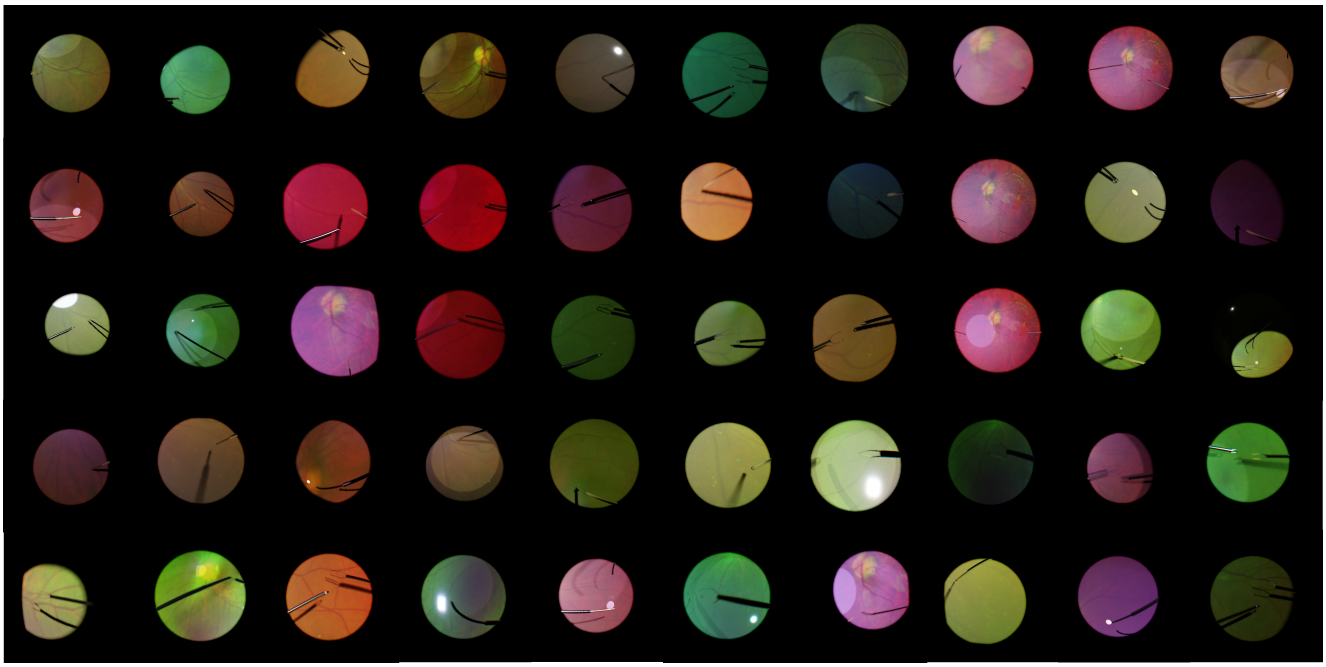
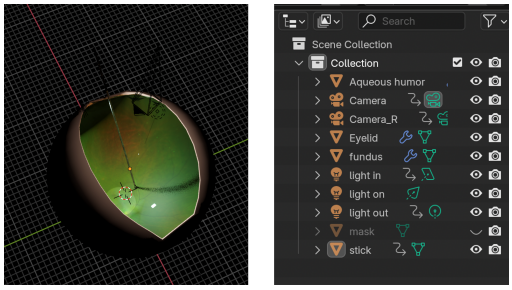


Figure 2. Examples of scene diversity in RetinalDepth are provided. The selection demonstrates variations in retinal pathology and coloration, instrument type and pose, tool-retina interaction, illumination direction and intensity, as well as aqueous humor tint. In total, 896 unique surgical scenarios are created.



(a) Blender Scene (b) Scene Collection

Figure 3. An Example of Blender Scene

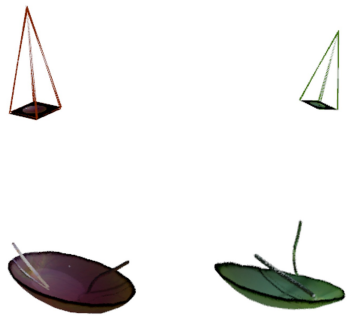


Figure 4. Point Clouds Derived from the Depth Data

2. Experiments Setting Details

2.1. Dataset Division

For the single-image depth estimation experiments, we prepare the data by randomly selecting 2–3 frames per scene to minimize redundancy while ensuring that no scene appears in multiple splits. This procedure yields an approximately 80 percent training set with 797 pairs, a 10 percent validation set with 100 pairs, and a 10 percent test set with 100 pairs. Scenes are uniquely assigned to each split to maintain independence and to ensure a balanced distribution across instrument types and scene variations. For fine-tuning in the single-image setting, models are initialized with pre-trained weights.

For the video depth estimation experiments, we divide the 896 scenes into an 80 percent training set with 716 scenes, a 10 percent validation set with 89 scenes, and a 10 percent test set with 91 scenes. Each scene retains all 50 frames in order to preserve temporal information. Because the evaluation focuses on zero-shot testing for video-specific methods, no additional fine-tuning is applied to these methods. Instead, they are evaluated directly alongside the models previously fine-tuned in the single-image setting, and performance is assessed on the full temporal sequences. Further details are provided in the supplementary material.

2.2. Training Details

We fine-tune the models on the training set for 20 epochs using the AdamW optimizer with an initial learning rate of 1e-4, which is decayed to 1e-6 after 20 epochs. The batch size is set to 4 or 8, depending on the model's memory requirements, and is performed on an NVIDIA RTX A6000 GPU (48GB VRAM). Loss functions adhere to the model-specific defaults. Zero-shot evaluations employ the pre-trained models without any fine-tuning on our dataset.

2.3. Evaluation Metrics

For single-image depth estimation of both monocular and stereo methods, we report three accuracy thresholds: $\delta < 1.25^{0.5}$, $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$ (higher is better), which measure the percentage of pixels where the ratio between predicted and ground-truth depth is below the given factor. We also include six error metrics: Absolute Relative Error (AbsRel), Squared Relative Error (SqRel), Root Mean Squared Error (RMSE), Root Mean Squared Logarithmic Error (RMSELog), Log10 Error, and Scale-Invariant Logarithmic Error (SIlog) (lower is better). Monocular predictions are aligned with ground truth via min-max normalization prior to metric computation. All standard metrics are computed in this clipped space.

3. More Qualitative Experiment results

In the main paper, Fig. 4 and Fig. 5 show qualitative results obtained with the foundation model after fine-tuning on RetinalDepth. To better appreciate the impact of this RetinalDepth fine-tuning stage, we present here direct side-by-side comparisons with the same foundation model *without* any fine-tuning on our synthetic dataset and realistic images. The additional figures in this appendix clearly demonstrate that fine-tuning on RetinalDepth yields noticeably sharper depth maps, more accurate instrument localization, and fewer estimation artifacts on both synthetic and real posterior segment surgical sequences.

Notice that the experiments are divided into single-image depth estimation and video depth estimation. We will analyze the qualitative results seperatively as well.

3.1. Comparison of Single-image Depth Estimation

Marigold [10]. As shown in Fig. 5, the original Marigold model without fine-tuning on RetinalDepth exhibits significant limitations on synthetic posterior segment scenes. Although it can sometimes detect the presence of surgical instruments, it typically fails to capture any depth variation along the instrument shift, treating the entire instrument as a flat, constant-depth object. For instance, only the edges of images in columns 3 and 4 are roughly recognized. Moreover, when the instrument has low contrast with the retinal background, the model often fails to detect the instrument and predicts an almost uniform depth across the entire scene, thereby losing the natural spherical curvature of the retina.

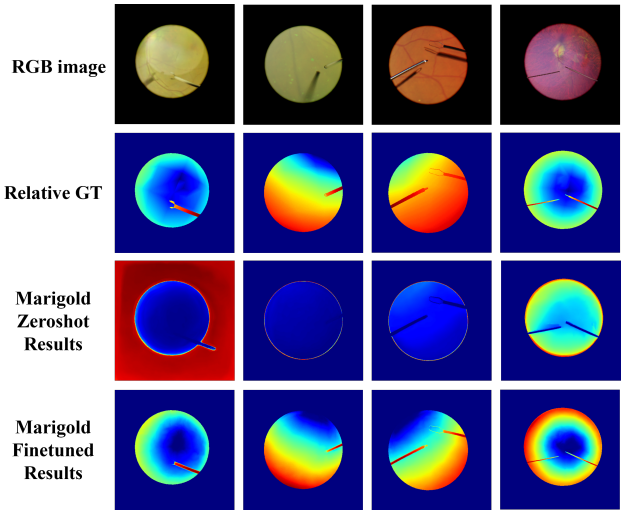


Figure 5. Inference Results on **Synthetic** RetinalDepth Images Before and After Fine-tuning of **Marigold**

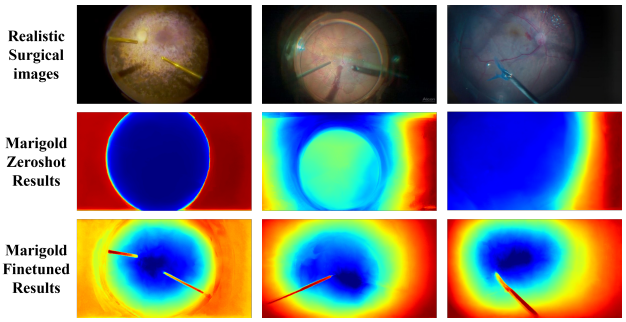


Figure 6. Inference Results on **Realistic** Surgical Images Before and After Fine-tuning of **Marigold**

In contrast, after fine-tuning on RetinalDepth, Marigold produces markedly richer and more accurate depth maps. The instrument shafts now exhibit clear depth gradients, particularly visible in the fourth column, and the retinal background correctly recovers the expected spherical shape of the eye globe with smooth depth variation from the center to the periphery. These improvements demonstrate that our synthetic data effectively teaches the model the characteristic geometry and appearance of posterior segment surgery.

The benefits of fine-tuning transfer strongly to real surgical videos, as illustrated in Fig. 6. Without fine-tuning, Marigold tends to produce overly flat predictions, failing to distinguish instruments or retinal curvature. After fine-tuning, the model successfully detects surgical instruments, assigns realistic depth variation along their shafts, and accurately reconstructs the curved retinal surface.

ZoeDepth [4]. As shown in Fig. 7, the zero-shot

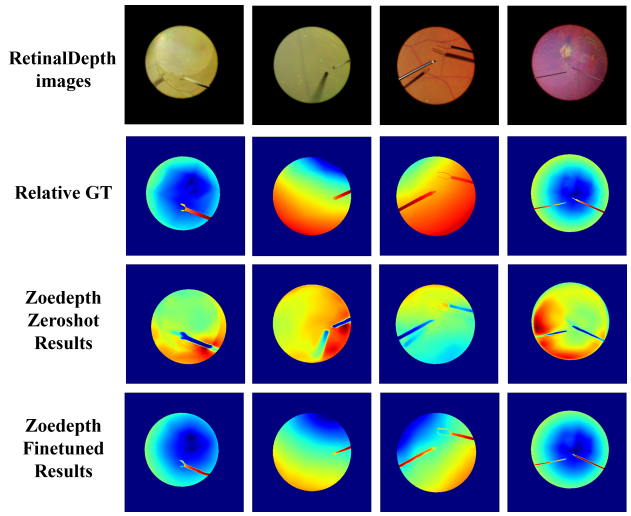


Figure 7. Inference Results on **Synthetic** RetinalDepth Images Before and After Fine-tuning of **ZoeDepth**

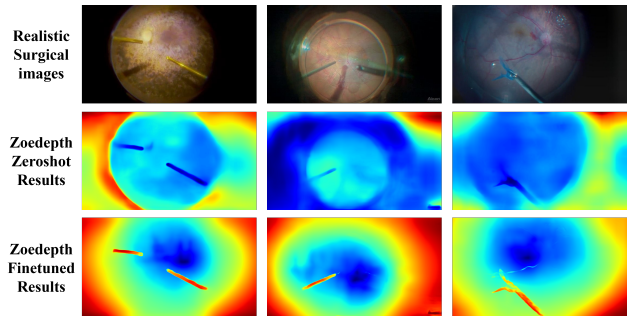


Figure 8. Inference Results on **Realistic** Surgical Images Before and After Fine-tuning of **ZoeDepth**

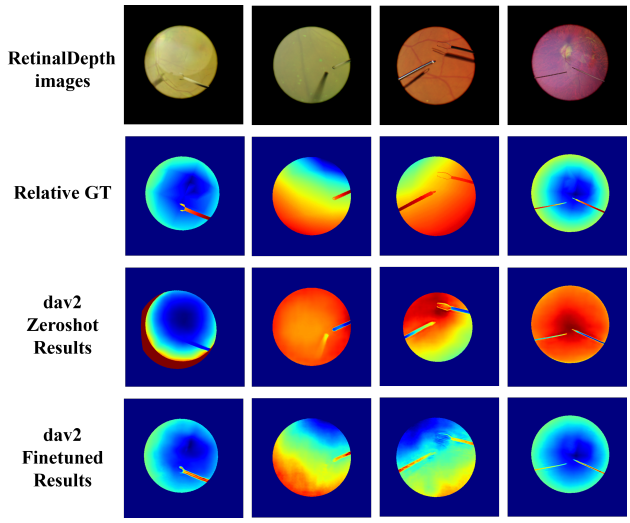


Figure 9. Inference Results on **Synthetic** RetinalDepth Images Before and After Fine-tuning of **Depth Anything V2**

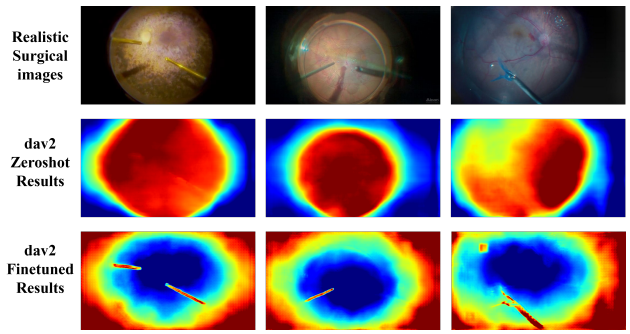


Figure 10. Inference Results on **Realistic** Surgical Images Before and After Fine-tuning of **Depth Anything V2**

ZoeDepth model already outperforms zero-shot Marigold on synthetic data: instrument boundaries and even cast shadows are clearly delineated. However, similar to Marigold, it fails to recover meaningful depth gradients along instrument shafts or the natural spherical curvature of the retina, producing relatively flat predictions overall.

After fine-tuning on RetinalDepth, ZoeDepth exhibits substantial improvement. Instrument details, especially depth variation at the tool tips and along the shafts, are now accurately captured, and the retinal surface correctly recovers its expected concave spherical shape.

This gain is effectively transferred to real surgical videos, as illustrated in Fig. 8. The fine-tuned model yields significantly sharper and more structured depth maps compared to its zero-shot version. Nevertheless, limitations remain in very challenging cases, as shown in column 2, where motion blur or low contrast at instrument edges still leads to partial misestimation of tool shape and depth.

Depth Anything V2 [34]. As shown in Fig. 9, the zero-shot DAV2 reliably detects surgical instruments and captures noticeable depth gradients along their shafts, particularly evident in columns 3 and 4. However, it still exhibits clear failure modes. Especially when the instrument's appearance resembles the retinal background, as shown in column 1, it fails to separate the two. More critically, from column 2 to column 4, it incorrectly inverts the global shape of the eye, predicting a convex bulge where the posterior pole should be concave.

After fine-tuning on RetinalDepth, these issues are largely resolved. Instrument-background separation becomes robust, and the model consistently recovers the correct concave spherical geometry of the retina while preserving fine-grained depth variation on the instruments.

This correction transfers effectively to real surgical videos as depicted in Fig. 10. The fine-tuned model produces geometrically plausible depth maps with accurate retinal concavity and clear instrument structuring, representing a substantial improvement over the zero-shot ver-

sion in both domain-specific shape priors and foreground-background separation.

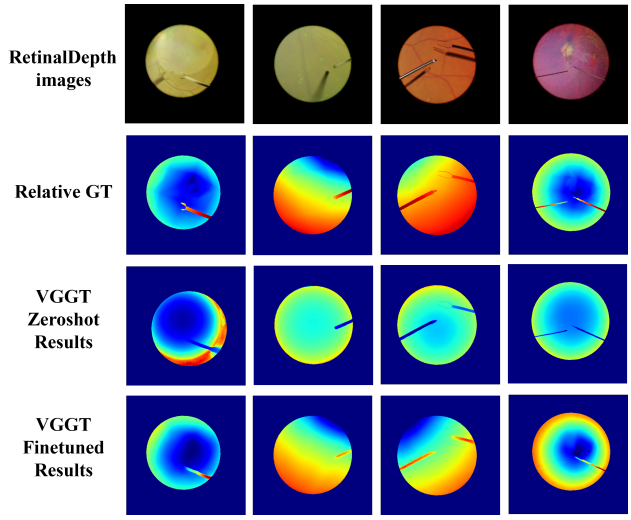


Figure 11. Inference Results on **Synthetic** RetinalDepth Images Before and After Fine-tuning of **VGGT**

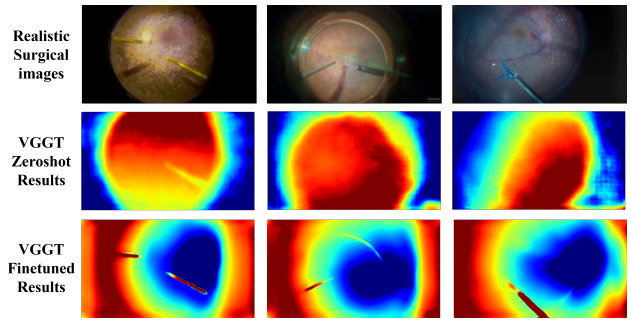


Figure 12. Inference Results on **Realistic** Surgical Images Before and After Fine-tuning of **VGGT**

VGGT [27]. As shown in Fig. 11, the zero-shot VGGT model reliably delineates instrument contours on synthetic data but similar to most monocular baselines, fails to recover meaningful depth gradients along the instrument shafts or the natural concave curvature of the posterior retina.

After fine-tuning on RetinalDepth, VGGT successfully captures depth variation along the instrument bodies. However, fine details at the instrument tips—regions with limited distinctive texture or features—remain under-resolved in several cases.

On real surgical videos as illustrated in Fig. 12, the fine-tuned VGGT clearly separates instruments from the retinal background and assigns realistic depth gradients to the tools, outperforming its zero-shot version. That said, recovery of the overall retinal concavity is less pronounced than with the best monocular methods, suggesting that VGGT still relies more heavily on local cues than on strong global spherical priors even after fine-tuning.

MoGE [28]

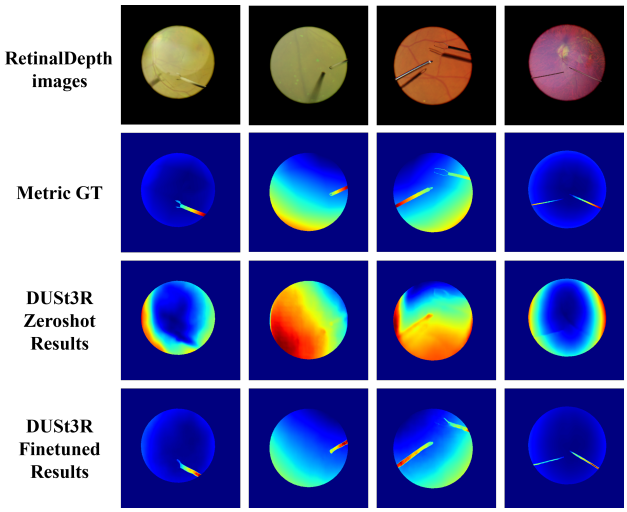


Figure 13. Inference Results on **Synthetic** RetinalDepth Images Before and After Fine-tuning of **DUSt3R**

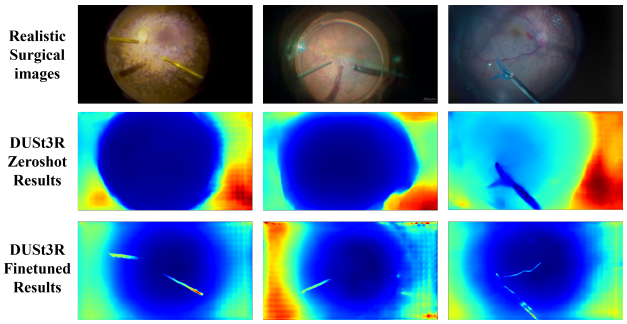


Figure 14. Inference Results on **Realistic** Surgical Images Before and After Fine-tuning of **DUSt3R**

DUSt3R [29]. As shown in Fig. 13, the zero-shot DUSt3R performs noticeably worse than zero-shot VGGT on synthetic data. It may be due to the different amount of training data. The surgical instruments are largely undetected, and the concave posterior retina is not recovered, resulting in overly smooth and unstructured depth maps.

After fine-tuning on RetinalDepth, DUSt3R improves dramatically. For instance, instruments become clearly visible with plausible depth structure, and the retinal background correctly exhibits the expected spherical concavity. However, similar to fine-tuned VGGT, fine-grained details at instrument tips remain partially under-resolved due to limited local texture.

On real surgical videos as shown in Fig. 14, the fine-tuned DUSt3R demonstrates strong generalization. It re-

liably reconstructs both instrument depth and overall retinal curvature. Notably, in column 3, it even captures subtle raised structures corresponding to retinal vessels, achieving the most detailed and anatomically faithful reconstruction among all tested methods.

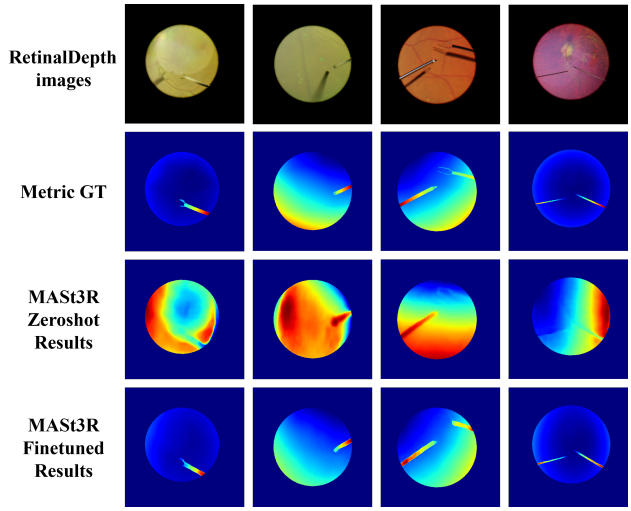


Figure 15. Inference Results on **Synthetic** RetinalDepth Images Before and After Fine-tuning of **MASt3R**

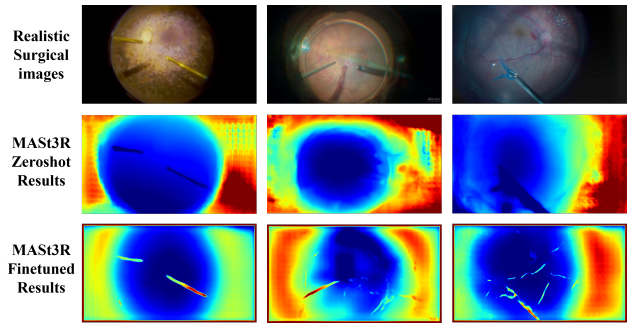


Figure 16. Inference Results on **Realistic** Surgical Images Before and After Fine-tuning of **MASt3R**

MASt3R [11]. As shown in Fig. 15, the zero-shot MASt3R behaves similarly to zero-shot DUSt3R. The surgical instruments are barely detectable, resulting in almost no recoverable depth structure for the tools, and the concave curvature of the posterior retina is completely absent.

After fine-tuning on RetinalDepth, MASt3R improves markedly, with instruments that are now clearly segmented and exhibit realistic depth variation along their shafts, and the retinal background correctly displays the expected spherical concavity. As with the other stereo-based methods, however, fine details at instrument tips, such as the forceps head in column 3, remain under-resolved owing to limited local texture cues.

On real surgical videos as illustrated in Fig. 16, the fine-tuned MASt3R generalizes very well: it reliably detects instruments, assigns plausible depth gradients to them, and faithfully reconstructs the overall retinal curvature, achieving results comparable to or slightly better than fine-tuned DUSt3R while sharing the same minor limitations on ultra-fine tool-tip geometry.

3.2. Comparison of Video Depth Estimation

We further evaluate video depth estimation on synthetic sequences from RetinalDepth, which contains 50 frames per scene. Fig. 17 shows relative depth predictions and ground truth for 10 representative frames of one challenging sequence. Among the tested models, ZoeDepth achieves the sharpest instrument details and excellent temporal consistency, but noticeably underestimates the concave retinal curvature. Depth Anything V2 and Marigold accurately recover the overall spherical shape of the posterior pole compared with Zoodepth and maintain a smooth temporal evolution, yet struggle with fine instrument structures, such as those at t=15 and t=20. VGGT, despite being stereo-based, fails to resolve fine instrument tips and shafts, although its temporal consistency remains reasonable. These results highlight that, even after fine-tuning, current foundation models still trade off between accurate global retinal geometry and precise local instrument reconstruction in dynamic posterior segment scenes.

We additionally evaluate the two stereo-based reconstruction models, DUSt3R and MASt3R, on the other 50-frame synthetic sequences. Fig. 18 displays ground-truth depth alongside predictions for 10 selected frames. Both models successfully recover the concave retinal curvature and clearly separate the instruments from the background even when the tools appear as thin, low-contrast structures. Although fine details at the forceps tips are sometimes smoothed due to the small stereo baseline and limited texture, the overall geometry of both the posterior pole and the instrument shafts is reconstructed with high fidelity and excellent frame-to-frame stability.

We finally apply the fine-tuned models to real posterior-segment surgical videos as illustrated in Fig. 19. ZoeDepth, Depth Anything V2, Marigold, VGGT, DUSt3R, and MASt3R after finetuning all yield plausible and temporally coherent results: the concave retinal surface is relatively clearly recovered, instruments are well separated from the background, and depth varies smoothly along the tool shafts. Some residual high-frequency artifacts and minor burrs remain visible on the instruments—particularly under strong specular reflections or motion blur. However, the overall 3D geometry is sufficiently faithful to support downstream tasks, such as instrument tracking and collision-aware navigation in real surgery. This confirms that fine-tuning on RetinalDepth effectively bridges the sim-

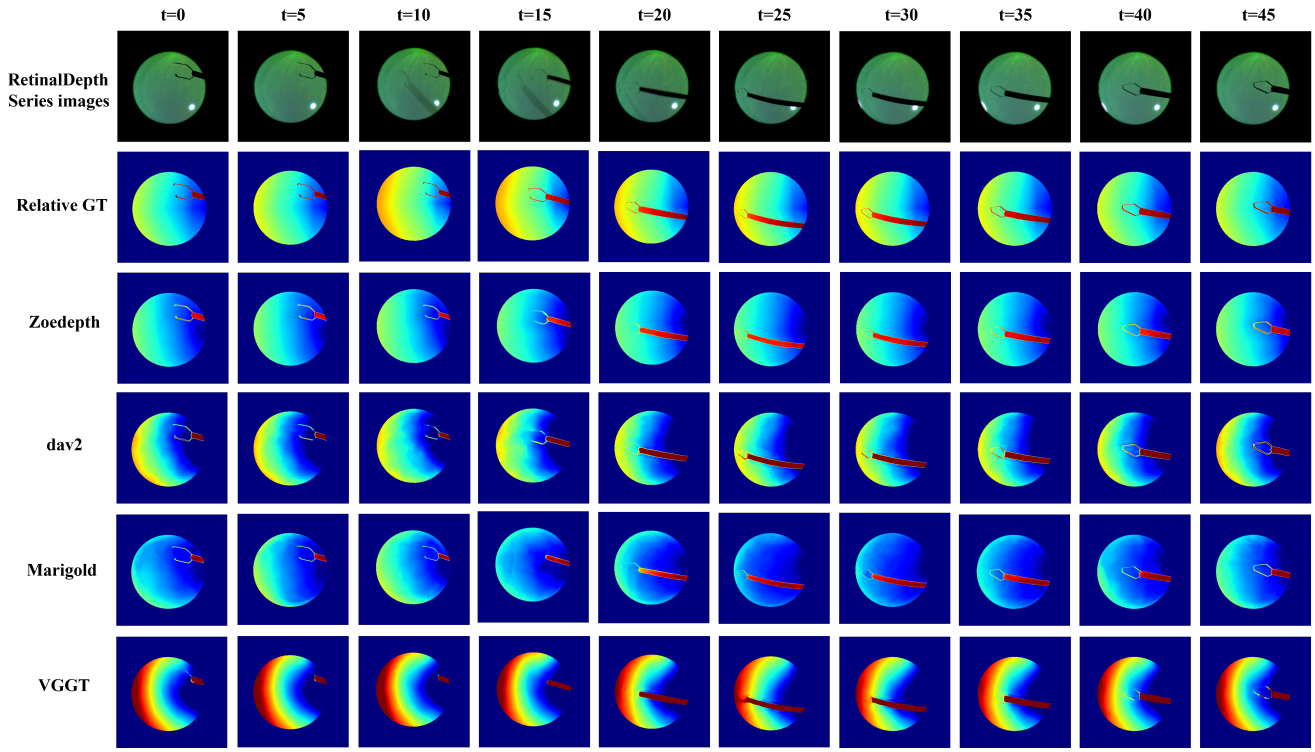


Figure 17. Inference Results on **Synthetic** Surgical Videos After Fine-tuning of various foundation models

to-real gap for dynamic posterior-segment depth estimation.

4. More Related Works

4.1. Medical Depth Datasets

Medical depth datasets are essential for training and validating depth estimation models in surgical applications, with a predominant focus on endoscopy and a nascent emphasis on microscopy, as illustrated in the Table 1. These datasets are broadly classified into real-world and synthetic categories, reflecting diverse imaging modalities and data acquisition strategies tailored to specific anatomical regions.

Real-world medical depth datasets are primarily captured using endoscopic systems, with depth maps generated through three main methods: structured light projection for sparse depth maps, computed tomography (CT) for dense depth maps, and stereo vision algorithms applied to binocular images for dense depth estimation and treated as groundtruths. EndoAbs [19] utilizes structured light projection to derive depth labels from 120 stereo image pairs of liver, kidney, and spleen surgeries recorded in 2018, offering high-fidelity representations but constrained by its small scale. SCARED [2], introduced as a stereo correspondence and reconstruction challenge at MICCAI 2019, expands on this with 22,950 stereo image pairs from porcine abdominal cadavers, employing structured light for sparse depth maps to facilitate 3D reconstruction, though it lacks normal and segmentation annotations. SCARED-C [24], derived from the original SCARED test set as part of the EndoDepth benchmark, evaluates the robustness of monocular depth prediction models in endoscopic scenarios by applying 16 types of image corruptions, such as lens distortion, specular reflection, resolution changes, and color variations across five severity levels, of which depth maps are from the original SCARED. The Hamlyn Centre Laparoscopic/Endoscopic Video Dataset [7], produced by the Hamlyn Centre at Imperial College London, contains a collection of laparoscopic and endoscopic videos. Recasens et al. [23] processed a subset of the Hamlyn Dataset using the Libelas stereo matching software, yielding 30 videos of ex-vivo porcine torsos with stereo-derived depth maps. SERV-CT [6], consisting of 16 stereo image pairs from 2021, integrates cone-beam CT with stereo endoscopy for disparity validation, providing camera intrinsics and extrinsics but limited by its small sample size. EndoNeRF [31] includes 807 frames extracted from 6 clips of in-house DaVinci robotic prostatectomy stereo videos captured at 15 fps and lasting 4–8 seconds per clip, featuring challenging scenes with non-rigid deformation and tool occlusion, including traction on thin structures, tissue pushing/pulling, and cutting, with depths estimated using the STTR-light algorithm [12]. StereoMIS [8], an in vivo dataset recorded

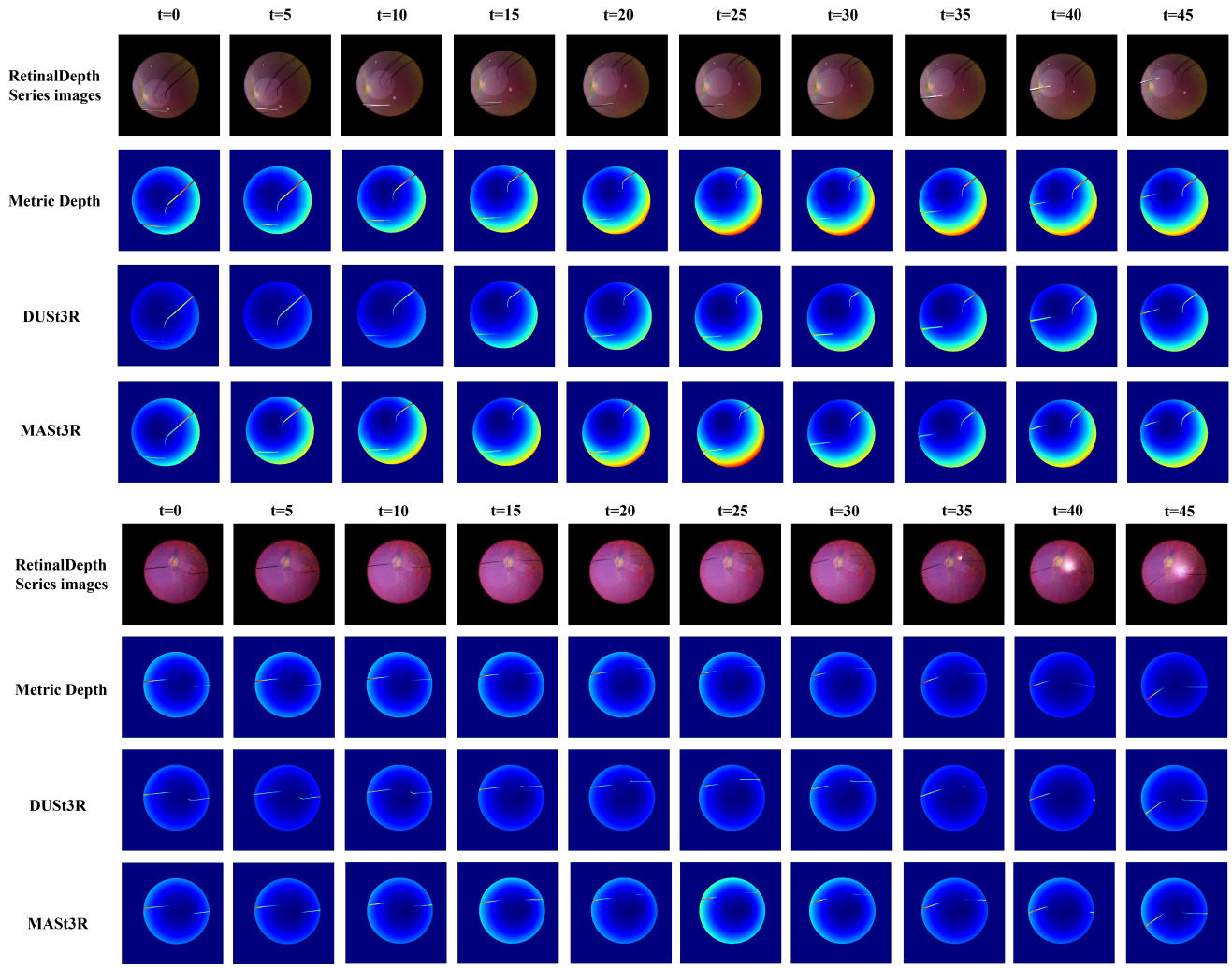


Figure 18. Inference Results on **Synthetic** Surgical Videos After Fine-tuning of DUSt3R and MASt3R

with a da Vinci Xi surgical robot, comprises 16 sequences from 3 porcine and 3 human subjects with durations ranging from 50 seconds to 30 minutes, incorporating ground-truth camera poses from endoscope forward kinematics and depicting scenes with breathing motions, tissue deformations, resections, bleeding, and smoke; while no depths are originally provided, Deform3DGS [35] induced depths using the RAFT algorithm [26] on the binocular images.

Synthetic datasets address these limitations of real-world data, such as annotation inconsistencies and scale constraints, by providing controlled environments for generating high-fidelity depth annotations. EndoUCL [21] (16,016 images, 2019) employs generative adversarial networks to synthesize colon images with pixel-level depth maps, enabling domain adaptation but lacking stereo support. Hybrid datasets like EndoSLAM [17] (35,993 images, 2021) and Endomapper [3] (96 videos, 2023) integrate real and synthetic intestinal tract data, offering depth maps and, in Endomapper's case, segmentation masks, to facilitate visual odometry and 3D reconstruction. Colon-focused Simcol3D [20, 22] (36,000 images, 2023) supports bimodal pose prediction with depth maps, while C3VD [5] (10,015 images, 2023) provides paired depth via 2D-3D registration, though both lack extensive normal labels. The microscope-based SMDE [39] (2,350 images, 2023) offers monocular depth maps for anterior eye cataract surgery assistance but omits stereo, camera parameters, and segmentation, restricting its utility for complex ocular depth tasks.

In summary, existing medical depth datasets demonstrate significant progress in endoscopic depth estimation but are limited by small scales, inconsistent annotations, and a lack of fundus-specific content, particularly for the posterior eye in synthetic formats. Our RetinalDepth addresses these gaps as the first large-scale synthetic dataset

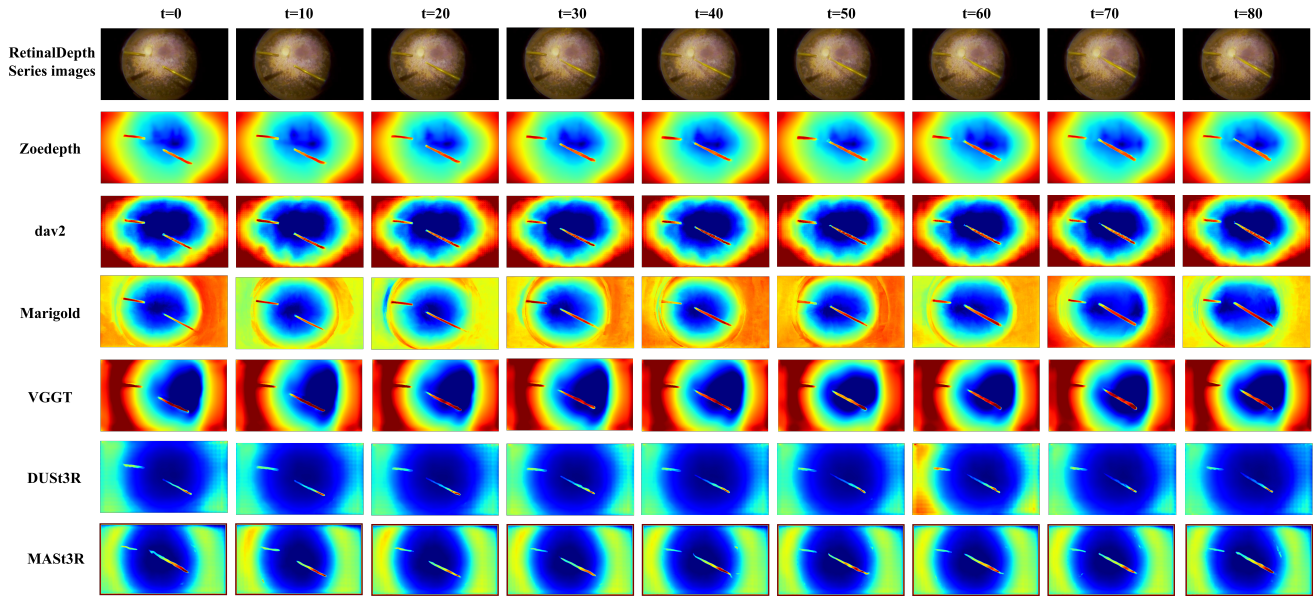


Figure 19. Inference Results on **Realistic** Surgical Videos After Fine-tuning of various foundation models

for retinal surgery, providing 64,000 stereo image pairs with comprehensive annotations including depth maps, normal labels, and camera parameters through the Real2Sim2Real pipeline, thereby enabling robust monocular or binocular depth estimation, enhancing surgical precision, and supporting novice surgeon training in ophthalmic procedures.

4.2. Depth Estimation Models

Depth estimation is a fundamental component of computer vision, enabling precise positioning for surgical robots and accurate 3D reconstruction of surgical scenes. Its origins trace back to the 1950s and 1960s, with foundational work by pioneers such as David Marr and Geoffrey E. Hinton, who developed methods to extract depth from stereo image pairs, laying the groundwork for modern techniques. Current research in depth estimation spans traditional geometric approaches and advanced deep learning methods. This subsection provides a comprehensive analysis of these approaches, emphasizing their relevance to medical applications.

Traditional depth estimation relies on geometric principles, primarily through stereo vision and structure-from-motion (SfM) techniques. Stereo vision processes left and right images through four key steps: matching cost calculation, cost aggregation, disparity calculation, and optimization. These steps produce a disparity map, which is converted to a depth map using the equation $z = \frac{b \cdot f}{d}$, where b represents the baseline distance, f is the focal length, and d is the disparity. The SfM algorithm [25] enhances multi-view depth estimation by extracting and matching feature points using descriptors like SIFT or ORB across images from multiple viewpoints. It estimates camera motions via essential and fundamental matrices, recovers 3D points through triangulation, and refines results with bundle adjustment. SfM is particularly effective for multi-view inputs, such as endoscopic sequences, but struggles with real-world challenges like limited viewpoints, noisy data, occlusions, and adaptation to dynamic or constrained imaging conditions, such as those in ophthalmic surgery.

The advent of deep learning has revolutionized depth estimation, enabling monocular, binocular, and continuous-viewpoint approaches that address challenges like single-view depth inference. State-of-the-art deep learning models fall into two categories: supervised and self-supervised learning.

Supervised depth estimation models leverage diverse architectures, such as vision transformers (ViTs) and diffusion models, to extract rich semantic and geometric features. ViT-based models, which benefit from large-scale data, offer faster training compared to diffusion models. For example, the Depth Anything series [33, 34] employs a ViT architecture with a hybrid training paradigm, combining pseudo-labeled data with supervised learning. By scaling to 1.5 million images, it improves monocular depth accuracy, supporting the "data scaling hypothesis" for geometric perception. Metric3D [36] introduces a camera-space decoupling strategy, projecting images into a scale-invariant space for depth prediction and recovering metric scales, which enhances cross-dataset generalization. DUSt3R [29] uses a regression-based grid model to compute dense scene representations from image pairs without prior scene knowledge, improving efficiency and ro-

Table 1. Inference Results on Synthetic Surgical Videos. After fine-tuning of various foundation models.

Dataset	Imaging	Category	Organs	Year	Volume	B.	D.	N.	CI.	CE.	Seg.	T.
EndoAbs [19]	Endoscopy	Real	Abdominal organs	2018	120 image pairs	✓	P	×	✓	✓	×	×
SCARED [2]	Endoscopy	Real	Porcine abdomen	2019	22,950 image pairs	✓	P	×	✓	✓	×	✓
Hamlyn [7, 16]	Endoscopy	Real	Porcine (ex-vivo)	2021	30 videos	✓	S	×	✓	✓	×	✓
SERV-CT [6]	Endoscopy	Real	Porcine (ex-vivo)	2021	16 image pairs	✓	M	×	✓	✓	×	×
EndoNeRF [31]	Endoscopy	Real	Prostatectomy	2022	807 image pairs	✓	✓	×	✓	✓	✓	✓
StereoMIS [8]	Endoscopy	Real	Porcine (in-vivo)	2023	10 videos	✓	S	×	✓	✓	✓	✓
SCARED-C [24]	Endoscopy	Real	Porcine abdomen	2024	551 images	✓	P	×	✓	✓	×	✓
EndoUCL [21]	Endoscopy	Syn.	Colon	2019	16,016 images	×	M	×	×	×	×	✓
EndoSLAM [17]*	Endoscopy	Syn.	Colon, intestine	2021	35,993 images	×	M	×	✓	✓	×	✓
EndoMapper [3]*	Endoscopy	Syn.	Intestinal tract	2023	96 videos	×	M	×	✓	✓	✓	✓
SimCol3D [22]	Endoscopy	Syn.	Colon	2023	36,000 images	×	M	×	✓	✓	×	✓
C3VD [5]	Endoscopy	Syn.	Colon	2023	10,015 images	✓	M	✓	✓	✓	×	✓
SMDE [39]	Microscope	Syn.	Anterior eye	2023	2,350 images	×	M	×	×	×	✓	×
RetinalDepth(Ours)	Microscope	Syn.	Posterior eye	2025	44,800 image pairs	✓	M	✓	✓	✓	✓	✓

* EndoSLAM and EndoMapper include both realistic and synthetic data; depth maps are available only for the synthetic portion.

bustness. Extensions like MonST3R [38] adapt DUSt3R for dynamic scenes by estimating per-timestep geometry, while MASt3R adds a descriptor head for precise matching. Diffusion-based models, such as ECoDepth [18], condition diffusion models for monocular depth estimation, efficiently handling complex scenes. GenPercept [32] optimizes visual perception with single-step inference and pixel-level loss, while Marigold [10] repurposes Stable Diffusion for affine-invariant monocular depth estimation, achieving zero-shot generalization through denoising U-Net fine-tuning and geometric alignment. MiDaS-based models, such as ZoeDepth [4], enhance zero-shot generalization with a lightweight decoder for metric depth prediction, while PatchFusion [13] refines ZoeDepth using multi-scale patch-wise fusion and cross-dataset training to mitigate scale drift.

Self-supervised depth estimation has gained traction by leveraging unlabeled data to infer depth through consistency losses, such as photometric reconstruction or geometric constraints, enabling robust performance in challenging scenarios without costly ground-truth annotations [9]. These advancements highlight the trend toward efficiency and generalization, making self-supervised models highly promising for medical imaging, where labeled data is often scarce.

Building upon general advancements, depth estimation models tailored for medical datasets address unique challenges such as low texture, specular reflections, and dynamic environments in endoscopic and surgical scenes. Pioneering works like Mahmood et al. [14] introduced deep learning with conditional random fields for depth estimation from conventional endoscopy images, enabling topographical reconstruction. Self-supervised approaches have been particularly effective due to the scarcity of labeled medical data. Recent surveys [30] review these developments, highlighting models that incorporate camera models for improved accuracy in reflective surfaces. Stereo-based methods [1, 15, 37] have been adapted for medical applications, enhancing generalization across datasets, paving the way for robot-assisted surgery and 3D scene reconstruction.

Our RetinalDepth, with 44,800 stereo image pairs, addresses these limitations by providing depth maps, normal labels, and segmentation masks, designed to support a range of depth estimation models—monocular such as ZoeDepth, binocular such as DUSt3R, and video-based such as MonST3R, enabling comprehensive evaluation in subsequent experiments. This versatility enhances surgical precision while also advancing ophthalmic research by providing a fundus-specific benchmark for cutting-edge computer vision techniques.

References

[1] M. Allan, J. Mcleod, C. Wang, J. C. Rosenthal, Z. Hu, N. Gard, P. Eisert, K. X. Fu, T. Zeffiro, W. Xia, et al. Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133, 2021. 10

[2] Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, et al. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133*, 2021. 7, 10

[3] Pablo Azagra, Carlos Sostres, Ángel Ferrández, Luis Riazuelo, Clara Tomasini, O León Barbed, Javier Morlana, David Recasens, Victor M Batlle, Juan J Gómez-Rodríguez, et al. Endomapper dataset of complete calibrated endoscopy procedures. *Scientific Data*, 10(1):671, 2023. 8, 10

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3, 10

[5] Taylor L Bobrow, Mayank Golhar, Rohan Vijayan, Venkata S Akshintala, Juan R Garcia, and Nicholas J Durr. Colonoscopy 3d video dataset with paired depth from 2d-3d registration. *Medical image analysis*, 90: 102956, 2023. 8, 10

[6] PJ Eddie"Edwards, Dimitris Psychogyios, Stefanie Speidel, Lena Maier-Hein, and Danail Stoyanov. Serv-ct: A disparity dataset from ct for validation of endoscopic 3d reconstruction. *arXiv e-prints*, pages arXiv–2012, 2020. 7, 10

[7] Stamatia Giannarou, Danail Stoyanov, David Noonan, George Mylonas, Jim Clark, Marco Visentini-Scarzanella, Pete Mountney, and Guang-Zhong Yang. Hamlyn centre laparoscopic / endoscopic video datasets. 7, 10

[8] Michel Hayoz, Christopher Hahne, Mathias Gallardo, Daniel Candinas, Thomas Kurmann, Maximilian Allan, and Raphael Sznitman. Learning how to robustly estimate camera pose in endoscopic videos. *International journal of computer assisted radiology and surgery*, 18(7):1185–1192, 2023. 7, 10

[9] Wei Jiang, Hao Zhang, and Yang Liu. Self-supervised depth estimation: A comprehensive survey. *arXiv preprint arXiv:2401.03456*, 2024. 10

[10] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 10

[11] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 6

[12] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6197–6206, 2021. 7

[13] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10016–10025, 2024. 10

[14] F. Mahmood and N. J. Durr. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. *Med. Image Anal.*, 48:230–243, 2018. 10

[15] N. Mayer, E. Ilg, P. H"ausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048. IEEE, 2016. 10

[16] Peter Mountney, Danail Stoyanov, and Guang-Zhong Yang. Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Processing Magazine*, 27 (4):14–24, 2010. 10

[17] Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L Bobrow, Gulfize Coskun, Kagan Incetan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, et al. Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical image analysis*, 71:102058, 2021. 8, 10

[18] Suraj Patni, Aradhye Agarwal, and Chetan Arora. Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28285–28295, 2024. 10

[19] Veronica Penza, Andrea S Ciullo, Sara Moccia, Leonardo S Mattos, and Elena De Momi. Endoabs dataset: Endoscopic abdominal stereo image dataset for benchmarking 3d stereo reconstruction algorithms. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 14(5):e1926, 2018. 7, 10

[20] Anita Rau, Sophia Bano, Yueming Jin, Pablo Azagra, Javier Morlana, Rawen Kader, Edward Sanderson, Bogdan J Matuszewski, Jae Young Lee, Dong-Jae Lee, et al. Simcol3d—3d reconstruction during colonoscopy challenge. *Medical Image Analysis*, 96: 103195. 8

[21] Anita Rau, PJ Eddie Edwards, Omer F Ahmad, Paul Riordan, Mirek Janatka, Laurence B Lovat, and Danail Stoyanov. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. *International journal of computer assisted radiology and surgery*, 14(7):1167–1176, 2019. 8, 10

[22] Anita Rau, Binod Bhattarai, Lourdes Agapito, and Danail Stoyanov. Bimodal camera pose prediction for endoscopy. *IEEE Transactions on Medical Robotics and Bionics*, 5(4):978–989, 2023. 8, 10

[23] David Recasens, José Lamarca, José M Fácil, JM Martinez Montiel, and Javier Civera. Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE Robotics and Automation Letters*, 6(4): 7225–7232, 2021. 7

[24] Ivan Reyes-Amezcua, Ricardo Espinosa, Christian Daul, Gilberto Ochoa-Ruiz, and Andres Mendez-Vazquez. Endodepth: A benchmark for assessing robustness in endoscopic depth prediction. In *MICCAI Workshop on Data Engineering in Medical Imaging*, pages 84–94. Springer, 2024. 7, 10

[25] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 9

[26] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 8

[27] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 5

[28] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5261–5271, 2025. 5

[29] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 5, 9

[30] X. Wang, B. Yang, M. Wei, L. Liu, J. Zhang, and Y. Nie. Deep learning for endoscopic depth estimation: A review. *Displays*, page 103086, 2025. 10

[31] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 431–441. Springer, 2022. 7, 10

[32] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. Diffusion models trained with large data are transferable visual models. *arXiv preprint arXiv:2403.06090*, 2024. 10

[33] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 9

[34] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 4, 9

[35] Shuojue Yang, Qian Li, Daiyun Shen, Bingchen Gong, Qi Dou, and Yueming Jin. Deform3dgs: Flexible deformation for fast surgical scene reconstruction with gaussian splatting, 2024. 8

[36] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 9

[37] J. Žbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(65):1–32, 2016. 10

[38] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 10

[39] Yingquan Zhou, Zhongxi Qiu, Mingming Yang, Yan Hu, and Jiang Liu. Synthetic monocular depth estimation dataset for cataract surgery assistance. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1812–1817. IEEE, 2023. 8, 10