Towards Dynamic 3D Reconstruction of Hand-Instrument Interaction in Ophthalmic Surgery

Ming Hu[1,2,3]* **Zhengdi Yu**[4]* **Feilong Tang**[1,2,3]* **Kaiwen Chen**[5] **Yulong Li**[3]
Imran Razzak[3] **Junjun He**[2] **Tolga Birdal**[4] **Kaijing Zhou**[5]† **Zongyuan Ge**[1]†

[1]Monash University [2]Shanghai AI Laboratory [3]MBZUAI
[4]Imperial College London [5]Eye Hospital, Wenzhou Medical Univeristy
ming.hu@monash.edu, z.yu23@imperial.ac.uk
https://ophnet-3d.github.io/

Abstract

Accurate 3D reconstruction of hands and instruments is critical for vision-based analysis of ophthalmic microsurgery, yet progress has been hampered by the lack of realistic, large-scale datasets and reliable annotation tools. In this work, we introduce OphNet-3D, the first extensive RGB-D dynamic 3D reconstruction dataset for ophthalmic surgery, comprising 41 sequences from 40 surgeons and totaling 7.1 million frames, with fine-grained annotations of 12 surgical phases, 10 instrument categories, dense MANO hand meshes, and full 6-DoF instrument poses. To scalably produce high-fidelity labels, we design a multi-stage automatic annotation pipeline that integrates multi-view data observation, data-driven motion prior with cross-view geometric consistency and biomechanical constraints, along with a combination of collision-aware interaction constraints for instrument interactions. Building upon OphNet-3D, we establish two challenging benchmarks—bimanual hand pose estimation and hand–instrument interaction reconstruction—and propose two dedicated architectures: H-Net for dual-hand mesh recovery and OH-Net for joint reconstruction of two-hand–two-instrument interactions. These models leverage a novel spatial reasoning module with weak-perspective camera modeling and collision-aware center-based representation. Both architectures outperform existing methods by substantial margins, achieving improvements of over 2mm in Mean Per Joint Position Error (MPJPE) and up to 23% in ADD-S metrics for hand and instrument reconstruction, respectively.

1 Introduction

Modern ophthalmic microsurgery represents one of the most delicate surgical paradigms in medicine, requiring sub-millimeter precision in instrument manipulation under restricted workspace conditions [63, 75]. While advances in robotic tools and surgical training platforms have improved treatment outcomes, current skill assessment methods still rely heavily on expert supervision and subjective feedback, limiting their scalability and objectivity [15, 39, 54]. In current ophthalmic surgical training paradigms, trainees predominantly rely on direct supervision from instructors for skill acquisition and performance evaluation. However, this approach imposes significant demands on instructional resources, particularly considering the time-intensive nature of surgical mentorship and the critical requirement for real-time feedback in complex microsurgical procedures. Recent studies in computer-assisted surgery [37, 73, 17, 21] reveal that kinematic analysis of surgical tools and operator hand movements could enable objective skill evaluation, personalized training feedback,

---

*Equal contribution, †Corresponding author

Table 1: **Comparison with existing 3D hand reconstruction datasets.** OphNet-3D is the first surgical RGB-D dataset, offering high-resolution videos and rich annotations of complex hand–instrument interactions. It far exceeds prior datasets in scale and diversity and uniquely supports real-time dual-hand and multi-object reconstruction tasks. Task abbreviations: HPE: Hand Pose Estimation, OPE: Object Pose Estimation, HR: Hand Reconstruction, HOI: Hand-Object Interaction Reconstruction, HMOI: Hand and Multi-Object Interaction Reconstruction, Video: Video-level Reconstruction.

Datasets	Modality	Source	Views	Resolution	Participants	Objects	Motions	Frames	HPE	OPE	HR	HOI	HMOI	Video
FreiHAND [ICCV'19] [96]	General	Real RGB	8	224×224	32	-	-	130.2K	✓	✗	✓	✗	✗	✓
ObMan [CVPR'19] [27]	General	Real RGB	1	256×256	-	8	-	150K	✓	✓	✓	✓	✗	✗
InterHand2.6M [ECCV'20] [56]	General	Real RGB	80-140	512×334	26	32	-	2.6M	✓	✗	✓	✗	✗	✓
ContactPose [ECCV'20] [5]	General	Real RGB-D	3	256×256	50	25	2	2.9M	✓	✓	✓	✓	✗	✓
H2O [ICCV'21] [40]	General	Real RGB-D	5	1280×720	4	8	36	571.6K	✓	✓	✓	✓	✗	✓
DexYCB [CVPR'21] [7]	General	Real RGB-D	8	640×480	10	20	-	582K	✓	✓	✓	✓	✗	✓
ARCTIC [CVPR'23] [16]	General	Real RGB	9	2800×2000	10	11	2	2.1M	✓	✓	✓	✓	✗	✓
HOT3D [CVPR'25] [1]	General	Real RGB(mocap)	3	1408×1408	19	33	-	1.5M	✓	✓	✓	✓	✗	✓
Hein et al. [IJCARS'21] [31]	Clinical	Synth RGB	2	256×256	2	1	-	10.5K	✓	✓	✓	✓	✗	✗
POV-Surgery [MICCAI'23] [81]	Clinical	Synth RGB-D	3	1920×1080	-	3	3	88.3K	✓	✓	✓	✓	✗	✗
HUP-3D [MICCAI'24] [3]	Clinical	Synth RGB	90	848×480	-	1	11	31.7K	✓	✓	✓	✓	✗	✗
OphNet-3D (Ours)	Clinical	Real RGB-D	8	848×480	40	10	12	7.1M	✓	✓	✓	✓	✓	✓

and even real-time intraoperative guidance [73, 80]. Goodman et al. [21] curated the 1,997-video AVOS corpus and trained a real-time multitask model that parses hands, tools, and actions. Building on AVOS, Vaid et al. [79] reframed surgeon-hand recognition as a semi-supervised, single-class 2D detection task that mixes many noisy unlabeled frames with a few labeled ones. Both efforts still rely on 10-frame 2D snippets and lack long-range temporal, depth, pose, or multi-view cues. Meanwhile, other approaches [43, 58] continue to depend on external motion-capture rigs or wearable sensors, introducing constraints that conflict with sterile surgical environments and disrupt natural workflow.

Recent advances in monocular [13, 94, 82, 59] and multi-view [84, 85] RGB-based 3D hand-object interaction reconstruction have demonstrated notable success in general-purpose scenarios, offering a promising avenue for contactless skill assessment in ophthalmic surgical training. Estimating surgeons' hand and instrument poses from a single RGB image enables the quantification of critical operational details—such as grip posture and tool orientation—that are closely linked to surgical quality and clinical outcomes. However, directly applying these methods to ophthalmic microsurgical settings remains challenging due to the highly constrained operating space, fine-grained motion scale, and frequent occlusions and complex interactions between both hands and multiple instruments. These factors pose significant difficulties for accurate motion structure reconstruction using existing algorithms. Moreover, the lack of high-precision, realistically annotated 3D datasets specific to ophthalmic surgery further limits methodological development and evaluation in this domain.

To address the aforementioned limitations, we introduce OphNet-3D, the first large-scale dataset capturing dynamic 3D hand–instrument interactions in real-world ophthalmic surgeries. Collected with eight synchronized RGB-D cameras, it comprises 41 cataract surgery sequences by 40 surgeons (avg. >12 min/sequence), totaling over 7.1 M aligned RGB–D frames annotated for 12 surgical phases and 10 instrument categories. We apply a multi-stage automatic annotation pipeline to recover dense 3D hand meshes and 6D tool poses from these multi-view videos. Finally, we define two evaluation benchmarks—bimanual hand-pose estimation and hand–instrument interaction—and propose a unified baseline, OH-Net, for joint reconstruction of two-hand–two-tool interactions. Our contributions are:

- We present OphNet-3D, the first large-scale, real-world, high-quality dataset for 3D reconstruction of hand–instrument interactions in clinical surgical settings. OphNet-3D delivers an unprecedented combination of dataset scale, camera views, participant diversity, instrument, motion and object variety, and supported task types, at **2.5×** the size of the largest existing general 3D hand reconstruction dataset and **70×** that of the largest surgical 3D hand reconstruction dataset.

- We propose a multi-stage automatic annotation pipeline that reconstructs 3D hand meshes and 6D tool poses from multi-view RGB-D videos using optimization with data-driven hand motion prior combined with biomechanical constraints and interaction-aware refinement.

- Based on the proposed dataset, we establish two benchmarks: one for bimanual hand pose estimation and another for hand-tool interaction. We further introduce a unified baseline, OH-Net, which jointly reconstructs two-hand–two-tool interactions with effective spatial reasoning, and demonstrate its performance through extensive quantitative and qualitative results.
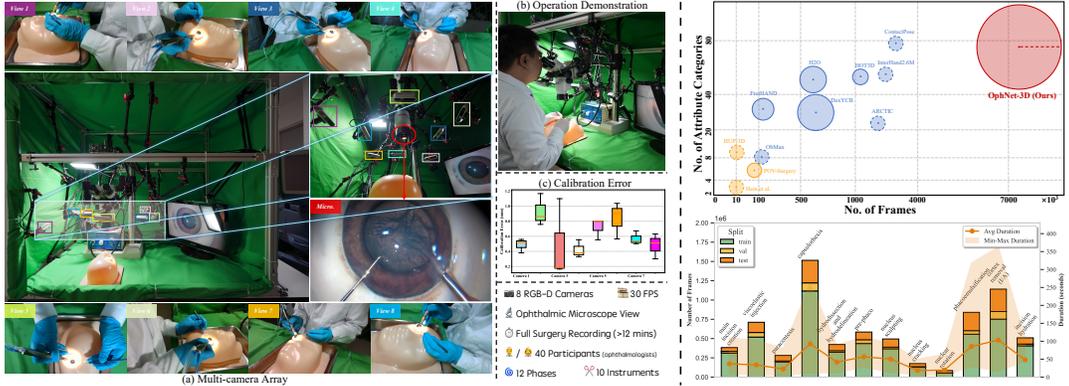
Figure 1: **OphNet-3D's acquisition framework, comparisons with other datasets, and phase–frame distributions.** *Left:* (a) a synchronized multi-camera rig with 8 calibrated RGB-D cameras and 3 directional LED lights; (b) participants perform standardized cataract surgery maneuvers on pig-eye simulators under an ophthalmic microscope; (c) boxplots of pixel errors for eight cameras across three calibration runs. *Upper Right:* comparative visualization of OphNet-3D and existing 3D hand datasets. The horizontal axis denotes the total number of RGB frames, while the vertical axis indicates the number of categories across the participants, objects, and motion settings. Circle diameters encode the number of provided segmentation-mask instances; datasets without mask annotations are represented by dashed outlines. *Lower Right:* distribution of frame counts and clip durations for each phase.

2 OphNet-3D Dataset

Data Collection. OphNet-3D is captured in a multi-camera studio consisting of 8 Intel® RealSense™ D435 RGB-D cameras recording at 30 FPS, along with 3 high-powered directional LED lights aimed at the hands to ensure uniform illumination (Fig. 1: left). The cameras capture at a resolution of 848×480, and the multi-view system is calibrated using an ArUco calibration board. The detailed setup of the recording platform can be found in B.1.

We adhered to standard cataract surgical protocols, segmenting the procedure into 12 distinct phases. Detailed definitions and demonstration for each phase are provided in B.3. During the procedures, 10 different surgical instruments were employed and all instruments were scanned using a ZEISS ATOS Q blue-light 3D scanner, with corresponding images of the physical instruments and their 3D CAD models presented in B.5. During each recording session, two additional assistants were present—one operated the recording system, while the other assisted with the surgical workflow, such as instrument handover, to ensure procedural continuity. All surgical actions were recorded on an ophthalmic surgical simulation platform utilizing pig eyes, accompanied by synchronized video captured from an ophthalmic surgical microscope perspective. Finally, all videos were temporally aligned between the hand-view and microscope-view by an ophthalmologist, who also annotated the surgical phase locations and performed a secondary verification.

Data Statistics. We recorded a total of 41 sequences from 40 unique participants (one participant contributed two sequences, wearing blue and white gloves respectively), of whom 20 have more than one year of surgical experience and 20 have less than one year. Raw videos in our dataset have an average duration of 16 minutes, comprising over 9.5M RGB frames. After filtering out transitional segments via phase localization annotation (B.4), the final OphNet-3D contains 565 phase segments with a total duration of 300 hours and over 7.1M RGB frames, as detailed in Tab. 5. In addition, OphNet-3D provides segmentation annotations for more than 21M instances.

3 Automatic Annotation Method

Given the input videos $\{\mathcal{V}^i\}_{i=1}^I$ from multiple views with T frames containing two hands interacting with possibly two surgical instruments, we aim to reconstruct the 3D hand-instrument interacting motions by recovering the hand meshes and 6D instrument poses. To efficiently auto-label the captured RGB-D videos with accurate mesh and pose annotations from multiple views, we design an optimization-based multi-stage automatic annotation pipeline as shown in Fig. 2.
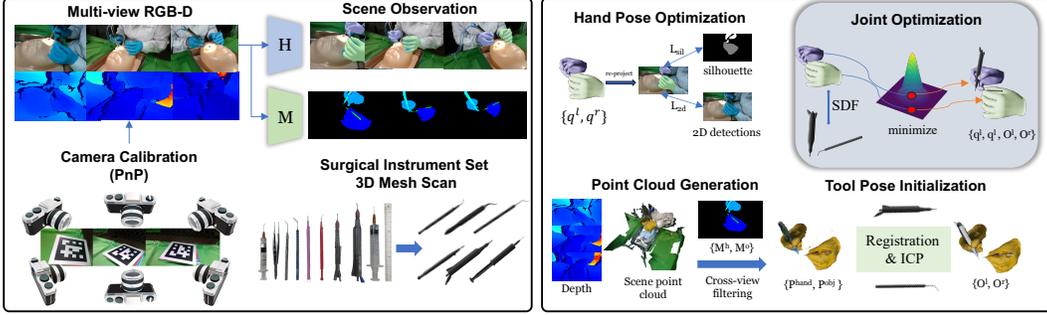
3

Figure 2: **Our automatic annotation pipeline.** Given a multi-view RGB-D video sequence as input, our pipeline reconstructs the 3D hand mesh and 6D instrument pose in a multi-stage manner. **H** and **W** represent the initialization network for hands [61] and instance segmentation masks [67].

In the first stage, we use the 3D CAD models scanned as described in Sec. 2 to track per-frame 6D instrument poses. Moreover, we leverage the state-of-the-art SAM2 [67] with manual corrections to obtain per-frame accurate instance masks for hands ($^i\mathcal{M}_h$) and instruments ($^i\mathcal{M}_o$) for further point cloud segmentation. For per-frame scene point cloud generation, we first compute a point cloud from depth images for each view and merge the point clouds across views. To segment out the region of interest (*i.e.* two hands and instruments) with cross-view filtering, we project the merged point cloud back to the 8 views and keep the points that project onto the hand-instrument region for more than half of the views to get the final per-frame scene point cloud $\mathbf{P}_t$, which can be further split into hand $\mathbf{P}_t^{hand}$ and instrument $\mathbf{P}_t^{obj}$. In the second stage, our goal is to reconstruct the 3D hand and 6D instrument pose from the multi-view RGB-D videos recorded by 8 calibrated cameras. To this end, we leverage the state-of-the-art 2D & 3D hand pose estimation method [61, 52] to initialize per-frame hand motion state in the camera coordinate system, as well as utilizing the masks for global registration and ICP [42] to estimate an initial instrument pose for each camera view. Recovering accurate hand-instrument interaction is challenging due to frequent occlusions, truncation and mutual confusion. As a remedy, we propose a hand-instrument joint optimization scheme with a hand motion prior model HMP [14] and biomechanical constraints in the third stage inspired by [91].

3.1 Hand Motion Annotation

Hand Representation. We parametrize the hand shape and pose using the MANO hand model [69], which uses standard vertex-based linear blend skinning with learned blend shapes. At each time step t, the hand motion state is represented as:

$$\boldsymbol{q}_t^h = \{\boldsymbol{\theta}_t^h, \boldsymbol{\beta}_t^h, \boldsymbol{\phi}_t^h, \boldsymbol{\tau}_t^h\}, \tag{1}$$

where $\boldsymbol{\theta}_t^h \in \mathbb{R}^{3\times15}$ denotes the local pose of 15 hand joints, $\boldsymbol{\beta}_t^h \in \mathbb{R}^{10}$ represents the hand shape parameters, and $(\boldsymbol{\phi}_t^h, \boldsymbol{\tau}_t^h)$ define the global wrist state. Specifically, the orientation $\boldsymbol{\phi}_t^h \in \mathbb{R}^3$ is expressed using the axis-angle representation, while the translation $\boldsymbol{\tau}_t^h \in \mathbb{R}^3$ specifies the wrist position in 3D space. The handedness is indicated by $h \in \{l, r\}$, representing left or right hand, respectively. Using these MANO parameters and the skinning function $\mathcal{W}(\cdot)$, we can reconstruct the 3D hand mesh $\mathbf{V}_t^h \in \mathbb{R}^{3\times778}$ and the 3D hand keypoints $\mathbf{J}_t^h \in \mathbb{R}^{3\times21}$ with $\mathbf{V}_t^h = \mathcal{W}(\mathcal{H}(\mathbf{J}_t^h, \boldsymbol{\beta}^h), \mathcal{P}(\boldsymbol{\beta}^h), \mathbf{S}) + \boldsymbol{\tau}_t^h \mathbf{1}_{778}$ and $\mathbf{J}_t^h = \mathbf{L}\mathbf{V}_t^h$. where $\mathcal{W}(\cdot)$ denotes the skinning function, $\mathcal{H}$ represents the posed parametric hand template, and $\mathbf{1}_{778} \in \mathbb{R}^{1\times778}$ is a row vector of ones. The function $\mathcal{P}(\cdot)$ returns the hand joint positions in the rest pose, $\mathbf{S}$ defines the skinning weights, and $\mathbf{L}$ is a pre-trained linear regressor for estimating joint locations from mesh vertices.

Hand Initialization. For each camera view, we initialize per-frame 3D hand motion state in camera coordinate system leveraging an efficient two-hand motion tracking system based on [61] with hallucinated detection handling. We further obtain per-frame 2D hand keypoints from ViTPose [83] and hand bounding box to extract image patches, feeding into [61] for a coarse-to-fine 3D motion state prediction. Finally, we compute the weighted sum of the motion state based on the visibility to obtain the final initilization of the global motion state $^i\boldsymbol{q}_t^h$ in the world coordinate system. Next, we convert the motion state into the world coordinate system using the calibrated camera information $\{\mathbf{R}_t^i, \boldsymbol{\tau}_t^i\}$ of camera view i:

$$^{\mathrm{w}}\boldsymbol{\phi}_t^h = \mathbf{R}_t^{-1}\cdot{}^{\mathrm{c}}\boldsymbol{\phi}_t^h \qquad \text{and} \qquad {}^{\mathrm{w}}\boldsymbol{\tau}_t^h = \mathbf{R}_t^{-1}\cdot{}^{\mathrm{c}}\boldsymbol{\tau}_t^h - \mathbf{R}_t^{-1}\cdot\boldsymbol{\tau}_t, \tag{2}$$

4

where $^\mathrm{w}\phi_t^h$ and $^\mathrm{c}\phi_t^h$ are the hand wrist orientation in world and camera space. $^\mathrm{w}\boldsymbol{\tau}_t^h$ and $^\mathrm{c}\boldsymbol{\tau}_t^h$ are the translation. Here we omit i for simplicity. For 2D observations, we initialize from ViTPose [83], MediaPipe [52] and the 2D re-projection from [61] with a confidence filter to extract final per-frame 2D keypoint $^i\hat{\mathbf{J}}_\mathbf{t}^\mathbf{h} \in \mathbb{R}^{2\times21}$ for each view as observation for following optimization, where the 2D re-projection is performed with weak-perspective camera parameters predicted by [61]. Moreover, we provide more details regarding the post-processing of the observations, including hallucination handling and missing detection infilling in the D.1.

Optimization. To recover the hand meshes from multi-view RGB-D videos, we propose an iterative fitting algorithm by minimizing the following objective with regularization and *biomechanical hand constraints* [71], as well as hand motion prior [14].

$$E_\mathrm{I}(\boldsymbol{\theta}_t^h, \boldsymbol{\beta}_t^h, \boldsymbol{\phi}_t^h, \boldsymbol{\tau}_t^h) = \sum_{i=1}^{N_i}(\lambda_\mathrm{2d}\mathcal{L}_\mathrm{2d} + \lambda_\mathrm{sil}\mathcal{L}_\mathrm{sil}) + \lambda_\mathrm{s}\mathcal{L}_\mathrm{smooth} + \lambda_\mathrm{3d}\mathcal{L}_\mathrm{3d} + \mathcal{L}_\mathrm{bio} + \mathcal{L}_\mathrm{prior}, \quad (3)$$

where N_i is the number of camera views and $\mathcal{L}_\mathrm{2d}$ is the joint 2D re-projection loss minimizing the difference between 2D hand keypoints observations $\{\hat{\mathbf{J}}_t^h\}_{t=0}^T$ and the re-projection of the 3D keypoints obtained from MANO model with parameters $\{\boldsymbol{\theta}_t^h, \boldsymbol{\beta}_t^h, \boldsymbol{\phi}_t^h, \boldsymbol{\tau}_t^h\}$ in current global state $\boldsymbol{q}_t^h$:

$$\mathcal{L}_\mathrm{2d} = \sum_{h\in\{l,r\}}\sum_{t=0}^T \rho\left(\mathbf{C}_t^h\left(^i\tilde{\mathbf{J}}_t^h - {}^i\hat{\mathbf{J}}_t^h\right)\right). \quad (4)$$

where $\rho(\cdot)$ is the Geman-McClure robust function [19]. $\mathbf{C}_t^h$ is a confidence filter mask for joint visibility. $\tilde{\mathbf{J}}_t^h = \Pi(^\mathrm{w}\mathbf{J}_t^h, \mathbf{R}_t, \boldsymbol{\tau}_t, \mathbf{K})$ is the re-projected 3D keypoints, and Π is the perspective camera projection for each view with collected camera intrinsics $\mathbf{K} \in \mathbb{R}^{3\times3}$ and extrinsics $\{\mathbf{R}_t^i, \boldsymbol{\tau}_t^i\}$. To improve the pixel alignment across views, we propose a silhouette-based error term $\mathcal{L}_\mathrm{sil}$:

$$\mathcal{L}_\mathrm{sil} = \sum_{t=0}^T \| \,^i\mathbf{M}_t^h - \mathbf{NR}(V_t^h, K^i, \mathbf{R}_t^i, \boldsymbol{\tau}_t^i)\| \quad (5)$$

where $\mathbf{NR}(\cdot)$ is a differentiable renderer that renders the 3D mesh for the two hands into the 2D mask and $\mathbf{M}_t^h$ is the clean segmentation mask for the hand h at timestep t. To have a precise 3D alignment between the mesh reconstruction and the real scene, we compute the 3D mesh loss and minimize the distance between the frame point cloud $\mathbf{P}_t^h$ and the MANO hand mesh:

$$\mathcal{L}_\mathrm{3d} = \sum_{h\in\{l,r\}}\sum_{t=0}^T\sum_{i=1}^{N_V}\frac{1}{N_V}\|\left(\mathbf{p}_t^{j(i)} - \mathbf{v}_t^i\right)\cdot\mathbf{n}_i\|, \quad \text{where} \quad j(i) = \arg\min_j \|\mathbf{p}_t^j - \mathbf{v}_t^i\| \quad (6)$$

where $\mathbf{p}_t^j$ is the j-th point of the hand point cloud $\mathbf{P}_t^h$ at timestep t and v_t^i is the i-th vertex of the hand mesh $\mathbf{V}_t^h$. The reconstructed hand mesh normal of the vertex is represented as n_t^i. Moreover, we reduce the jitter of the hand motion and improve the temporal smoothness by integrating $\mathcal{L}_\mathrm{smooth}$:

$$\mathcal{L}_\mathrm{smooth} = \sum_{h\in\{l,r\}}\sum_{t=0}^T \|\mathbf{J}_{t+1}^h - \mathbf{J}_t^h\|^2 + g(\boldsymbol{\theta}_{t+1}^h, \boldsymbol{\theta}_t^h)^2 \quad (7)$$

where $g(\cdot)$ represents the geodesic distance. To further improve the plausibility of hand motion qualityand reduce jitter for natural movement, we compute the $\mathcal{L}_\mathrm{prior} = \mathcal{L}_\mathbf{z} + \lambda_\theta\mathcal{L}_{\boldsymbol{\theta}} + \lambda_\beta\mathcal{L}_{\boldsymbol{\beta}}$ by utilizing a data-driven motion prior [14] where the latent code is $\mathbf{z}^h$ Inspired by [91]. It ensures the motion is constrained under the learnt prior space, by penalizing the negative log-likelihood:

$$\mathcal{L}_\mathbf{z} = \sum_{h\in\{l,r\}}\sum_{t=0}^T -\log\mathcal{N}(\mathbf{z}^h; \mu^h(\{\mathbf{J}_t^h\}), \sigma^h(\{\mathbf{J}_t^h\})).$$

To explicitly prevent implausible poses produced during optimization, we further propose $\mathcal{L}_\mathrm{bio} = \sum_{t=0}^T(\lambda_\mathrm{ja}\mathcal{L}_\mathrm{ja} + \lambda_\mathrm{bl}\mathcal{L}_\mathrm{bl} + \lambda_\mathrm{palm}\mathcal{L}_\mathrm{palm} + \lambda_\mathrm{angle}\mathcal{L}_\mathrm{angle})$, which consists of angle regularization terms with biomechanical constraints [71] and and an angle limitation constraint. More details regarding the loss calculation are provided in D.2.

3.2 Instrument Motion Annotation

By leveraging multi-view RGB-D frames together with camera pose information, our method can accurately annotate per-frame 6D instrument pose of the instruments. To recover 3D hand motion under the challenging surgical scenario (*e.g.* light, occlusion).

Obtaining Canonical Local Instrument Geometry. We laser-scanned each instrument to obtain high-resolution 3D meshes (Sec. 2). For articulated instruments (*e.g. phacoemulsification handpiece*), we additionally separate and scan them into two articulated parts, as well as their rest pose and maximum relative pose articulation as shown in Fig. 2. Please see 16 for the detailed visualization of the instruments.

Acquiring Instrument Articulation. The articulated instrument (*e.g.* handpiece) surface is parameterised by the 6D pose of each base part and a 1D articulation relative to a canonical pose. In particular, the 6D instrument pose can be represented as $\{\mathbf{R}_t^o, \boldsymbol{\tau}_t^o\}$. For each instrument, we define a 3D parametric model $\mathcal{O}(\cdot)$ leveraging the scanned instrument parts and relative pose state. Given the 6D pose $\theta_t^o \in \mathbb{R}^6$ and the 1D relative articulation factor $\alpha_t \in \mathbb{R}^1$, where $\alpha \in [0, 1]$ uniformly parameterizes pose deformation across different models. Here, $\alpha = 0$ corresponds to the rest pose, and $\alpha = 1$ to the maximally articulated pose, representing the deformation state. The instrument 3D mesh $\mathcal{O}(\theta_t^o, \alpha_t) \in \mathbb{R}^{3 \times N_o}$ can be reconstructed, where N_o is the instrument vertices number.

Initialization of 6D instrument pose. We obtain accurate per-frame 6D instrument pose leveraging the multi-view RGB-D information. As mentioned, we first perform instrument segmentation and 2D tracking for each camera view using SAM2 [67] with manual correction to obtain clean 2D segmentation masks for both hands and instruments. Moreover, we merge the depth image across views to generate the point cloud for the whole scene. After that, we segment out the region of interest leveraging a cross-view filter, which projects the point cloud into all camera views and keeps the points projected onto the hand-instrument region for more than half of the views to get the point cloud for instruments $\mathbf{P}_t^{obj}$. By running RANSAC-based global registration, we have a coarse global alignment of the 3D instrument mesh and $\mathbf{P}_t^{obj}$. After that, we refine the alignment with ICP [42] to obtain the initial rigid transformation from instrument canonical coordinate system to the world coordinate system. Finally, we run a simple Chamfer distance-based optimizer for better alignment and to obtain the articulation α_t and the final 6D instrument pose $\{\mathbf{R}_t^o, \boldsymbol{\tau}_t^o\}$. By applying the 6D pose to the instrument model, we can obtain the 3D mesh in the world coordinate system. Note that our surgical scenario contains various two-hand-two-instrument interactions, thus we represent the instrument in each hand as $\mathbf{O}_t^h$.

3.3 Joint Optimization

Naively putting the hand and instrument together may result in implausible hand-instrument interactions such as inter-penetration and unnatural contact. To jointly optimize 3D hand pose and instrument pose and introduce more constraints for the interaction, we propose the following objectives:

$$E_{\text{II}}(\boldsymbol{\theta}_t^h, \boldsymbol{\beta}_t^h, \boldsymbol{\phi}_t^h, \boldsymbol{\tau}_t^h, \mathbf{R}_t^o, \boldsymbol{\tau}_t^o, \alpha_t) = E_I + \sum_{i=1}^{N_i} (\lambda_{\text{sil}} \mathcal{L}_{\text{sil}}) + \lambda_{\text{3d}} \mathcal{L}_{\text{3d}} + \lambda_{\text{inter}} \mathcal{L}_{\text{inter}} + \lambda_{\text{sdf}} \mathcal{L}_{\text{sdf}} \quad (8)$$

Specifically, $\mathcal{L}_{\text{sil}}$ represents the silhouette loss term that is computed between the combined hand-object mask ${}^i\mathbf{M}_t^{h,o}$ and the rendered mask $\mathbf{NR}(V_t^h, O_t^h, K^i, \mathbf{R}_t^i, \boldsymbol{\tau}_t^i)$ of the 3D hand-object mesh. $\mathcal{L}_{\text{3d}}$ is calculated between ground truth point cloud $\mathbf{P}_t$ and the predicted hand and object mesh. Moreover, we leverage the interaction loss $\mathcal{L}_{\text{inter}}$ to constrain the hand-instrument contact following [28] as $\mathcal{L}_{\text{inter}} = \lambda_R \mathbf{L}_R + \lambda_A \mathbf{L}_A$, where $\mathbf{L}_R$ and $\mathbf{L}_A$ are the attraction loss and the repulsion loss, which penalize the interpenetration between hand and instruments, and minimize the distance between the interacting hand and instrument in the possible contact region, respectively. We provide more details regarding the calculation in the Appendix. Finally, we refine the interaction by applying a modified version of Signed Distance Field (SDF) loss $\mathcal{L}_{\text{sdf}}$, for which we provide more details and an ablation study in D.2.

4 Baseline and Experiments

Our dataset can enable various downstream tasks for pose estimation and recognition. In this section, we introduce two benchmarks built upon our dataset. We first propose the evaluation protocols of

each benchmark and provide a detailed analysis of our dataset. Furthermore, we present baseline methods corresponding to the benchmarks with comparison against the state-of-the-art methods to demonstrate the effectiveness of our approach. More implementation details are provided in D.

4.1 Evaluation Protocol

Data Split. To ensure each phase has balanced samples, we split our dataset into training, validation, and test sets by subjects, which have 30, 3, 8 subjects separately. Based on the data split, bimanual hand pose estimation and (2) hand-instrument interactions. We provide more details regarding the data distribution and data quality analysis in the Appendix. Note that in our experiments on both benchmarks, we train the model on the monocular training images from all 8 views, including both egocentric and allocentric for rich supervision.

Evaluation Metrics. Our goal is to reconstruct accurate 3D motion of hands and instruments during complex surgical operations from video. Specifically, we propose metrics to quantify estimate quality and compare our baseline method against state-of-the-art hand-object pose approaches.
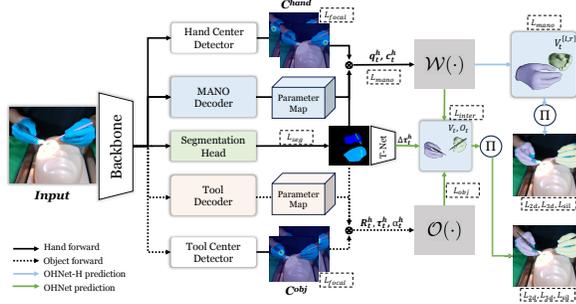

Figure 3: **Overview of the OH-Net.** The backbone image encoder outputs the image feature, which is then used to decode the hand/instrument centre heatmap and segmentation mask. MANO decoders predict their corresponding weak-perspective camera parameters. Decoupling the instrument branch forms **H-Net**.

- **Bimanual Hand Pose Estimation:** To evaluate the accuracy and plausibility of the hand reconstruction pipeline, we report the Mean Per Joint Positional Error (MPJPE), Mean Per Vertex Positional Error (MPVPE) in mm after root (hand wrist) joint alignment. To explicitly measure the relative translation error, we report Mean Relative Root Translation Error (MRRTE).
- **Hand-instrument Interaction:** To quantify the reconstruction quality, we report the same evaluation metrics (MPJPE, MPVPE, MRRTE) as in bimanual hand pose estimation benchmarks. For instrument pose estimation quality, we evaluate the commonly used ADD-S score, which measures the average distance between the model vertices transformed by the ground truth and the estimated poses, following [48, 41, 72, 76, 92]. Specifically, we report the percentage of the transformed instruments with a vertex positional error less than 10% of the instrument diameter. We further report the Mean Articulation Error (MAE) to evaluate the articulation of instruments, calculating the absolute error between the ground-truth articulation factor and the prediction in percentage, excluding the rigid instruments without articulation. For the interaction quality, we evaluate Mean Per Joint Positional Error between each hand-instrument interaction pair (MRRTEh, o) and Mean Inter-penetration Volume (Pen) in cm^3.

4.2 Bimanual Hand Motion Estimation

We now set up the benchmark for bimanual hand pose estimation from a monocular RGB input image. Acquiring accurate 3D hand pose is essential in the surgical scenarios during manipulation.

Parametric Representation. In the task of (monocular) bimanual hand pose estimation, our goal is to reconstruct the 3D pose of the two hands from an RGB input video. In order to obtain the detailed geometry of the two hands, we adopt the parametric model MANO [69] as our mesh representation to predict $\{\boldsymbol{\theta}_t^h, \boldsymbol{\beta}_t^h, \boldsymbol{\phi}_t^h, \boldsymbol{\tau}_t^h\}$ following the dataset settings. Given the MANO parameters, the 3D hand mesh V_t^h and the 21 hand keypoints

Table 2: **Quantitative evaluation results for bimanual hand motion estimation.** We compare our method with state-of-the-art hand reconstruction methods on local hand poses. MPJPE, MPVPE, and MRRTE are reported in millimeters (mm) after root alignment.

Method	Val			Test		
	MPJPE ↓	MPVPE ↓	MRRTE ↓	MPJPE ↓	MPVPE ↓	MRRTE ↓
DIR [68]	18.63	18.91	34.76	18.89	18.75	35.17
InterWild [55]	19.48	19.87	37.27	20.19	20.34	38.76
IntagHand [44]	18.92	17.96	32.43	19.38	19.16	32.77
ACR [90]	18.18	18.57	33.29	18.86	19.28	33.59
H-Net (w/o T-Net)	18.57	18.48	33.92	18.98	19.14	34.18
H-Net	17.39	18.72	31.66	17.66	18.66	31.89
H-Net-D	**15.28**	**16.37**	**26.86**	**15.97**	**16.18**	**26.59**

7

Table 3: **Quantitative evaluation results for two-hand-instrument interactions.** We compare our method with the state-of-the-art hand reconstruction methods on local hand poses.

Split	Method (mm)	MPJPE (mm) ↓	MPVPE (mm) ↓	MRRTE (mm) ↓	ADD-S (%) ↑	MAE (%) ↓	Pen (mm) ↓	MRRTE^{h,o} (mm) ↓
	Hasson et al. [26]	19.87	21.45	39.78	56.66	-	7.87	27.98
	HFL-Net [48]	17.51	18.27	37.43	58.64	-	7.06	25.67
Val	HOISDF [65]	17.05	18.22	34.36	60.91	-	5.69	24.13
	OH-Net (w/o T-Net)	17.48	18.33	32.61	66.87	14.83	6.41	21.86
	OH-Net	17.12	18.43	31.36	71.52	11.14	5.87	19.94
	OH-Net-D	15.23	16.41	26.78	76.68	9.67	5.14	17.33
	Hasson et al. [26]	20.04	21.33	40.56	56.89	-	7.76	28.69
	HFL-Net [48]	17.45	17.66	38.65	59.78	-	7.14	25.49
Test	HOISDF [65]	17.36	17.91	35.88	61.32	-	5.77	23.82
	OH-Net (w/o T-Net)	17.89	18.06	33.25	65.94	14.31	6.45	21.92
	OH-Net (Ours)	17.34	18.36	31.58	70.79	11.17	5.91	20.11
	OH-Net-D (Ours)	15.94	16.13	26.44	76.31	9.62	5.18	17.45

$\mathbf{J}_t^h$ can be regressed. We adopt the commonly used weak-perspective camera model following [61, 44, 90, 56, 94] to estimate the 3D translation.

Baseline. To address the problem of reconstructing bimanual hands from a monocular RGB(-D) image, we introduce H-Net as the baseline approach, without the instrument branch. As shown in Fig. 3, the model takes as input the image at timestep t to extract the image feature $f_t \in \mathbb{R}^{D \times H \times W}$ where D is the dimension of the feature map. Subsequently, the following 3 regression heads predict the segmentation mask $\mathbf{M}^h \in \mathbb{R}^{3 \times H \times W}$ (*i.e.* left hand, right hand, background), MANO parameter map $\mathbf{M}_t^{mano} \in \mathbb{R}^{218 \times H \times W}$, hand center heatmap $\mathbf{C}_t^{hand} \in \mathbb{R}^{2 \times H \times W}$ respectively. Leveraging the collision aware center-based representation [74, 90] for hands, we disentangle the bimanual hand features while pushing away the centers that are too close in the repulsion field. In the following, the MANO parameters q_t^h for each hand is extracted by combining with the Hand Center map and the instance segmentation mask. After obtaining the 3D mesh by MANO model $\mathcal{W}(\cdot)$ with keypoints, we use the output relative translation $\Delta\tau \in \mathbb{R}^3$ from T-Net to model the fine-grained relative transformation from the left hand to the right hand, incorporating the strong spatial features as prior knowledge. The weak-perspective camera parameter is represented as c_t^h. Moreover, we denote the RGB-D input based version as H-Net-D. Finally, the network is supervised by the weighted sum of the hand center loss and the mesh parameter loss:

$$\mathcal{L} = \lambda_{\text{focal}}\mathcal{L}_{focal} + \lambda_{\text{pj2d}}\mathcal{L}_{\text{pj2d}} + \lambda_{3\text{d}}\mathcal{L}_{3\text{d}} + \lambda_{\text{sil}}\mathcal{L}_{\text{sil}} + \mathcal{L}_{\text{mano}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} \quad (9)$$

where $\mathcal{L}_{focal}$ is the focal loss for the hand center map. $\mathcal{L}_{\text{mano}} = \lambda_\theta \mathcal{L}_{\boldsymbol{\theta}} + \lambda_\beta \mathcal{L}_{\boldsymbol{\beta}}$ is the weighted sum of L2 loss of the MANO parameters. We provide the training and implementation details in App. D.3.

Results. We evaluate the performance of our baseline models on surgical bimanual hand reconstruction tasks, and compare them against state-of-the-art hand pose estimation methods. As shown in Tab. 2, our method significantly outperforms prior approaches such as InterWild [55], DIR [68], and IntagHand [44], achieving the best overall performance across different metrics, including MPJPE, MPVPE, and MRRTE. These results highlight the importance of domain-specific design: our hand center detector and tailored parameterization improve robustness in surgical environments, where factors such as gloves, occlusions, and instrument-induced hand articulation pose unique challenges. Notably, the T-Net module contributes to finer pose refinement by learning spatial alignment from segmentation masks, leading to consistent improvements across both joint-wise and vertex-level metrics. These findings validate the effectiveness of our baseline in modeling surgical hands with high precision, serving as a strong foundation for subsequent hand-instrument reasoning.

4.3 Two-Hand-Instrument Interactions

In this section, we propose the benchmark for hand-instrument interaction, which aims to reconstruct the 3D mesh for two hands and the 6D pose for the surgical instruments.

Parametric Representation In terms of two-hand-instrument Interaction baseline, our task is to reconstruct the 3D meshes of the two hands as well as the 6D pose of the in-hand surgical instruments. We keep the hand representation as MANO [69] for consistency. For 6D instrument pose estimation, we leverage the parametric model $\mathcal{O}(\theta_t^o, \alpha_t)$ introduced in Sec. 3.2 to represent the surgical instruments with articulation. Specifically, the 3D mesh is regressed using the parameters of 6D pose $\theta_t^o \in \mathbb{R}^6$ and the 1D relative articulation factor $\alpha_t \in [0, 1]$ which controls the articulation.

Baseline. As discussed in Sec. 4.2, integrating the instrument 6D pose estimation branch forms the full model of OH-Net. As the first method to reconstruct two-hand-two-object, we propose to disentangle the features and mesh representation along with an explicit handler for interaction. As

illustrated in Fig. 3, the extracted feature map is followed by 5 regression heads, yielding hand center map $\mathbf{C}_t^{hand} \in \mathbb{R}^{2 \times H \times W}$ and object center map $\mathbf{C}_t^{obj} \in \mathbb{R}^{2 \times H \times W}$. The segmentation head predicts the instance mask $\mathbf{M}_t^{h,o} \in \mathbb{R}^{5 \times H \times W}$. We extend the collision-aware center-based representation mechanism to fit in the task and push away the two close instrument center and hand centers. The instrument parameter map is regressed as $\mathbf{M}^{obj} \in \mathbb{R}^{14 \times H \times W}$, which contains the 6D pose and 1D articulation factor for both instruments. Next, T-Net predicts $\Delta\tau \in \mathbb{R}^9$ for the relative translation between the left hand, right hand, and between their interacting instruments. OH-Net is supervised with the objectives below:

$$\mathcal{L} = \lambda_{\text{focal}}\mathcal{L}_{focal} + \lambda_{\text{pj2d}}\mathcal{L}_{\text{pj2d}} + \lambda_{\text{3d}}\mathcal{L}_{\text{3d}} + \lambda_{\text{sil}}\mathcal{L}_{\text{sil}} + \mathcal{L}_{\text{mano}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{obj}}\mathcal{L}_{\text{tool}} \quad (10)$$

where $\mathcal{L}_{\text{tool}}$ represents the loss function for instrument supervision. Specifically, $\mathcal{L}_{\text{tool}}$ is composed of $\mathcal{L}_R$ and $\mathcal{L}_\tau$ for the 6D instrument pose and $\mathcal{L}_\alpha$ for the 1D articulation. Moreover, $\mathcal{L}_{\text{pj2d}}$ and $\mathcal{L}_{\text{3d}}$ is also defined for instrument and optimized with pre-defined bounding box of the instrument mesh following [50]. We refer the readers to the appendix for detailed implementation details.

Results. We evaluate the proposed OH-Net framework on the hand-instrument interaction benchmark. As shown in Tab. 3, our method achieves state-of-the-art performance across all evaluated metrics, including MPJPE, MPVPE, MRRTE, and ADD-S. The full OH-Net benefits significantly from the joint modeling of hands and instruments, with joint training improving both interaction accuracy and articulation consistency. Notably, OH-Net is the first method capable of reconstructing two hands and two interacting instruments simultaneously, with explicit disentanglement and spatial reasoning. The RGB-D version, OH-Net-D, shows further gains across all metrics, especially in interaction-specific scores such as Penetration (Pen) and MRRTE^{h,o}. The MAE and ADD-S results highlight reliable articulation estimation and instrument localization. Moreover, Fig. 4 showcases qualitative examples of predicted meshes and multi-
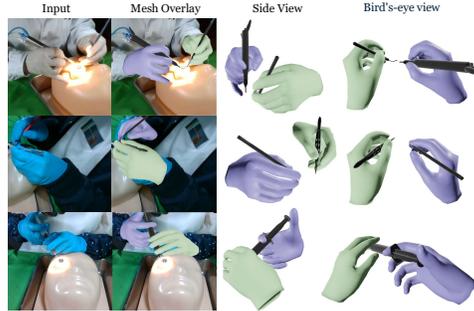

Figure 4: **Qualitative results on the Hand-instrument interaction benchmark.** Each row shows a sample from the test set, with columns displaying: (1) input image, (2) mesh prediction, (3) rendered mesh from a side view, and (4) bird's-eye view.

view renderings, demonstrating our model's ability to preserve hand-instrument contact and recover detailed interactions even under occlusions. Additional results and visualizations are provided in D.3.

5 Discussion

Related Work. Surgical vision research has recently evolved beyond traditional tasks [57, 33, 23] toward 3D perception, including scene reconstruction [93, 11], depth estimation [12, 87], and navigation [66, 86]. Despite advances in datasets like AVOS [21] and MM-OR [58], most systems remain limited to passive monitoring rather than real-time surgical assistance. Meanwhile, 3D hand reconstruction has progressed from general benchmarks [96, 56] to synthetic medical datasets [81, 3], though lacking clinical realism. Monocular approaches have evolved from single-hand models [13, 61] to two-hand systems [94, 46] and hand-object interactions [82, 89], using template-based methods [18, 25] and geometric constraints [5, 95]. Our work bridges the clinical gap with a targeted 3D perception framework for ophthalmic microsurgery using real RGB-D data capturing fine-grained, two-hand, multi-instrument interactions. See Tab. 1 and A for details.

Conclusion. In this paper, we introduce OphNet-3D, the first large-scale, real-world RGB-D dynamic 3D reconstruction dataset for ophthalmic microsurgery, comprising 41 sequences from 40 surgeons (7.1 M frames), annotated with 12 surgical phases, 10 instrument classes, detailed MANO hand meshes, and 6D instrument poses. We develop a multi-stage automatic annotation pipeline that integrates monocular hand-prior models, segmentation–point-cloud alignment, and biomechanics-based joint hand–instrument optimization. Leveraging this dataset, we establish two new benchmarks—bimanual hand pose estimation and hand–instrument interaction—and propose H-Net and OH-Net, which achieve state-of-the-art performance on all metrics ().

Limitations & Future Work. This study offers new insights into dynamic 3D reconstruction in ophthalmic surgery but has three main limitations: it relies on data from a single center (requiring

multi-center validation for better generalizability); strong microscope illumination causes instrument-tip overexposure (which could be addressed with synchronized motion-capture RGB or infrared imaging); and it has not yet integrated the microscope view for joint reconstruction of the ocular surface, hands, and instruments (may be a key direction for future work to boost clinical relevance).

6 Acknowledgements

References

[1] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol, R. Newcombe, R. Wang, J. J. Engel, and T. Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. *CVPR*, 2025.

[2] R. A. Bartholomew, H. Zhou, M. Boreel, K. Suresh, S. Gupta, M. B. Mitchell, C. Hong, S. E. Lee, T. R. Smith, J. P. Guenette, et al. Surgical navigation in the anterior skull base using 3-dimensional endoscopy and surface reconstruction. *JAMA Otolaryngology–Head & Neck Surgery*, 150(4):318–326, 2024.

[3] M. Birlo, R. Caramalau, P. J. Edwards, B. Dromey, M. J. Clarkson, D. Stoyanov, et al. Hup-3d: A 3d multi-view synthetic dataset for assisted-egocentric hand-ultrasound pose estimation. *arXiv preprint arXiv:2407.09215*, 2024.

[4] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283, 2020.

[5] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020.

[6] Z. Cao, I. Radosavovic, A. Kanazawa, and J. Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021.

[7] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9044–9053, 2021.

[8] P. Chen, W. Li, N. Gunderson, J. Ruthberg, R. Bly, Z. Sun, W. M. Abuzeid, and E. J. Seibel. Endoperfect: A hybrid nerf-stereo vision approach pioneering monocular depth estimation and 3d reconstruction in endoscopy, 2025.

[9] Y. Chen, Z. Tu, D. Kang, R. Chen, L. Bao, Z. Zhang, and J. Yuan. Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. *IEEE Transactions on Image Processing*, 30:4008–4021, 2021.

[10] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020.

[11] C. Cotsoglou, S. Granieri, S. Bassetto, V. Bagnardi, R. Pugliese, G. L. Grazi, A. Guglielmi, A. Ruzzenente, L. Aldrighetti, F. Ratti, et al. Dynamic surgical anatomy using 3d reconstruction technology in complex hepato-biliary surgery with vascular involvement. results from an international multicentric survey. *HPB*, 26(1):83–90, 2024.

[12] B. Cui, M. Islam, L. Bai, and H. Ren. Surgical-dino: adapter learning of foundation models for depth estimation in endoscopic surgery. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–8, 2024.

[13] H. Dong, A. Chharia, W. Gou, F. V. Carrasco, and F. De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *arXiv preprint arXiv:2407.09646*, 2024.

[14] E. Duran, M. Kocabas, V. Choutas, Z. Fan, and M. J. Black. Hmp: Hand motion priors for pose and shape estimation from video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6353–6363, 2024.

[15] P. Entezami, L. E. Franzblau, and K. C. Chung. Mentorship in surgical training: a systematic review. *Hand*, 7(1):30–36, 2012.

[16] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[17] R. Fujii, R. Hachiuma, H. Kajita, and H. Saito. Surgical tool detection in open surgery videos. *Applied Sciences*, 12(20):10473, 2022.

[18] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018.

[19] S. Geman. Statistical methods for tomographic image restoration. *Bull. Internat. Statist. Inst.*, 52:5–21, 1987.

[20] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018.

[21] E. D. Goodman, K. K. Patel, Y. Zhang, W. Locke, C. J. Kennedy, R. Mehrotra, S. Ren, M. Guan, O. Zohar, M. Downing, et al. Analyzing surgical technique in diverse open surgical videos with multitask machine learning. *JAMA surgery*, 159(2):185–192, 2024.

[22] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021.

[23] M. Grammatikopoulou, E. Flouty, A. Kadkhodamohammadi, G. Quellec, A. Chow, J. Nehme, I. Luengo, and D. Stoyanov. Cadis: Cataract dataset for surgical rgb-image segmentation. *Medical Image Analysis*, 71:102053, 2021.

[24] H. Hamer, J. Gall, T. Weise, and L. Van Gool. An object-dependent hand pose prior from sparse training data. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 671–678. IEEE, 2010.

[25] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020.

[26] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 571–580, 2020.

[27] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019.

[28] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019.

[29] M. Hayoz, C. Hahne, T. Kurmann, M. Allan, G. Beldi, D. Candinas, P. Márquez-Neila, and R. Sznitman. Online 3d reconstruction and dense tracking in endoscopic videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 444–454. Springer, 2024.

[30] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[31] J. Hein, M. Seibold, F. Bogo, M. Farshad, M. Pollefeys, P. Fürnstahl, and N. Navab. Towards markerless surgical tool and hand pose estimation. *International journal of computer assisted radiology and surgery*, 16:799–808, 2021.

[32] M. Hu, L. Wang, S. Yan, D. Ma, Q. Ren, P. Xia, W. Feng, P. Duan, L. Ju, and Z. Ge. Nurvid: A large expert-level video database for nursing procedure activity understanding. *Advances in Neural Information Processing Systems*, 36:18146–18164, 2023.

[33] M. Hu, P. Xia, L. Wang, S. Yan, F. Tang, Z. Xu, Y. Luo, K. Song, J. Leitner, X. Cheng, et al. Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding. *arXiv preprint arXiv:2406.07471*, 2024.

[34] M. Hu, K. Yuan, Y. Shen, F. Tang, X. Xu, L. Zhou, W. Li, Y. Chen, Z. Xu, Z. Peng, et al. Ophclip: Hierarchical retrieval-augmented learning for ophthalmic surgical video-language pretraining. *arXiv preprint arXiv:2411.15421*, 2024.

[35] D. Huang, X. Ji, X. He, J. Sun, T. He, Q. Shuai, W. Ouyang, and X. Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia Conference Proceedings*, 2022.

[36] K. Karunratanakul, J. Yang, Y. Zhang, M. J. Black, K. Muandet, and S. Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020.

[37] S. Khalid, M. Goldenberg, T. Grantcharov, B. Taati, and F. Rudzicz. Evaluation of deep learning models for identifying surgical actions and measuring performance. *JAMA network open*, 3(3):e201664–e201664, 2020.

[38] J. Kim, M.-G. Gwon, H. Park, H. Kwon, G.-M. Um, and W. Kim. Sampling is matter: Point-guided 3d human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 12880–12889, 2023.

[39] S. V. Kotsis and K. C. Chung. Application of the "see one, do one, teach one" concept in surgical training. *Plastic and reconstructive surgery*, 131(5):1194–1201, 2013.

[40] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, October 2021.

[41] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10138–10148, 2021.

[42] V. Lepetit, F. Moreno-Noguer, and P. Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81:155–166, 2009.

[43] C. Li, Y. Tong, Y. Long, W. Si, D. C. M. Yeung, J. Y.-K. Chan, and Q. Dou. Extended reality with hmd-assisted guidance and console 3d overlay for robotic surgery remote mentoring. *IEEE Robotics and Automation Letters*, 2024.

[44] M. Li, L. An, H. Zhang, L. Wu, F. Chen, T. Yu, and Y. Liu. Interacting attention graph for single image two-hand reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2761–2770, 2022.

[45] M. Li, H. Zhang, Y. Zhang, R. Shao, T. Yu, and Y. Liu. Hhmr: Holistic hand mesh recovery by enhancing the multimodal controllability of graph diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 645–654, 2024.

[46] D. Lin, Y. Zhang, M. Li, Y. Liu, W. Jing, Q. Yan, Q. Wang, and H. Zhang. 4dhands: Reconstructing interactive hands in 4d with transformers. *arXiv preprint arXiv:2405.20330*, 2024.

[47] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[48] Z. Lin, C. Ding, H. Yao, Z. Kuang, and S. Huang. Harmonious feature learning for interactive hand-object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12989–12998, 2023.

[49] D. Liu, Q. Li, T. Jiang, Y. Wang, R. Miao, F. Shan, and Z. Li. Towards unified surgical skill assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9522–9531, 2021.

[50] S. Liu, H. Jiang, J. Xu, S. Liu, and X. Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021.

[51] X. Liu, A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath. Dense depth estimation in monocular endoscopy with self-supervised learning methods. *IEEE Transactions on Medical Imaging*, 39(5):1438–1447, 2020.

[52] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[53] G. Manni, C. Lauretti, F. Prata, R. Papalia, L. Zollo, and P. Soda. Bodyslam: A generalized monocular visual slam framework for surgical applications. *arXiv preprint arXiv:2408.03078*, 2024.

[54] R. Q. Mao, L. Lan, J. Kay, R. Lohre, O. R. Ayeni, D. P. Goel, et al. Immersive virtual reality for surgical training: a systematic review. *Journal of Surgical Research*, 268:40–58, 2021.

[55] G. Moon. Bringing inputs to shared domains for 3d interacting hands recovery in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17028–17037, 2023.

[56] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020.

[57] C. I. Nwoye, T. Yu, C. Gonzalez, B. Seeliger, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.

[58] E. Özsoy, C. Pellegrini, T. Czempiel, F. Tristram, K. Yuan, D. Bani-Harouni, U. Eck, B. Busam, M. Keicher, and N. Navab. Mm-or: A large multimodal operating room dataset for semantic understanding of high-intensity surgical environments. *arXiv preprint arXiv:2503.02579*, 2025.

[59] Y. L. Pang, C. Oh, and A. Cavallaro. Sparse multi-view hand-object reconstruction for unseen environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–810, 2024.

[60] A. Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[61] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024.

[62] T.-H. Pham, N. Kyriazis, A. A. Argyros, and A. Kheddar. Hand-object contact force estimation from markerless visual tracking. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2883–2896, 2017.

[63] J. D. Pitcher, J. T. Wilson, T.-C. Tsao, S. D. Schwartz, and J.-P. Hubschman. Robotic eye surgery: past, present, and future. *J Comput Sci Syst Biol*, 3(1):137, 2012.

[64] A. Prakash, M. Chang, M. Jin, R. Tu, and S. Gupta. 3d reconstruction of objects in hands without real world 3d supervision. In *European Conference on Computer Vision*, pages 126–145. Springer, 2024.

[65] H. Qi, C. Zhao, M. Salzmann, and A. Mathis. Hoisdf: Constraining 3d hand-object pose estimation with global signed distance fields. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10392–10402. IEEE, 2024.

[66] L. Qiu and H. Ren. Endoscope navigation with slam-based registration to computed tomography for transoral surgery. *International Journal of Intelligent Robotics and Applications*, 4(2):252–263, 2020.

[67] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[68] P. Ren, C. Wen, X. Zheng, Z. Xue, H. Sun, Q. Qi, J. Wang, and J. Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8014–8025, 2023.

[69] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017.

[70] D. Shan, J. Geng, M. Shu, and D. F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.

[71] A. Spurr, U. Iqbal, P. Molchanov, O. Hilliges, and J. Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European conference on computer vision*, pages 211–228. Springer, 2020.

[72] S. Stevsic and O. Hilliges. Spatial attention improves iterative 6d object pose estimation. In *2020 international conference on 3D vision (3DV)*, pages 1070–1078. IEEE, 2020.

[73] T. Sugiyama, S. Lama, L. S. Gan, Y. Maddahi, K. Zareinia, and G. R. Sutherland. Forces of tool-tissue interaction to assess surgical skill level. *JAMA surgery*, 153(3):234–242, 2018.

[74] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, and T. Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11179–11188, 2021.

[75] S. Tamai. History of microsurgery—from the beginning until the end of the 1970s. *Microsurgery*, 14(1):6–13, 1993.

[76] B. Tekin, F. Bogo, and M. Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4511–4520, 2019.

[77] Q. Tian, Z. Chen, H. Liao, X. Huang, L. Li, S. Ourselin, and H. Liu. Endoomni: Zero-shot cross-dataset depth estimation in endoscopy by robust self-learning from noisy labels. *arXiv preprint arXiv:2409.05442*, 2024.

[78] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118:172–193, 2016.

[79] P. Vaid, S. Yeung, and A. Rau. Robust semi-supervised detection of hands in diverse open surgery environments. In K. Deshpande, M. Fiterau, S. Joshi, Z. Lipton, R. Ranganath, I. Urteaga, and S. Yeung, editors, *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pages 736–753. PMLR, 11–12 Aug 2023.

[80] H. Wang, Y. Long, Y. Chen, H.-C. Yip, M. Scheppach, P. W.-Y. Chiu, Y. Yam, H. M.-L. Meng, and Q. Dou. Learning dissection trajectories from expert surgical videos via imitation learning with equivariant diffusion. *Medical Image Analysis*, 103:103599, 2025.

[81] R. Wang, S. Ktistakis, S. Zhang, M. Meboldt, and Q. Lohmeyer. Pov-surgery: A dataset for egocentric hand and tool pose estimation during surgical activities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 440–450. Springer, 2023.

[82] J. Wu, G. Pavlakos, G. Gkioxari, and J. Malik. Reconstructing hand-held objects in 3d. *arXiv preprint arXiv:2404.06507*, 2024.

[83] Y. Xu, J. Zhang, Q. Zhang, and D. Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022.

[84] J. Yang, J. Li, G. Li, Z. Shen, H.-Y. Wu, Z. Fan, and H. Huang. Mlphand: Real time multi-view 3d hand mesh reconstruction via mlp modeling. *arXiv preprint arXiv:2406.16137*, 2024.

[85] L. Yang, J. Xu, L. Zhong, X. Zhan, Z. Wang, K. Wu, and C. Lu. Poem: reconstructing hand in a point embedded multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21108–21117, 2023.

[86] Y. Yang, L. Guoliang, Q. Li, and R. Song. A slam framework based spinal endoscopic localization method. *Procedia Computer Science*, 250:81–87, 2024.

[87] Z. Yang, J. Pan, J. Dai, Z. Sun, and Y. Xiao. Self-supervised endoscopy depth estimation framework with clip-guidance segmentation. *Biomedical Signal Processing and Control*, 95:106410, 2024.

[88] Y. Ye, A. Gupta, and S. Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3895–3905, 2022.

[89] Y. Ye, P. Hebbar, A. Gupta, and S. Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19717–19728, 2023.

[90] Z. Yu, S. Huang, F. Chen, T. P. Breckon, and J. Wang. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[91] Z. Yu, S. Zafeiriou, and T. Birdal. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. *arXiv preprint arXiv:2412.12861*, 2024.

[92] S. Zakharov, I. Shugurov, and S. Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1941–1950, 2019.

[93] R. Zha, X. Cheng, H. Li, M. Harandi, and Z. Ge. Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In *International conference on medical image computing and computer-assisted intervention*, pages 13–23. Springer, 2023.

[94] B. Zhang, Y. Wang, X. Deng, Y. Zhang, P. Tan, C. Ma, and H. Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11354–11363, 2021.

[95] J. Y. Zhang, S. Pepose, H. Joo, D. Ramanan, J. Malik, and A. Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 34–51. Springer, 2020.

[96] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019.

A   Related Work

Surgical 3D Perception. Beyond conventional tasks in surgical assistance such as phase recognition [57, 33, 32, 34], anatomical segmentation [4, 23] and instrument detection [34, 57], recent efforts have increasingly focused on 3D perception for open surgical environments, addressing challenges in scene reconstruction [93, 11, 29, 8], depth estimation [12, 87, 77, 51], navigation [66, 2, 86, 53], and skill assessment [49, 21]. A growing body of work leverages multi-view camera arrays, RGB-D sensors, and IMUs to reconstruct dense surgical scenes with both static anatomical structures and dynamic hand-tool interactions. Goodman et al. [21] proposed AVOS, a large-scale annotated video dataset of open surgeries, and developed a multitask model to extract procedural signatures and quantify surgeon skill from real-world surgical videos. MM-OR [58] introduces a comprehensive multimodal operating room dataset featuring RGB-D, audio, speech transcripts, and robotic logs, annotated with panoptic segmentation and semantic scene graphs. While these systems demonstrate technical progress on specific tasks and in multimodal fusion capabilities, their clinical impact remains limited, with most systems designed primarily for passive action recognition or offline documentation. Few are optimized for real-time use or tailored to support specific training objectives in microsurgical procedures. In this work, we address this gap by developing a clinically-oriented 3D perception framework for ophthalmic microsurgery, integrating task-specific data, model design, and system-level considerations for real-world deployment.

3D Hand Dataset. Recent 3D hand reconstruction datasets have driven progress in hand pose, object pose, and interaction reconstruction (see Tab. 1). General benchmarks such as FreiHAND [96] and InterHand2.6M [56] provide multi-view real RGB images for reliable 3D pose recovery. ContactPose [5] further leverages RGB-D to capture detailed contact patterns. Banerjee et al. present HOT3D [1], a large-scale egocentric, multi-view hand–object interaction (HOI) dataset comprising ≈1.5 M synchronized frames captured with head-mounted Aria and Quest 3 cameras across 19 participants manipulating 33 objects; each frame is paired with mocap-grade 3D hand and object poses. Medical datasets extend these ideas to surgery. Hein et al. [31] released a synthetic surgical RGB set with limited views and simple actions. POV-Surgery [81] and HUP-3D [3] enrich viewpoint count and interaction diversity, yet their synthetic origin still curbs realism. Ophthalmic surgery, notably cataract removal, follows fixed workflows but requires both hands to manipulate several tools in a confined field—for instance, holding an iris retractor while guiding capsulorhexis forceps. These dense, dual-hand motions strain current reconstruction methods. OphNet-3D addresses this gap with a real RGB-D clinical dataset that records fine-grained, two-hand, multi-instrument interactions during live ophthalmic operations.

Monocular 3D Hand Mesh Reconstruction. Compared with multi-view approaches, monocular reconstruction is more practical in clinical theatres: a single, non-contact RGB camera minimises equipment, preserves sterility, and avoids obstructing the surgeon's workspace. Research has progressed from single-hand models [13, 45, 61, 38, 90] to two-hand systems [94, 90, 46, 68] and, most recently, HOI reconstruction [82, 64, 35, 89]. Monocular HOI remains difficult because occlusion is severe and annotated data are scarce. Many studies therefore assume known instance-specific templates [18, 24, 25]; with the template in place, object recovery reduces to 6D pose estimation and joint hand–object pose inference. Joint reasoning is implemented via implicit feature fusion [9, 20, 50, 70, 76], explicit geometric constraints such as contact or collision [5, 6, 10, 22, 95], or physics-based consistency [62, 78]. To lift the template assumption, newer work directly predicts object shape—either as explicit genus-0 meshes [28] or via a joint hand–object implicit field [36]. Conditional reconstruction strategies further exploit hand joint cues to refine object geometry [88].

B   OphNet-3D Construction

B.1   Synchronized Recording Configuration

We capture using the Intel RealSense SDK and synchronize all cameras with the official software-trigger method. Because we simultaneously acquire and store high-resolution RGB and depth streams from eight cameras, the system's USB bandwidth and I/O performance are pushed to their limits, and sustained data throughput may exceed its capacity, causing occasional frame drops. Therefore, on the hardware side, we've configured a high-core-count host machine, enterprise-grade storage drives with high write speeds, and fiber-optic USB cables to ensure stable data transmission.

In high-frequency image I/O scenarios, traditional storage formats such as JPEG and PNG—despite their widespread compatibility and ease of visualization—incur compression artifacts and processing overhead that can significantly degrade overall system performance. To address these limitations and improve data throughput, we adopt a binary-format–based image encoding approach that directly serializes cv::Mat matrix data into raw binary files. Specifically, our method first writes essential matrix metadata (rows, columns, and cv::Mat type) as integers in the file header to ensure accurate reconstruction of the image's dimensions and layout; it then appends the unaltered pixel byte stream to the file body without any compression or encoding conversion. During decoding, the file is read in the same order to restore the original cv::Mat structure exactly. In our implementation, color frames are stored as 848 × 480 three-channel matrices of type CV_8UC3, while depth frames are 848 × 480 single-channel matrices of 16-bit unsigned integers (UINT16), with each color–depth pair spatially aligned. For acquisition, the main thread initializes eight camera-capture subthreads; upon receiving a capture command, image saving begins and the preview display is disabled to conserve resources, whereas the preview is re-enabled when capture is inactive.
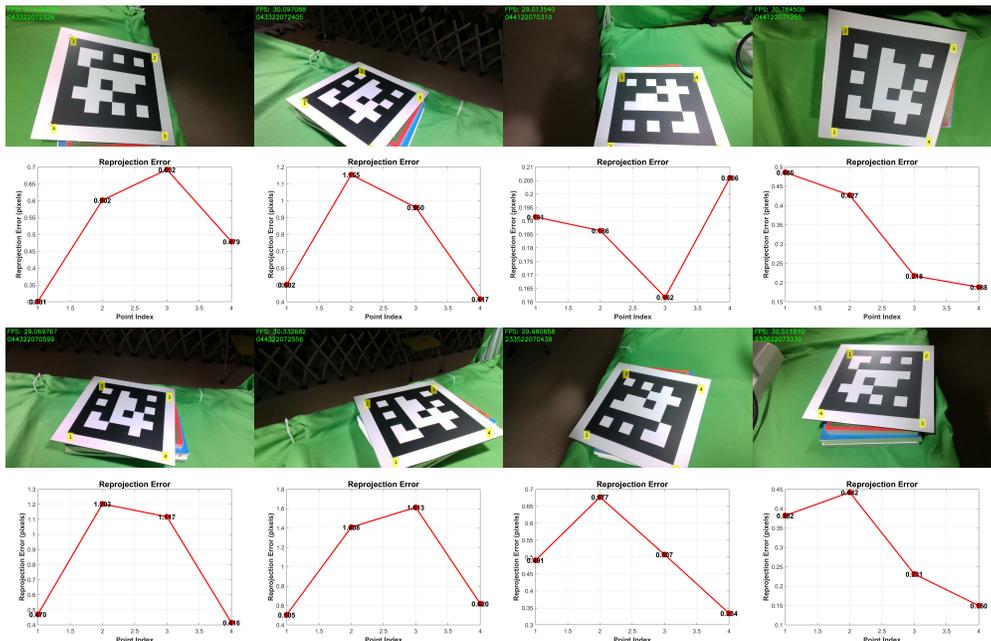

Figure 5: Synchronized calibration of 8 cameras.

B.2 Synchronized Calibration

In our multi-camera calibration pipeline, each Intel RealSense D435's intrinsic parameters—including focal lengths, principal point coordinates and lens distortion coefficients—are retrieved at runtime via the official SDK and assembled into the 3×3 camera matrix K; for extrinsic calibration, a planar ArUco marker board of precisely known marker size and layout serves as the world reference, and for each synchronized color capture a simple green-channel mask first isolates the board region, from which the 2D corner coordinates of each detected ArUco marker are paired with their exact 3D positions on the board plane. A non-linear Perspective-n-Point solver then computes the rigid-body rotation and translation that best align the 3D marker positions to their 2D observations, yielding a 4×4 homogeneous transform from board (world) to camera coordinates. During the recording process, due to force majeure the platform was moved three times. After each move, we performed a recalibration; Fig. 5 shows one such calibration example.

B.3 Phase Definition and Demonstration

Referring to the standard cataract surgical procedure, we divided the entire surgical process into 12 phases: *(1) main incision creation, (2) viscoelastic injection, (3) paracentesis, (4) capsulorhexis, (5) hydrodissection and hydrodelineation, (6) pre-phaco, (7) nucleus sculpting, (8) nucleus cracking, (9)*

17

nuclear rotation, (10) phacoemulsification, (11) cortex removal, and (12) incision hydration. Detailed definitions of each phase and the instruments used are listed in Tab. 4. Fig. 10 shows the 12 phases from two different camera viewpoints.

Index	Phase Label	Phase Definition	Instruments Used
1	main incision creation	Use keratome to create a precise entry point through the cornea or limbus to access the anterior chamber of the eye.	toothed forceps, keratome blade
2	viscoelastic injection	Injection of ophthalmic viscoelastic devices (OVDs) into the anterior chamber using a viscoelastic syringe to maintain chamber depth, protect intraocular tissues, and facilitate subsequent surgical steps.	viscoelastic syringe
3	paracentesis	Creation of a small, self-sealing side-port incision at the limbus using a 15° stab blade to allow access for second instruments such as the chopper into the anterior chamber.	15° stab blade, toothed forceps
4	capsulorhexis	Creating a continuous curvilinear opening in the anterior lens capsule using a capsulorhexis forceps to allow safe access to the lens nucleus for removal.	iris repositor, capsulorhexis forceps
5	hydrodissection and hydrodelineation	Inject balanced salt solution (BSS) using a 10ml syringe to separate the lens nucleus from the cortex and capsule, allowing easier rotation and removal.	10 mL syringe
6	pre-phaco	Use a phacoemulsfication handpiece gently remove loose cortical or epinuclear material from the anterior lens surface before nucleus sculpting begins.	nucleus chopper, phacoemulsification handpiece
7	nucleus sculpting	Using a phacoemulsification handpiece to carve grooves into the lens nucleus to facilitate nucleus division and removal.	nucleus chopper, phacoemulsification handpiece
8	nucleus cracking	Using the phacoemulsification handpiece in combination with a nucleus chopper to mechanically split the grooved nucleus into smaller fragments for easier phacoemulsification and removal.	nucleus chopper, phacoemulsification handpiece
9	nuclear rotation	Using a nucleus chopper to gently rotate the lens nucleus within the capsular bag, ensuring optimal positioning for continued phacoemulsification.	nucleus chopper, phacoemulsification handpiece
10	phacoemulsification	Involves using a phacoemulsification handpiece with ultrasonic tip to break up and emulsify the nucleus fragments, while simultaneously aspirating the lens material and maintaining anterior chamber stability.	nucleus chopper, phacoemulsification handpiece
11	cortex removal (I/A)	Use an irrigation-aspiration handpiece to gently remove the residual cortical material from the capsular bag following phacoemulsification of the lens nucleus.	iris repositor, irrigation-aspiration handpiece
12	incision hydration	Using a balanced salt solution (BSS) through a 10mL syringe to swell the corneal stroma at the incision (both main incision and paracentesis) edges, sealing the surgical incisions at the end of cataract surgery.	10 mL syringe

Table 4: Definition of each phase and the instruments used in each phase

B.4 Phase Localization Annotation

We performed phase–boundary annotation on each video sequence as follows. First, because the microscope and hand-view recordings were acquired on separate devices and thus lack intrinsic temporal synchronization, we aligned them manually. Immediately before each trial, a rigid printed marker bearing the legend "Start Recording" was displayed concurrently across all camera views; the frame in which this marker first appeared in each view served as the synchronization point. Next, an experienced ophthalmologist reviewed the temporally aligned microscope and hand-view videos to delineate the onset and offset of each surgical phase. To improve label purity, segments corrupted by visual noise—such as instrument exchanges during which the hand left the camera field of view—were excluded. A second ophthalmologist then independently verified and revised all boundary annotations. Finally, to ensure precise demarcation of phase transitions, we uniformly contracted each annotated interval by removing one second from both its start and end.

B.5 Instrument Demonstration

During the procedures, 10 different surgical instruments were employed: *(1) 15° stab blade, (2) keratome blade, (3) iris repositor, (4) nucleus chopper, (5) toothed forceps, (6) capsulorhexis forceps, (7) phacoemulsification handpiece, (8) irrigation-aspiration handpiece, (9) 10 mL syringe, and (10) viscoelastic syringe.* Fig. 6 shows photographs of the ten instruments used during surgery, and Fig. 16 presents the 3D model scan files. For the forceps, we scanned both the open and closed configurations. For the syringe, we scanned two states—plunger rod at its maximum and minimum extension—and additionally scanned the plunger rod and the syringe body separately as individual components.

C OphNet-3D Statistics

Tab. 5 reports, for each split of OphNet-3D, the number of video clips and the distribution of frame counts per surgical phase. Since some phase operations may involve a third hand entering the scene, we remove those segments during annotation; this can truncate a complete phase into multiple shorter clips. Such clips are counted separately in our statistics but are linked via sequential index identifiers.

Figure 6: 10 different instruments. From left to right they are: capsulorhexis forceps, viscoelastic syringe, toothed forceps, iris repositor, nucleus chopper, 15° stab blade, keratome blade, irrigation-aspiration handpiece, phacoemulsification handpiece, and 10 mL syringe.

Phase Label	No. of Clips				No. of Frames			
	train	val	test	all	train	val	test	all
main incision creation	32	3	8	43	312,224	21,600	48,960	382,784
viscoelastic injection	59	8	20	87	519,120	59,040	132,960	711,120
paracentesis	35	4	13	52	196,080	13,440	71,520	281,040
capsulorhexis	49	4	15	68	1,116,960	105,360	292,080	1,514,400
hydrodissection and hydrodelineation	30	4	8	42	321,864	21,120	79,920	422,904
pre-phaco	31	4	8	43	435,600	48,960	99,600	584,160
nucleus sculpting	30	3	8	41	367,920	20,400	106,080	494,400
nucleus cracking	29	3	8	40	127,200	8,160	39,840	175,200
nuclear rotation	12	1	5	18	51,120	2,640	31,440	85,200
phacoemulsification	29	4	8	41	556,080	46,320	237,600	840,000
cortex removal (I/A)	33	3	10	46	751,832	98,640	290,640	1,141,112
incision hydration	32	4	8	44	406,168	22,560	80,672	509,400
all	401	45	119	565	4,955,272	468,240	1,511,312	7,141,720

Table 5: Phase distribution across splits for clips and frames.

D More Implementation Details and Results

D.1 Data processing

Hand Mesh Initialization. To initialize the per-frame hand motion for each view, we adopt a three-stage process inspired by DynHaMR [91], incorporating 2D detection fusion, per-view tracking, and global fusion.

We first extract 2D hand keypoints by applying ViTPose [83] to each RGB frame. Keypoints below a confidence threshold of $\epsilon_j = 0.5$ are discarded. Cropped patches around detected hand regions are then reprocessed through ViTPose for local refinement. To address unreliable or missing joints, we additionally apply MediaPipe [52] and fuse the results: for each joint, we retain the ViTPose prediction if above threshold, and replace it with MediaPipe's output otherwise. If entire hands are undetected in some views, we infill the missing detections by copying hand motion from nearby frames with high visibility and smoothing their trajectory with a temporal window. This approach

ensures a full joint set $\hat{\mathbf{J}}_t^h$ per frame for each view. To reduce noise from hallucinated detections or wong handedness, we adopt a filtering strategy. For each frame, we keep only the bounding box with the highest IoU (> 0.9) among overlapping detections and discard those that appear in fewer than 10 frames across the sequence. Additionally, we track bounding box continuity to detect erroneous handedness flips or duplicated hands — if a bounding box IoU with the previous frame drops below 0.1, we mark the frame as invalid and exclude it from subsequent fitting. These invalid frames are later recovered via generative infilling.

Next, we estimate the 3D hand pose per view using the coarse-to-fine regression pipeline of [61], which returns MANO parameters $\{\boldsymbol{\theta}_t^h, \boldsymbol{\beta}_t^h, \boldsymbol{\phi}_t^h, \boldsymbol{\tau}_t^h\}$. The 3D wrist translation $\boldsymbol{\tau}_t^h$ is obtained via depth sampling and back-projection:

$$x = \frac{z(u - c_x)}{f_x}, \quad y = \frac{z(v - c_y)}{f_y},$$

where (u, v) are 2D keypoints, and (f_x, f_y) are focal lengths. We choose the optimal z minimizing reprojection error to initialize depth.

Finally, per-view MANO parameters are transformed to world coordinates using known camera extrinsics $\{\mathbf{R}_i, \mathbf{t}_i\}$, and merged across views using a weighted average. View weights are derived from the per-frame visibility scores computed from 2D keypoint confidence. This results in a globally consistent, temporally smooth initialization of hand pose across all frames and views. The resulting motion serves as the input to our multi-stage RGB-D optimization pipeline Sec. 3.2, where temporal, geometric, and interaction constraints are jointly optimized.

D.2 Annotation pipeline

Implementation details. We implement the annotation pipelien with PyTorch [60]. During the optimization of stage II and stage III (Sec. 3.2), we use L-BFGS algorithm with $lr = 1$ and optimizing the loss functions using below weights:

- For stage II, we have: $\lambda_{2d} = 0.001, \lambda_{\mathrm{smooth}} = 10, \lambda_{\boldsymbol{\theta}} = 0.04, \lambda_{\boldsymbol{\beta}} = 0.05$.
- For stage III, we have: $\lambda_z = 200, \lambda_\phi = 2, \lambda_\gamma = 10, \lambda_{pen} = 10, \lambda_\beta = 0.05, \lambda_{\mathrm{ja}} = 1, \lambda_{\mathrm{palm}} = 1, \lambda_{\mathrm{bl}} = 1$.

To better model the hand plausibility, we propose to leverage a biomechanical constraints and an angle limitation constraint to our objective function:

$$\mathcal{L}_{\mathrm{ja}} = \sum_j d_{\alpha, H}(\boldsymbol{\alpha}_t^j, \mathbf{H}^j), \quad \mathcal{L}_{\mathrm{bl}} = \sum_j \mathcal{I}(\|\mathbf{b}_t^j\|_2; b_{\min}^j, b_{\max}^j), \tag{11}$$

$$\mathcal{L}_{\mathrm{palm}} = \sum_j \mathcal{I}(\|\mathbf{c}_t^j\|_2; c_{\min}^j, c_{\max}^j) + \sum_j \mathcal{I}(\|\mathbf{d}_t^j\|_2; d_{\min}^j, d_{\max}^j), \tag{12}$$

$$\mathcal{L}_{\mathrm{angle}} = \|\hat{\theta}_t^h\|_2 + \mathcal{I}(\|\hat{\theta}_t^h\|_2; \theta_{\min}^h, \theta_{\max}^h) + \mathcal{I}(\|\hat{\theta}_t^b\|_2; \theta_{\min}^b, \theta_{\max}^b), \tag{13}$$

where j is the index of the hand joint. $\mathcal{L}_{\mathrm{ja}}$ is for joint angle priors that constrains the joint angle sequence $\boldsymbol{\alpha}_t^j = (\boldsymbol{\alpha}_t^f, \boldsymbol{\alpha}_t^a)$ by approximating the convex hull on $(\boldsymbol{\alpha}_t^f, \boldsymbol{\alpha}_t^a)$ plane with the point set $\mathbf{H}^j$. $\mathcal{I}(\cdot)$ is the interval loss that penalizes outliers. $\mathcal{L}_{\mathrm{bl}}$ represents the loss term for bone length penalizing the finger bone length b_j that lie outside valid bone length range $[b_{\min}^j, b_{\max}^j]$. Similarly, we further constrain the curvature $\|\mathbf{c}_t^j\|_2$ and angular distance $\|\mathbf{d}_t^j\|_2$ for the palm root bones, $\mathcal{L}_{\mathrm{palm}}$ by penalizing the outliers if the ranges $[c_{\min}^j, c_{\max}^j]$ and $[d_{\min}^j, d_{\max}^j]$. Moreover, we constrain a specific subset of hand poses $\hat{\theta}_t^h$ (e.g. twist rotation of Distal Interphalangeal (DIP) joints) and penalize the outliers of the pre-defined range $[\theta_{\min}^b, \theta_{\max}^b]$.

MANO Regularization. We regularize the predicted MANO parameters during optimization using a prior on both pose and shape:

$$\mathcal{L}_{\mathrm{mano}} = \lambda_\theta \mathcal{L}_{\boldsymbol{\theta}} + \lambda_\beta \mathcal{L}_{\boldsymbol{\beta}}. \tag{14}$$

The pose regularization term $\mathcal{L}_{\boldsymbol{\theta}}$ penalizes deviations from a rest pose (assumed to be all-zero) using an ℓ_2 norm over the pose parameters:

$$\mathcal{L}_{\boldsymbol{\theta}} = \sum_{h \in \{l, r\}} \sum_{t=0}^{T} \|\boldsymbol{\theta}_t^h\|_2^2. \tag{15}$$

The shape prior $\mathcal{L}_{\beta}$ similarly penalizes the shape coefficients β_t^h, encouraging plausible hand geometry:

$$\mathcal{L}_{\beta} = \sum_{h \in \{l,r\}} \|\beta^h\|_2^2. \tag{16}$$

These terms serve as soft constraints that prevent drift during optimization and help enforce physical realism.

Interaction Loss. To model physical plausibility and guide the relative spatial arrangement of the hand and tool, we incorporate an interaction loss $\mathcal{L}_{\text{inter}}$ comprising an attraction loss $\mathcal{L}_{\text{A}}$ and a repulsion loss $\mathcal{L}_{\text{R}}$, following prior work [28]. Specifically, the attraction term encourages contact between hand and object surfaces when interaction is expected, while the repulsion term penalizes interpenetration. We define the set of hand contact vertices $\mathcal{C}_{\text{ext}}^h$ by computing the proximity of each MANO vertex on the hand to the object mesh. For each frame, we mark as contact those hand vertices within $5\,\text{mm}$ of the object surface. From this set, we extract six regions of contact based on anatomical structure: five fingertips and the palm base, following [28]. These six anchor regions provide soft guidance for maintaining realistic contact.

The attraction loss is computed from the set of anchor points on the hand and their closest points on the object mesh V_{obj}:

$$\mathcal{L}_{\text{A}}(V_{\text{obj}}, V_{\text{hand}}) = \sum_{i=1}^{6} \mathcal{L}_{\text{dist}}(d(C_i^h(\text{Ext}(\text{Obj})), V_{\text{obj}})), \tag{17}$$

where $C_i^h(\text{Ext}(\text{Obj}))$ are the hand anchor vertices corresponding to region i, and $\mathcal{L}_{\text{dist}}$ is a distance loss (e.g., L1) between those vertices and the nearest points on the object. We define the distance loss $\mathcal{L}_{\text{dist}}$ as the average Euclidean distance between a set of hand contact vertices and their nearest neighbors on the object mesh:

$$\mathcal{L}_{\text{dist}}(C, V_{\text{obj}}) = \frac{1}{|C|} \sum_{\mathbf{v} \in C} \min_{\mathbf{u} \in V_{\text{obj}}} \|\mathbf{v} - \mathbf{u}\|_2, \tag{18}$$

where C is the set of contact vertices on the hand and V_{obj} is the object mesh. To discourage unnatural interpenetration, we define a repulsion loss $\mathcal{L}_{\text{R}}$ between the hand vertices and the inside of the object:

$$\mathcal{L}_{\text{R}}(V_{\text{obj}}, V_{\text{hand}}) = \sum_{\mathbf{v}_i \in V_{\text{hand}}} \mathbb{1}_{\text{in}}(\mathbf{v}_i) \cdot d(\mathbf{v}_i, V_{\text{obj}}), \tag{19}$$

where $\mathbb{1}_{\text{in}}(\cdot)$ is an indicator function marking vertices that lie inside the object, and $d(\cdot, V_{\text{obj}})$ is the shortest distance to the object surface. The final interaction loss is a weighted combination:

$$\mathcal{L}_{\text{inter}} = \lambda_{\text{R}}\mathcal{L}_{\text{R}} + (1 - \lambda_{\text{R}})\mathcal{L}_{\text{A}}, \tag{20}$$

where $\lambda_{\text{R}} = 0.5$ balances repulsion and attraction. Following [28], we empirically set $\lambda_{\text{R}} = 1.0$ during early training to resolve interpenetration first, then reduce it to allow attraction.

This loss encourages anatomically plausible contact while suppressing mesh collisions, improving the realism of hand-tool interaction.

Signed Distance Field Loss $\mathcal{L}_{\text{sdf}}$. To penalize interpenetration between the hand mesh and the tool surface, we adopt a signed distance field loss that queries the SDF defined over the tool volume. For each time step, we precompute a voxelized signed distance field $\phi_t^o(\cdot)$ around the tool mesh O_t^h. Then, the SDF loss is defined as:

$$\mathcal{L}_{\text{sdf}} = \sum_{\mathbf{v} \in V_t^h} \max(0, -\phi_t^o(\mathbf{v}))^2, \tag{21}$$

where V_t^h is the hand mesh and $\phi_t^o(\mathbf{v})$ returns the signed distance of a hand vertex $\mathbf{v}$ to the tool surface — negative values indicate penetration. The $\max(0, \cdot)^2$ term ensures that only intrusions (i.e., where $\phi < 0$) are penalized, encouraging the hand to remain outside the tool surface.

Runtime Our network is agnostic to the initialization method and is not restricted to using [61], which affects the processing time. Therefore we conduct runtime experiment excluding the stage I (hand and instance mask initialization time) and Pyrender offscreen rendering, which are not included in the optimization pipeline. On an NVIDIA A100 GPU the optimization pipeline can take 15 minutes for a video with 1000 frames, The stage II (the initialization of instrument 6D pose and hand pose optimization) takes around 10.9 minutes to process due to the bottle neck in ICP processing and registration. Finally, the last stage of joint optimization only takes around 4.1 minutes.

D.3 Baselines and Experiments

Implementation details. We implement our baseline methods based on PyTorch [60]. We use ResNet-50 [30] as the backbone network. All the input image and segmentation maps are resized to 512×512 while keeping the same aspect ratio with 0 paddings, which are then used to extract the feature maps $f \in \mathcal{R}^{(D+2)\times H \times W}$ with CoordConv []. We train our network using 1 A100 GPU with batchsize of 64. The size of our backbone feature is 128×128 and the size of our 4 pixel-aligned output maps is 64×64. We applied random scale, rotation, flip, and colour jitter augmentation during training.

Loss functions. We supervise our baseline models using a weighted sum of losses that account for 2D keypoint projection, 3D reconstruction accuracy, silhouette alignment, segmentation, and parameter regression. The total loss is formulated as:

$$\mathcal{L} = \lambda_{\text{focal}}\mathcal{L}_{\text{focal}} + \lambda_{\text{pj2d}}\mathcal{L}_{\text{pj2D}} + \lambda_{\text{3d}}\mathcal{L}_{\text{3d}} + \lambda_{\text{sil}}\mathcal{L}_{\text{sil}} + \mathcal{L}_{\text{MANO}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{obj}}\mathcal{L}_{\text{tool}}. \quad (22)$$

$\mathcal{L}_{\text{focal}}$ is focal loss [47] used to supervise the predicted hand and object center heatmaps. The projection loss $\mathcal{L}_{\text{pj2D}}$ penalizes the re-projection error between the predicted 3D keypoints (via MANO) and the 2D annotations using a robust Geman-McClure function. The 3D loss $\mathcal{L}_{\text{3d}}$ measures the vertex-to-surface distance between the predicted mesh and the observed point cloud from multi-view fusion. The $\mathcal{L}_{\text{pj2D}}$ and $\mathcal{L}_{\text{3d}}$ are also computed for the instrument vertex and pre-defined 3D bounding box around it. The silhouette loss $\mathcal{L}_{\text{sil}}$ compares the predicted hand and tool silhouette masks (from a differentiable renderer) against the ground-truth masks to enforce pixel-wise consistency. The MANO loss $\mathcal{L}_{\text{mano}}$ is composed of L2 losses over the predicted hand pose and shape parameters:

$$\mathcal{L}_{\text{mano}} = \lambda_\theta \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 + \lambda_\beta \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2, \quad (23)$$

where $\boldsymbol{\theta}^*$ and $\boldsymbol{\beta}^*$ denote pseudo ground-truth values from the annotation pipeline. The segmentation loss $\mathcal{L}_{\text{seg}}$ is a pixel-wise cross-entropy loss over the hand and tool instance masks. The tool loss $\mathcal{L}_{\text{tool}}$ supervises both the 6D pose and the 1D articulation parameters via parameter map regression and point cloud alignment. We use the following weights in all experiments: $\lambda_{\text{focal}} = 80$, $\lambda_{\text{pj2d}} = 400$, $\lambda_{\text{3d}} = 300$, $\lambda_{\text{sil}} = 50$, $\lambda_\theta = 80$, $\lambda_\beta = 10$, $\lambda_{\text{seg}} = 160$.

Figure 7: **Qualitative results on the hand pose estimation benchmark.** Each image is an overlay from each camera view.

D.4 More Qualitative Results

In this section, we provide qualitative visualizations of our model predictions on the Hand-Instrument Interaction benchmark for different phases. Each row in Fig. 8 illustrates three temporally adjacent frames from representative video clips, capturing various surgical manipulation phases and interaction types. For each frame, we show: (1) the input RGB image, (2) the mesh overlay with predicted hand and instrument meshes, and two alternative views to highlight the spatial relationship between hands and tools. These results demonstrate that our method generates consistent, physically plausible reconstructions across frames despite visual challenges such as occlusion, rapid tool motion, and complex hand articulation. The visual continuity across time confirms that our model not only produces accurate per-frame predictions but also maintains coherent temporal behavior, which is essential for understanding fine-grained surgical actions.

E Discussion

Limitation. While our dataset and method offer new insights into dynamic 3D reconstruction in ophthalmic surgery, several limitations remain. First, the data collection was conducted based on a single surgical procedure and within a single-center setting, potentially limiting the generalizability to datasets with varying surgical workflows, surgeon-specific operational habits, and illumination conditions; future work will expand to multi-center studies. Second, due to strong illumination from the surgical microscope, some instrument tips are overexposed in RGB views, affecting visibility and downstream pose estimation. Incorporating mocap-synchronized RGB capture or infrared cameras may help mitigate this issue. Third, we have not yet explored integrating the microscope view for joint reconstruction of the ocular surface, hands, and instruments, which may enable more clinically meaningful applications.
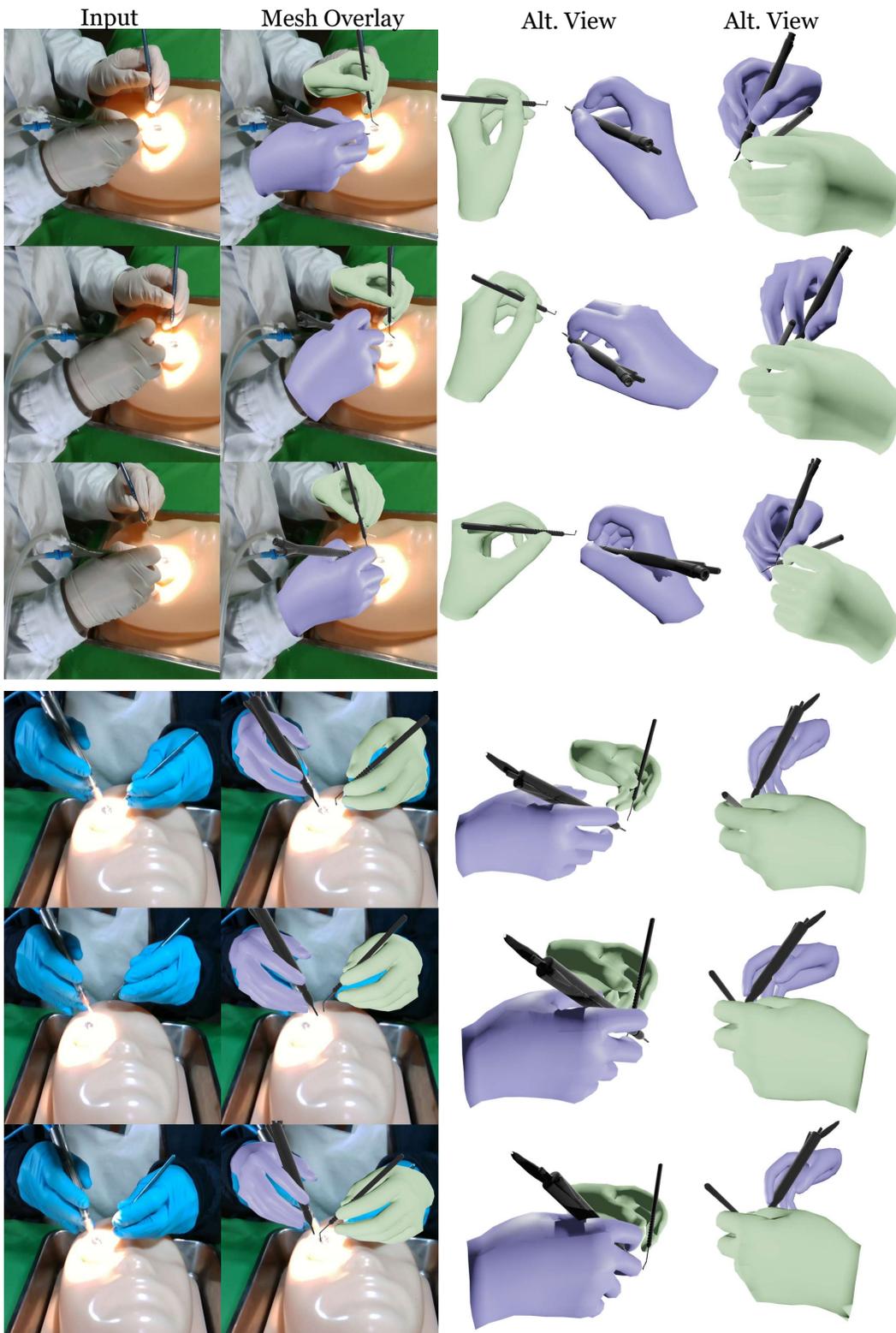
Figure 8: **Qualitative results on the hand-instrument interaction benchmark.** Each row shows a sample from the test set, with columns displaying: (1) input RGB image, (2) mesh overlay prediction, and (3)(4) for alternative view.
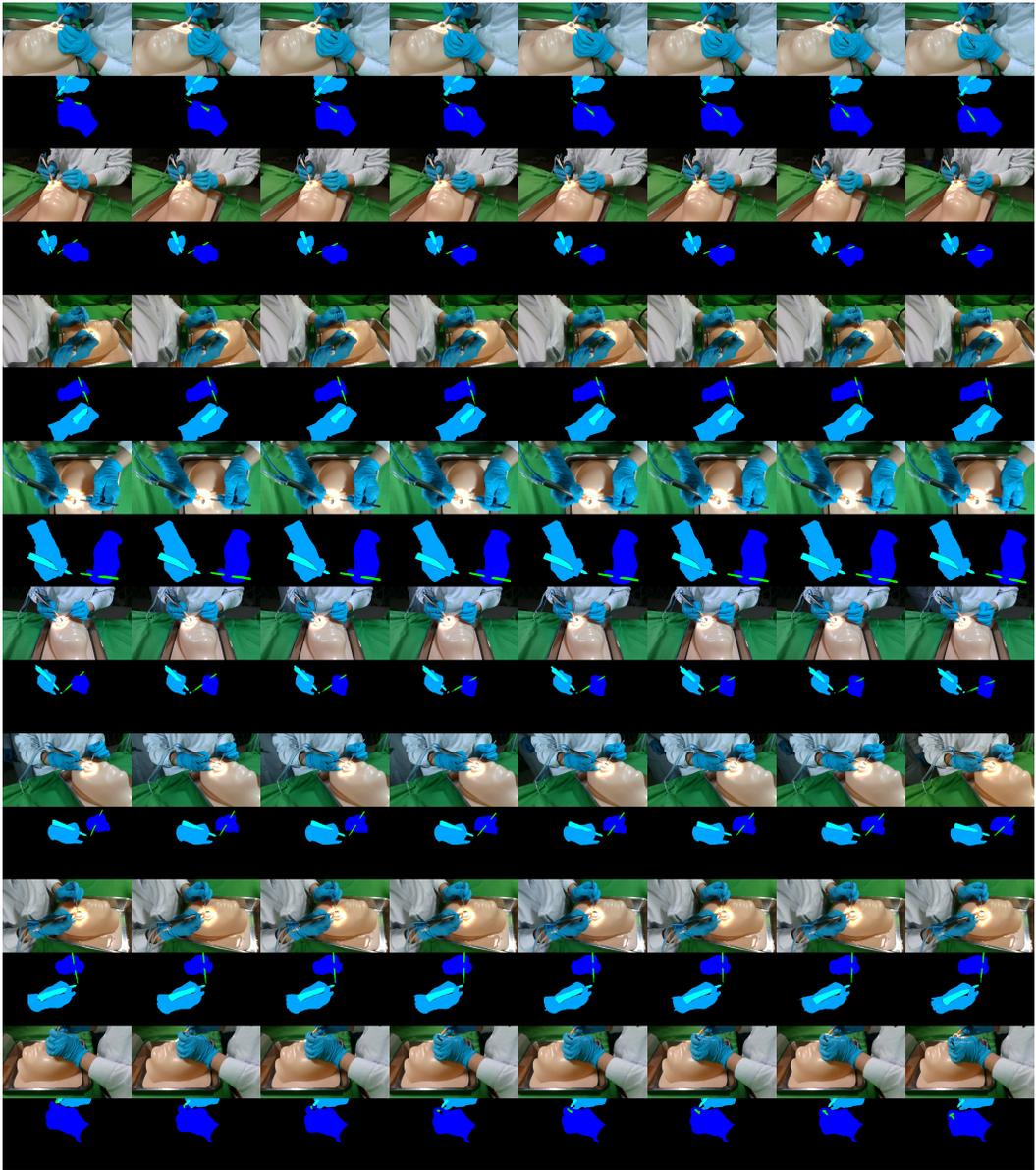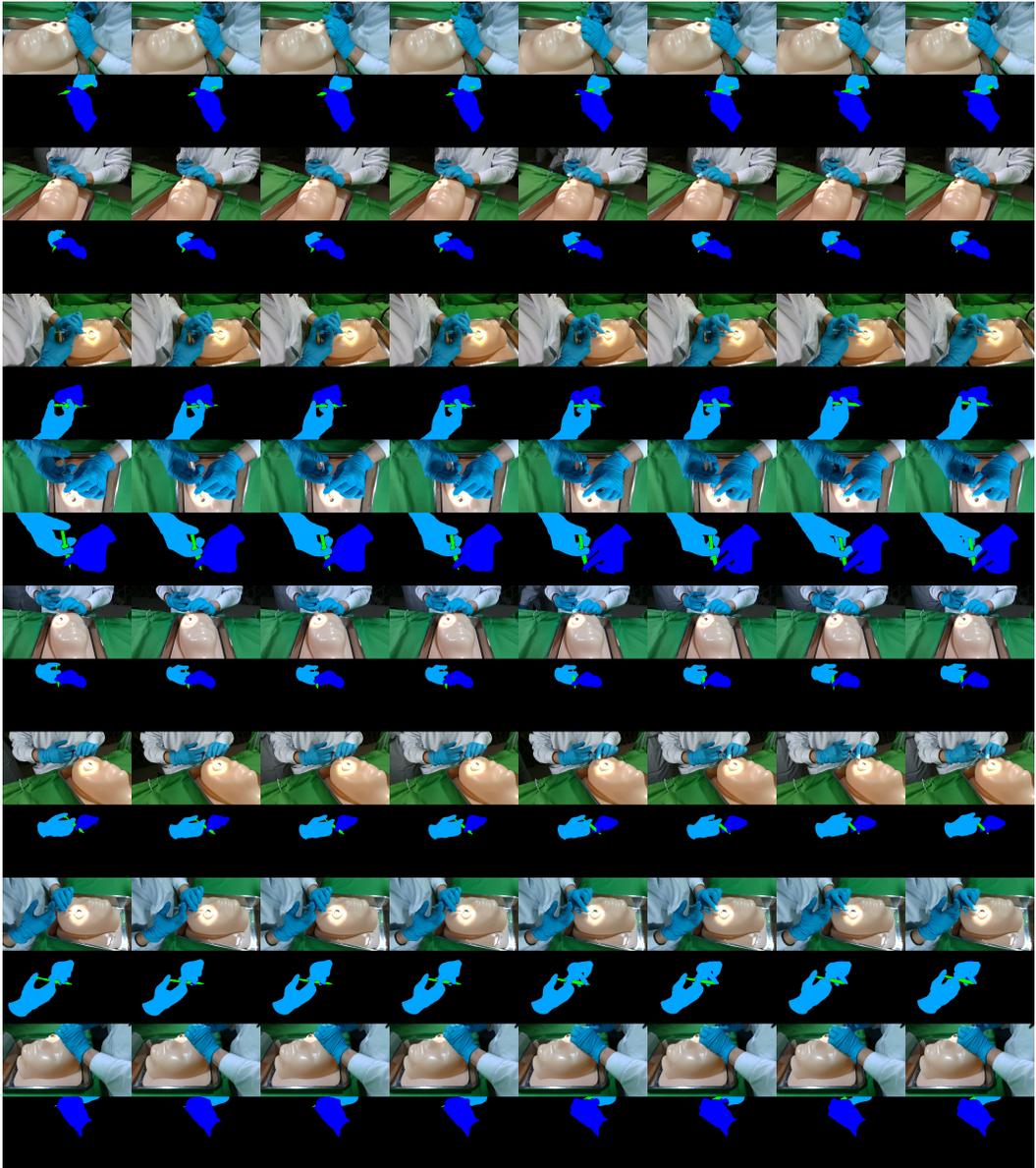
Figure 9: **Qualitative results on the hand-instrument interaction benchmark.** Each row shows a sample from the test set, with columns displaying: (1) input RGB image, (2) mesh overlay prediction, and (3)(4) for alternative view.

main incision creation

viscoelastic injection

paracentesis

capsulorhexis

hydrodissection and hydrodelineation

pre-phaco

nucleus sculpting

nucleus cracking

nuclear rotation

phacoemulsification

cortex removal (I/A)

incision hydration

main incision creation

viscoelastic injection

paracentesis

capsulorhexis

hydrodissection and hydrodelineation

pre-phaco

nucleus sculpting

nucleus cracking

nuclear rotation

phacoemulsification

cortex removal (I/A)

incision hydration
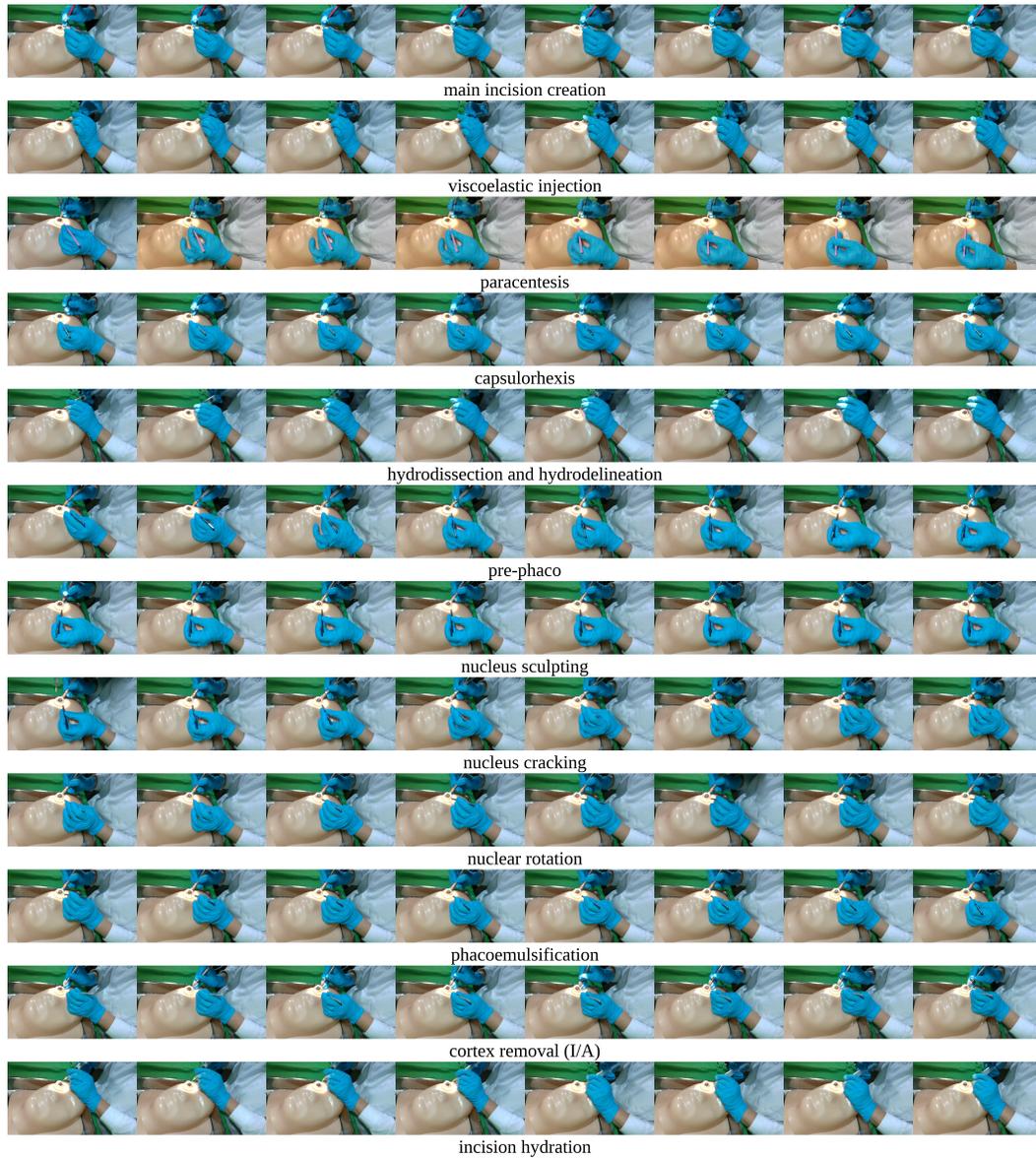
Figure 10: 12 phases from 2 different views.

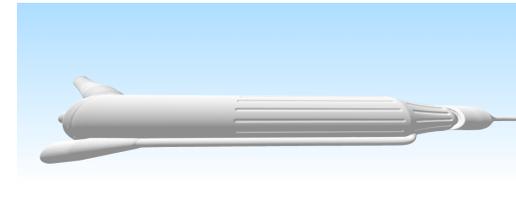Figure 11: Instance mask examples for phacoemulsification and viscoelastic injection.
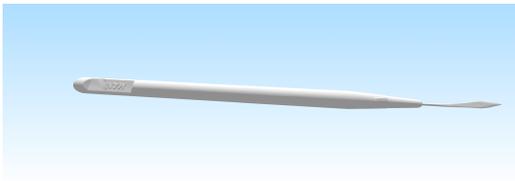
keratome blade
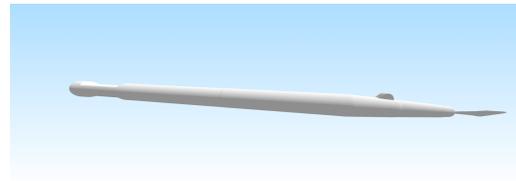
15° stab blade
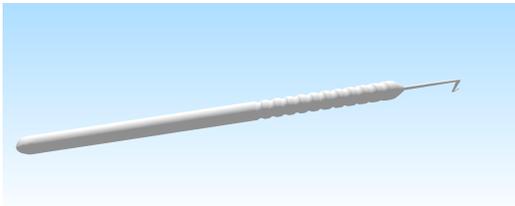
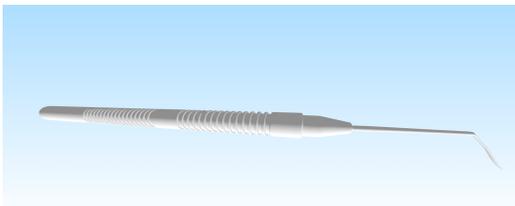nucleus chopper
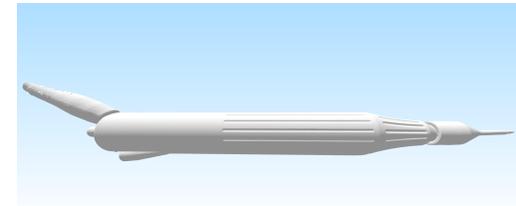
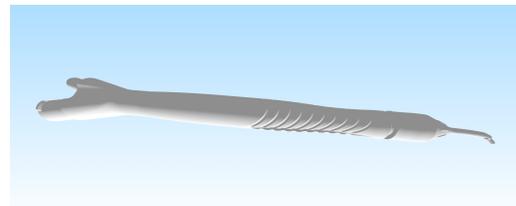phacoemulsification handpiece
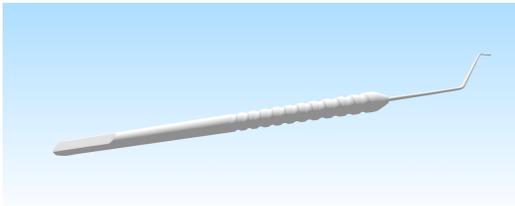
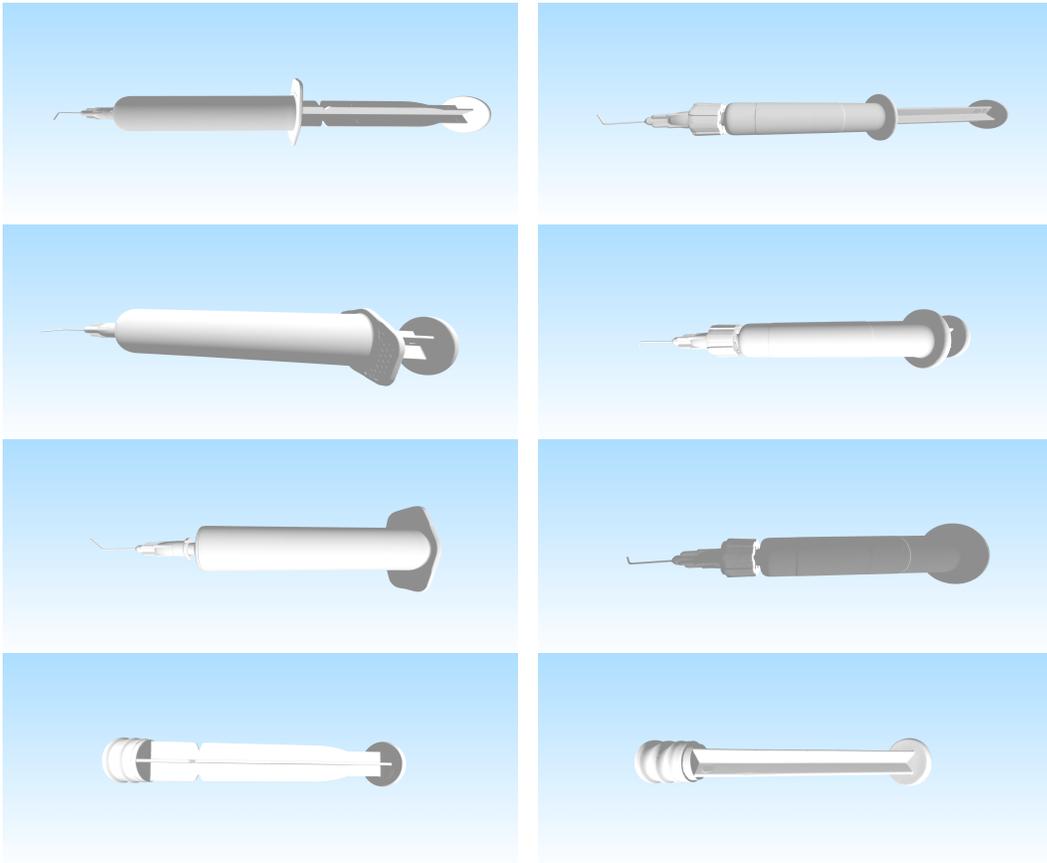iris repositor

irrigation-aspiration handpiece.

toothed forceps

capsulorhexis forceps

10 mL syringe   viscoelastic syringe

Figure 16: Scanned model files for 10 types of instruments.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, clearly outline our contributions to data collection, data annotation, and methodological innovation.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Sec. 5 and E, we analyze the current limitations and their corresponding solutions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For the data annotation pipeline and methodology, we provide clear mathematical proofs and citations.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We supplement each experiment with detailed model descriptions and experimental parameter settings to ensure reproducibility, and all datasets and code will be open-sourced.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Once the paper is accepted, we will open-source all data, code, and model weights.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We explain the data splits and some training and validation details in the main text, and provide more detailed experimental setup information in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We employ a fair evaluation protocol and metrics, and discuss the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide details of the model training platform and training time in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This study adheres to the NeurIPS ethical guidelines.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: This work has the potential to substantially benefit the field of ophthalmic surgery by enabling accurate, real-time 3D understanding of complex hand-instrument interactions. Our dataset and methods can facilitate objective skill assessment, personalized surgical training, and intraoperative guidance, potentially improving surgical outcomes and reducing the need for extensive expert supervision.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Our data and code will be hosted on major open-source platforms such as Hugging Face and GitHub.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: This study does not involve any data or code that require prior permission.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide visualizations for some of the datasets and ensure they are anonymized.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: In the data collection section, we describe the participants' basic information.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.