CARE3R: CONTACT-AWARE FEED-FORWARD 3D RECONSTRUCTION FOR SURGICAL ENDOSCOPY

Anonymous authors
Paper under double-blind review

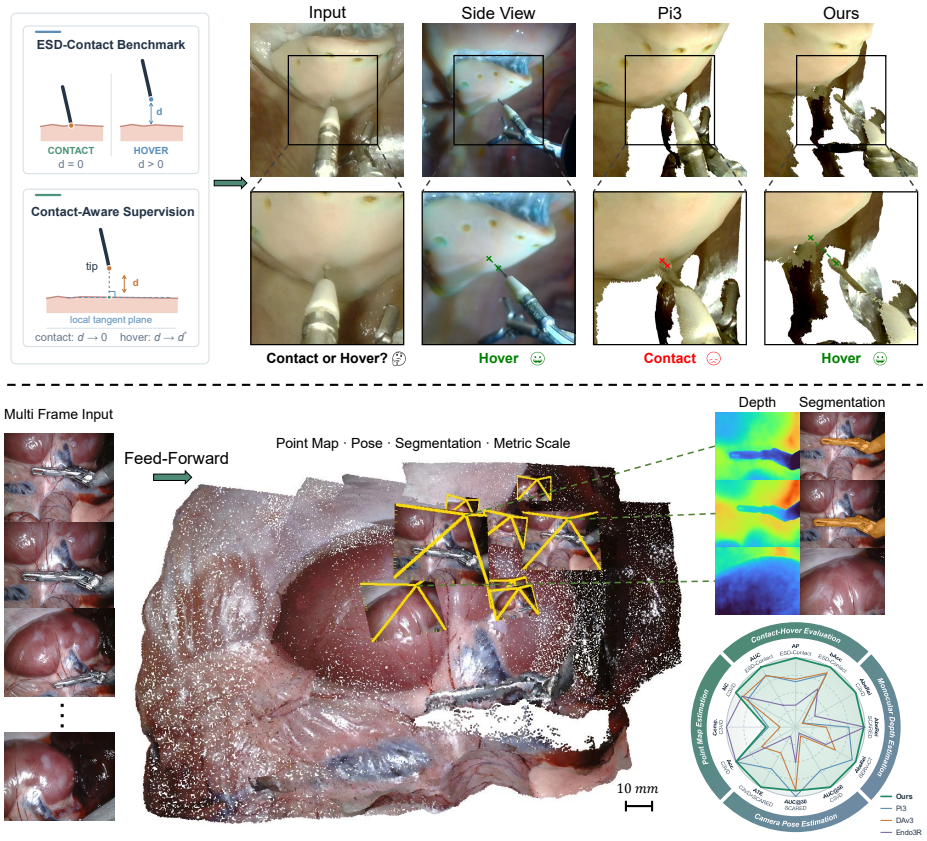


Figure 1: **CARE3R overview.** Trained with our multimodal ESD-Contact dataset and contact-aware supervision, CARE3R keeps the instrument tip at a physically plausible clearance above the tissue in the illustrated hover frame, where a general 3D model places it on the surface. From RGB frames alone, it predicts metric-scale point clouds, camera poses, and instrument segmentation in one feed-forward pass. The radar chart shows stronger contact discrimination and competitive conventional geometry metrics compared with state-of-the-art methods.

ABSTRACT

Three-dimensional reconstruction from endoscopic video supports surgical planning, intraoperative navigation, augmented visualization, and robotic assistance. Yet a reconstruction can be accurate pixel by pixel and still place the instrument tip at the wrong distance from the tissue. We show that this instrument–tissue clearance is dissociated from standard geometry metrics: three variants of our model trained with different instrument supervision have near-identical depth, pose, and reconstruction metrics, including the same tissue depth error on our stereo validation clips, yet their contact AUCs range from 0.74 to 0.93; and fine-tuning the prediction heads of Depth Anything 3 on our training data reduces its C3VD

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

depth error sevenfold without improving contact discrimination. To make clearance measurable and learnable, we introduce ESD-CONTACT, a multimodal ex-vivo endoscopic submucosal dissection (ESD) dataset with synchronized stereo and side views, verified contact/hover labels, and resistance recordings for a subset of clips, together with a contact benchmark that scores contact/hover discrimination from primary-view RGB. We further propose Contact-Aware Supervision (CAS), which supervises the instrument with independently measured targets rather than from the model's own instrument prediction: a stereo-referenced depth profile along the instrument, and perpendicular tip-to-tissue targets that a triangle construction derives from the side view. CARE3R, trained with CAS jointly across public and in-house datasets, predicts metric-scale depth, pointmaps, camera poses, and instrument segmentation from RGB frames in one feed-forward pass. On the validation clips of ESD-CONTACT it reaches a contact AUC of 0.933, compared with at most 0.806 for seven general and surgical 3D baselines, while remaining competitive on conventional depth, pose, reconstruction, and segmentation benchmarks. Code and weights will be released upon acceptance.

1 INTRODUCTION

Three-dimensional perception from endoscopic video is central to computer-assisted minimally invasive surgery. Depth, camera motion, and reconstructed surfaces support surgical planning, intraoperative navigation, augmented visualization, and robotic automation (Guo et al., 2025; Allan et al., 2021; Bobrow et al., 2023; Cui et al., 2024). Many of these uses depend on where the instrument is relative to the tissue, not only on the shape of either. A reliable estimate of instrument–tissue clearance could inform contact detection, proximity warnings, and geometric feedback for robotic assistance (Attanasio et al., 2021; Gruijthuijsen et al., 2018).

Standard metrics do not capture this clearance. Depth, pose, and reconstruction errors are averaged over the whole frame or trajectory, whereas clearance depends on a few millimetres between the tip and the tissue beneath it. We find that the two are dissociated. Three variants of our model trained with different instrument supervision have near-identical depth, pose, and reconstruction errors, and even the same tissue depth error on the frames being scored, yet their contact AUCs range from 0.74 to 0.93. Fine-tuning the prediction heads of Depth Anything 3 (Lin et al., 2025) on the same training data as our model cuts its C3VD depth error sevenfold without improving contact discrimination, and general 3D foundation models (Wang et al., 2025b;a; Lin et al., 2025) tend to compress the clearance of a hovering instrument towards zero, rendering it as touching the tissue. The cause is structural: a dense depth loss is dominated by pixels away from the tip, so without explicit supervision the tip can sink towards the tissue at almost no cost.

Existing work does not address this gap. Endoscopic depth estimation has progressed through self-supervision and domain adaptation (Shao et al., 2021; Rodriguez-Puigvert et al., 2023; Cui et al., 2024; Tian et al., 2024; Zhou et al., 2025; Shao et al., 2026), scale-aware and online reconstruction (Recasens et al., 2021; Guo et al., 2025; Cui et al., 2025), and neural or Gaussian scene representations (Liu et al., 2024; Xie et al., 2024; Huang et al., 2025). General feed-forward models (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025a;b; Lin et al., 2025) recover multi-view geometry in one pass but face a large domain gap in surgical scenes. None of them supervises or evaluates instrument–tissue clearance, and, to our knowledge, no public dataset provides the labels to do so. Clearance is also hard to label from the endoscope itself, where it is a millimetre-scale depth difference between a thin, specular tip and the tissue behind it; a side view roughly orthogonal to the endoscope axis turns it into a lateral displacement that is much easier to measure.

We address both measurement and learning. We first introduce ESD-CONTACT, which records ex-vivo endoscopic submucosal dissection (ESD) on porcine stomachs with a synchronized stereo endoscope and an auxiliary side-view camera, and provides verified contact/hover labels, resistance recordings for instrumented clips, stereo depth, instrument masks, tip locations, and tissue landmarks. On its validation clips we define a contact benchmark that scores contact/hover discrimination from primary-view RGB. We then propose Contact-Aware Supervision (CAS), a mesh-free supervision that constrains the instrument only with independently measured targets, never with its own prediction of the instrument, which would move with the tip. A profile term aligns the predicted

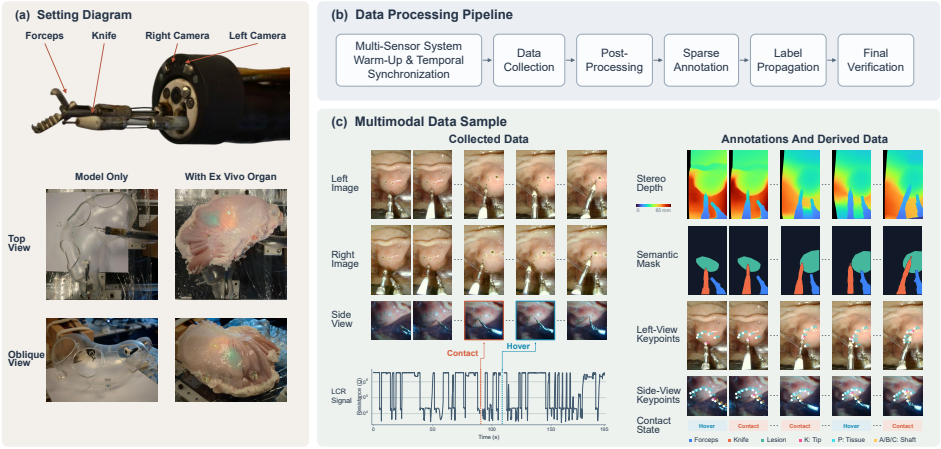


Figure 2: **ESD-Contact Dataset and Benchmark overview.** (a) Multimodal acquisition combines synchronized stereo and side-view imaging with resistance sensing on instrumented clips. (b) The processing pipeline aligns the recordings, propagates annotations, and verifies the labels. (c) Multimodal examples combine images, depth, masks, keypoints, and interaction labels to support our new contact benchmark.

depth profile along the instrument with the stereo profile, anchoring the tip to the measured shape of the instrument without any interaction labels. Contact and hover terms pull the tip onto a local tissue plane in contact frames and, in hover frames, toward a perpendicular clearance above the plane that a triangle construction derives from the side view. With CAS and joint training on public and in-house data, we obtain CARE3R, a feed-forward model that predicts metric-scale depth, pointmaps, camera poses, and instrument segmentation from primary-view RGB; side views are used only to construct training targets (Figure 1).

Our contributions are threefold:

- We identify and quantify a dissociation between instrument–tissue clearance and per-pixel geometric accuracy: models with near-identical depth, pose, and reconstruction metrics, and even the same tissue depth error on the frames being scored, can differ widely in contact discrimination, so standard metrics cannot reveal this failure.
- We introduce ESD-CONTACT, a multimodal endoscopic ESD dataset with synchronized stereo and side views, verified contact/hover labels, and local geometric annotations, together with a contact benchmark on its validation clips that measures contact/hover discrimination from primary-view RGB.
- We propose CAS, a mesh-free supervision built on independently measured instrument targets, and CARE3R, to our knowledge the first feed-forward surgical reconstruction model trained with explicit contact and hover supervision; it improves clearance discrimination while remaining competitive on conventional geometry benchmarks.

2 ESD-CONTACT DATASET AND BENCHMARK

We introduce ESD-CONTACT (*A Synchronized Stereo–Side-View Dataset for Instrument–Tissue Contact in Endoscopic Submucosal Dissection*), recorded during ex-vivo ESD on porcine stomachs while the knife is moved repeatedly between contact and hover (Figure 2). Its 14 clips (8,241 frames) pair a primary endoscope, stereo for 13 clips and monocular for one, with a synchronized side-view camera roughly orthogonal to the endoscope axis; three clips also record the knife–tissue electrical resistance. Frames are annotated with primary-view instrument masks, tip and tissue keypoints in both views, and an interaction state judged on the side view. Annotators label sparse frames, the labels are propagated through the video and manually reviewed, and on the three sensor clips the resistance trace provides initial states that an annotator verifies frame by frame. Primary-view depth is computed from the stereo pairs with FoundationStereo (FS) (Wen et al., 2025). These modalities

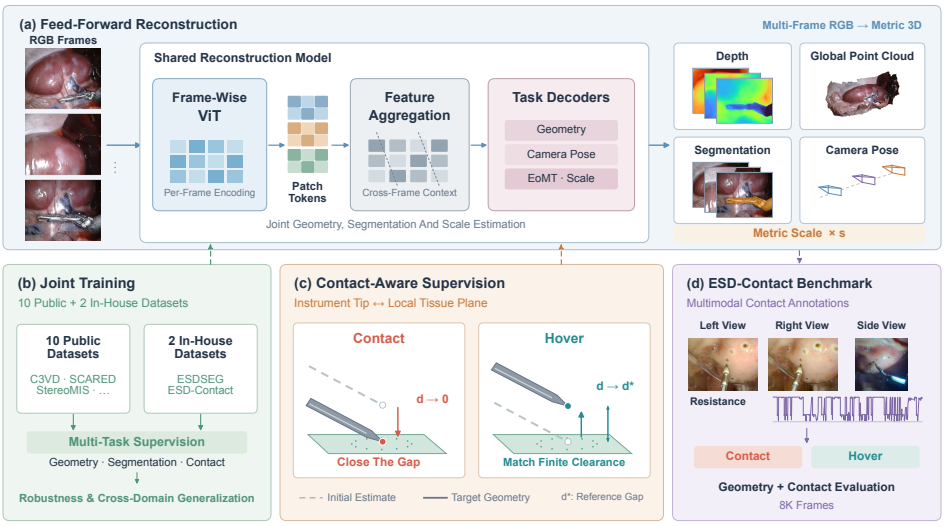


Figure 3: **CARE3R framework.** (a) A multi-frame encoder and task-specific heads predict metric geometry, camera poses, and instrument segmentation. (b) Joint training combines ten public and two in-house datasets. (c) Contact-Aware Supervision closes the tip–tissue gap on contact and matches a side-view-derived reference during hover. (d) ESD-Contact provides a multimodal benchmark for scene geometry and contact discrimination.

support both surgical 3D learning and direct evaluation of contact and hover discrimination from the primary view. Acquisition, annotation, and per-clip statistics are detailed in Appendix A.2.

Contact benchmark. Most prior surgical 3D methods evaluate scene geometry with pixel-level metrics such as RMSE, AbsRel, and δ accuracy (Guo et al., 2025; Cui et al., 2024), but do not test local instrument–tissue separation. Four clips, the monocular clip and three stereo clips, are excluded from training and serve as validation clips. On their contact and hover frames, we read a signed tip–tissue gap from each method's predicted depth: the median depth at the annotated tip minus that of the tissue just outside the annotated instrument mask near the tip, so that a hovering tip has a negative gap. Because the annotated tip and mask are given, the benchmark evaluates geometry rather than detection. On the stereo clips, each prediction is aligned per frame to FS by its tissue median, which puts all methods on a common millimetre scale and changes our AUC by less than 0.001; the monocular clip is scored without alignment. The primary metric is the ROC–AUC of contact against hover, computed per clip and averaged over clips. Average precision (AP) and balanced accuracy (bAcc) at a fixed gap threshold are secondary (Appendix A.7).

3 Method

Figure 3 gives an overview of CARE3R. Section 3.1 describes the model, Section 3.2 introduces Contact-Aware Supervision, and Section 3.3 gives the training objective.

3.1 Model

Given N primary-view RGB frames $\{I_i\}_{i=1}^{N}$, CARE3R predicts in one feed-forward pass a global pointmap $\widehat{P}_i^g$, a camera-to-world pose $\widehat{T}_i$, instrument mask and class logits $\widehat{M}_i$ and $\widehat{Y}_i$, and a dense confidence $\widehat{C}_i$ for every frame, together with one positive metric scale $\widehat{s}$ that converts the canonical pointmaps of the sequence to millimetres. The frames are patchified by a DINOv2-style encoder (Oquab et al., 2024), and every frame receives one camera token and 16 register tokens. A 24-level VGGT-Ω aggregator (Wang et al., 2026) alternates within-frame and inter-frame attention blocks, and features from four of its levels feed the heads. Dense and camera heads predict depth, confidence, pose, and intrinsics, from which the local and global pointmaps are formed; a learn-

able scale query attends to the register and patch tokens to predict $\hat{s}$; and an EoMT-style query head (Kerssies et al., 2025) predicts instrument masks and classes. Details are in Appendix A.1.

3.2 Contact-Aware Supervision

CAS supervises the instrument with targets that come from independent measurements, never from the model's own prediction of the instrument, which would move with the tip. Its contact and hover terms supervise the clearance between the instrument tip and the tissue beneath it, using the contact-state labels and, for hover frames, distances measured in the side view. Its profile term supervises the shape of the instrument with the depth profile that stereo measures along the blade.

Contact and hover terms. A pointmap has no explicit surface, unlike the meshes used to model hand–object contact (Grady et al., 2021), so we fit an oriented plane (c_i, n_i) to predicted tissue points sampled on two rings around the annotated tip, excluding instrument pixels. The fitted points are detached, so the plane does not move with the tip. The predicted clearance is the signed distance from the predicted tip point p_i^{tip} to this plane,

$$\widehat{d}_i^{\perp} = (p_i^{\text{tip}} - c_i)^{\top} n_i, \tag{1}$$

which is positive when the tip lies on the camera-facing side of the tissue. Contact frames drive $\widehat{d}_i^{\perp}$ to zero, and hover frames drive it to a clearance measured with the side view.

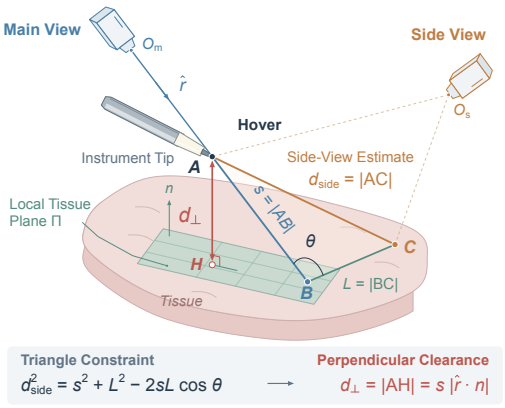


Figure 4: **Triangle construction of the hover target.** For one hover anchor (indices omitted), the diagram's $\Pi, n, d_{\text{side}}, d_\perp$ correspond to $\Pi_i^\star, n_i^\star, d_{ij}^{\text{side}}, d_{ij}^\perp$; $\Pi_i^\star$ is the offline reference plane, distinct from the predicted plane.

Hover target. The annotated primary-view tip defines a unit camera ray $\hat{\mathbf{r}}_i$, but its depth is unknown. Let A_i be the tip on this ray and C_{ij} a tissue landmark back-projected from the stereo depth. The side view provides their point-to-point distance $d_{ij}^{\text{side}} = |A_i C_{ij}|$. An offline reference plane $\Pi_i^\star = (c_i^\star, n_i^\star)$, fitted to the stereo depth around the tip, intersects the ray at B_i. With $L_{ij} = |B_i C_{ij}|$ and $\theta_{ij} = \angle A_i B_i C_{ij}$ (Figure 4),

$$B_i = \frac{(c_i^\star)^{\top} n_i^\star}{\hat{\mathbf{r}}_i^{\top} n_i^\star}\,\hat{\mathbf{r}}_i, \qquad \cos\theta_{ij} = \frac{(-\hat{\mathbf{r}}_i)^{\top}(C_{ij} - B_i)}{L_{ij}}. \tag{2}$$

With the tip on the camera-facing side of B_i, the larger law-of-cosines root (Appendix A.4) gives its displacement along the ray, whose projection onto the plane normal is the hover target:

$$\begin{aligned} s_{ij} &= L_{ij}\cos\theta_{ij} + \sqrt{(d_{ij}^{\text{side}})^2 - L_{ij}^2(1-\cos^2\theta_{ij})}, \\ d_{ij}^{\perp} &= s_{ij}\,|\hat{\mathbf{r}}_i^{\top} n_i^\star| = |A_i H_{ij}|, \end{aligned} \tag{3}$$

where H_{ij} is the foot of the perpendicular from A_i to $\Pi_i^\star$, so $d_{ij}^\perp$ has the same meaning as $\widehat{d}_i^{\perp}$. The targets are computed offline without primary-view tip depth and stay fixed during training; the side view is not needed at inference.

Let t_i be the target of the landmark closest to the tip in the side view, converted to pointmap units, and $e_i = \widehat{d}_i^{\perp} - t_i$. The contact and hover terms are

$$\begin{aligned} \mathcal{L}_{\text{contact}} &= \text{WM}\Big(\mathcal{H}_{0.1}(\widehat{d}^{\perp}),\, m^c\Big), \\ \mathcal{L}_{\text{hover}} &= \text{WM}\big(\mathcal{H}_{0.1}(\text{sign}(e)\,[\,|e| - 0.2|t|\,]_+),\, m^h\big), \end{aligned} \tag{4}$$

where WM is a masked mean, $\mathcal{H}_{0.1}$ the Huber loss (threshold in pointmap units), and m^c, m^h select valid contact and hover frames. The 20% dead band absorbs target noise, and larger errors are penalized symmetrically.

Profile term. A clearance loss on the tip alone is under-determined: the network can satisfy it by bending the tip instead of moving the instrument, especially when dense depth regression holds the rest of the blade in place. We therefore release the instrument from dense depth regression (Section 3.3) and supervise its shape instead. Along the knife axis, oriented towards the annotated tip, we average the predicted depth $\bar{z}_{ik}$ and the stereo depth $\bar{z}^{\mathrm{FS}}_{ik}$ in bins k, subtract from each profile its mean over the shaft bins, and compare the two profiles on the tip bins:

$$\mathcal{L}_{\mathrm{profile}} = \mathrm{WM}_{i,k}\left(\left[\left| (\bar{z}_{ik} - \bar{z}_i^{\mathrm{shaft}}) - (\bar{z}^{\mathrm{FS}}_{ik} - \bar{z}_i^{\mathrm{FS,shaft}}) \right| - \epsilon_{\mathrm{p}} \right]_+ ,\ \nu_{ik} \right), \tag{5}$$

where ν_{ik} selects valid tip bins and $\epsilon_{\mathrm{p}} = 0.2\,\mathrm{mm}$. Centring on the shaft makes the term constrain the shape of the instrument but not its placement, which is left to the contact and hover terms. It needs no interaction label; the annotated tip only orients the knife axis.

3.3 Training Objective

CARE3R is trained jointly on twelve datasets, ten public and two in-house (ESDSEG and ESD-CONTACT), each contributing the geometry, camera, and segmentation labels it provides (Appendix B.1). On ESD-CONTACT, FS supplies the dense depth and normals, the tissue geometry of the triangle construction, and the reference profile. The base objective is

$$\mathcal{L}_{\mathrm{base}} = \mathcal{L}_{\mathrm{point}} + 0.1\,\mathcal{L}_{\mathrm{cam}} + \lambda_{\mathrm{seg}}\mathcal{L}_{\mathrm{seg}}, \tag{6}$$

where $\mathcal{L}_{\mathrm{point}}$ combines a confidence-weighted, trimmed point loss with normal and scale supervision (Appendix A.1). Three settings complete the recipe. REL removes the instrument pixels of ESD-CONTACT from the point loss, so that the contact and hover terms can move the whole instrument; the normal term still covers these pixels. FLOOR adds a one-sided floor of 1.04 mm to the hover clearance, whose penalty vanishes once the floor is met. NORM raises the weight of the normal term from 1 to 20. The full objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{c}}\big(\mathcal{L}_{\mathrm{contact}} + \mathcal{L}_{\mathrm{hover}} + \mathcal{L}_{\mathrm{floor}}\big) + \lambda_{\mathrm{p}}\mathcal{L}_{\mathrm{profile}}, \tag{7}$$

with $\lambda_{\mathrm{seg}} = 1$, $\lambda_{\mathrm{c}} = 3$, and $\lambda_{\mathrm{p}} = 0.3$. Millimetre targets and tolerances are converted to pointmap units with a per-sequence normalization factor. FLOOR, REL, and NORM are ablated in Appendix E. At inference the model takes only RGB frames.

4 Experiments

We evaluate CARE3R on instrument–tissue contact (Section 4.1), analyse which parts of CAS produce the contact gain (Section 4.2), and test depth, camera pose, and reconstruction (Section 4.3). We compare with seven released geometry models used without project-specific fine-tuning, Pi3 (Wang et al., 2025b), DAv3 (Lin et al., 2025), VGGT-Ω (Wang et al., 2026), VGGT (Wang et al., 2025a), Fast3R (Yang et al., 2025), DUSt3R with global alignment (Wang et al., 2024), and Endo3R (Guo et al., 2025), and with DAv3 (FT), whose prediction heads we fine-tune on our training data without contact supervision. All results come from single training runs (Appendix B).

4.1 Instrument–Tissue Contact

Table 1 scores contact against hover on the four validation clips of Section 2; the three stereo clips come from the endoscope used for training (in-domain) and the monocular clip from a different endoscope (out-of-domain). CARE3R reaches an AUC of 0.933 over all clips, 12.7 points above the strongest baseline, and leads in both domain groups, including the out-of-domain clip (0.909). It also leads in AP (0.785 vs. at most 0.542), and on the stereo clips its gaps follow FS most closely (Spearman $\rho = 0.806$ vs. at most 0.521). Domain fine-tuning does not close the gap: DAv3 (FT) scores slightly below zero-shot DAv3 (0.777 vs. 0.800). The fixed-threshold bAcc of CARE3R is also the highest, but this measure is weaker: at $-0.5\,\mathrm{mm}$ even FS reaches only 0.650, and our model without the contact and hover terms (Section 4.2) scores higher (0.790 vs. 0.737), because the contact terms move the median contact gap from -0.17 to $-0.48\,\mathrm{mm}$, so that about half of the contact frames fall below the threshold (two-thirds for FS, median $-0.68\,\mathrm{mm}$). We therefore read the gain as a better ranking of contact and hover rather than as calibrated clearance.

Table 1: **Contact discrimination on the ESD-Contact validation clips.** AUC (ROC) and AP (average precision of contact) are averaged over clips; bAcc pools frames at a fixed threshold, -0.5 mm for In-domain (FS-aligned stereo clips) and -0.01 on the gap divided by the ring depth for All clips and Out-of-domain; ρ is the Spearman correlation between the predicted and FS gaps on the stereo clips. Higher is better for all metrics. For reference, FS reaches an AUC of 0.892, an AP of 0.562, and a bAcc of 0.650 on the stereo clips. Top-3 results are highlighted as first , second , and third .

	All clips			In-domain				Out-of-domain		
Method	AUC	AP	bAcc	AUC	AP	bAcc	ρ	AUC	AP	bAcc
VGGT	0.5481	0.331	0.591	0.5019	0.196	0.595	0.149	0.6867	0.733	0.611
Fast3R	0.5674	0.341	0.597	0.5045	0.196	0.573	0.038	0.7560	0.778	0.663
DUSt3R+GA	0.5795	0.373	0.540	0.5286	0.247	0.516	0.154	0.7323	0.753	0.675
Endo3R	0.3408	0.249	0.490	0.2957	0.149	0.486	-0.213	0.4759	0.548	0.513
VGGT-Ω	0.7611	0.476	0.685	0.7238	0.341	0.710	0.433	0.8732	0.883	0.771
DAv3 (FT)	0.7769	0.487	0.645	0.7522	0.363	0.690	0.521	0.8511	0.861	0.711
DAv3	0.7997	0.542	0.697	0.7750	0.434	0.709	0.449	0.8737	0.865	0.749
Pi3	0.8057	0.515	0.679	0.7992	0.414	0.700	0.462	0.8251	0.816	0.722
CARE3R (Ours)	0.9329	0.785	0.769	0.9408	0.745	0.737	0.806	0.9094	0.902	0.824

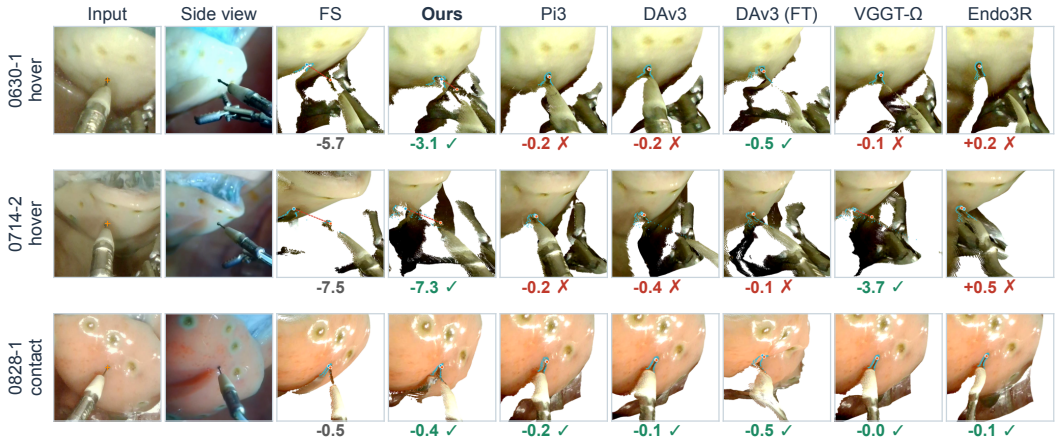


Figure 5: **Reconstructed tip–tissue geometry on validation frames.** In hover frames, CARE3R keeps a clear gap like the stereo reference (FS), whereas most baselines collapse it (values: gap in mm, rounded; ✓/✗: state correct at $\tau = -0.5$ mm on the unrounded gap); in contact, all methods are correct on this frame.

Qualitative and temporal behaviour. Figure 5 renders three of 15 candidate frames fixed before any model output was inspected (per stereo clip, the three hover and two contact frames whose FS gap is closest to the clip and state median); we picked the three for clarity, and the other 12 are in Appendix C. In the hover rows, FS measures -5.7 and -7.5 mm and CARE3R keeps a clear gap (-3.1 and -7.3 mm), whereas most baselines place the tip within 0.5 mm of the tissue; in the contact row every method is correct. Over all 15 candidates, CARE3R recovers 9 of 9 hover and 4 of 6 contact frames, and both misses are frames on which FS itself measures below -0.5 mm. In Figure 6, CARE3R tracks FS into and out of the contact episodes (Spearman ρ 0.86 and 0.79, vs. 0.33 and 0.70 for Pi3, which stays within 1 mm of zero on 98% of 0714-2 frames), though not exactly: it stays at -5.0 and -5.8 mm in two short contact episodes of 0714-2, and after frame 525, where the tool hovers beyond the tissue edge, it underestimates the clearance (-11.9 vs. -31.3 mm for FS) while still labelling all 193 frames as hover; Pi3 stays at a median of -0.26 mm there and labels 75.6% of these hover frames as contact.

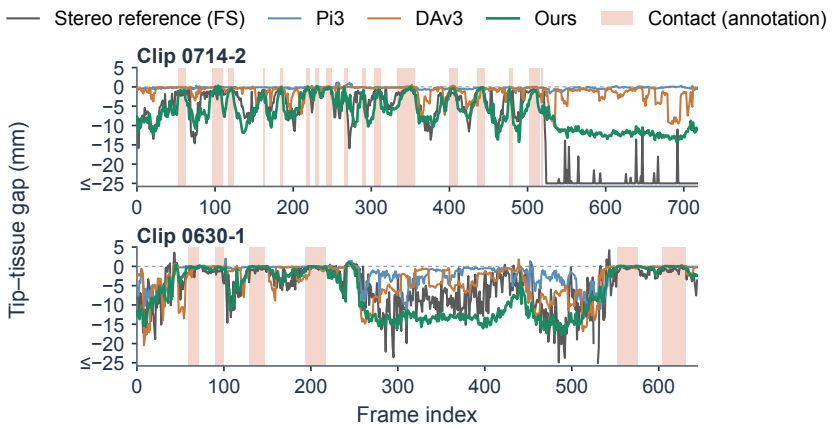


Figure 6: **Per-frame tip–tissue gap on two validation clips.** CARE3R follows FS into and out of the shaded contact bands, whereas Pi3 stays near zero gap on 0714-2 and follows FS only partly on 0630-1; after frame 525 of 0714-2, CARE3R underestimates the clearance but still reports hover.

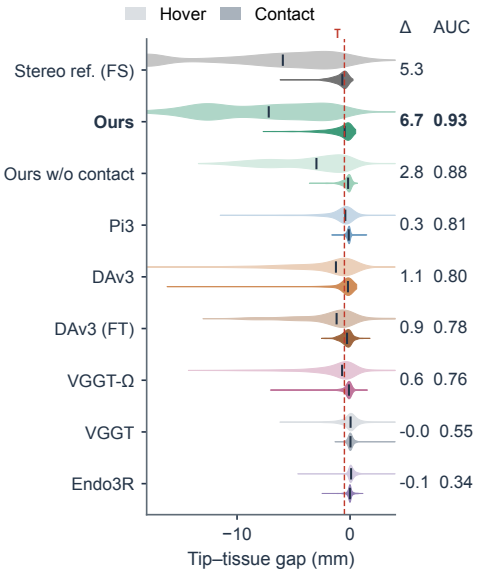


Figure 7: **Hover and contact gap distributions on the three stereo validation clips.** CARE3R separates the states more widely than FS; external models keep hover frames near zero. Δ (mm): contact minus hover median gap; AUC: four-clip mean (Tables 1–2); dashed: $\tau = -0.5$ mm.

Distribution-level separation. Figure 7 pools all hover and contact frames of the stereo clips. The contact–hover separation Δ is 5.3 mm for FS and 6.7 mm for CARE3R, whereas every external model, including DAv3 (FT), keeps hover frames near zero ($\Delta \leq 1.1$ mm and AUC ≤ 0.81); without the contact and hover terms our model reaches $\Delta = 2.8$ mm. As CARE3R exceeds FS, Δ indicates separation relative to the stereo reference, not calibrated clearance.

Table 2: **Instrument supervision and contact.** Stage 1 uses dense FS depth on all pixels and is scored without the per-frame FS alignment. The other rows raise the normal weight (NORM), remove the instrument from the dense loss (REL), add a blade profile referenced to the model's own shaft line or to FS, and add the contact and hover terms with the hover floor (C/H); the last row is CARE3R. Tissue AbsRel is measured against FS on the stereo clips. The line-profile row was trained before the data-pipeline corrections of Appendix B.1.

	Contact AUC ↑			Tissue
Supervision	All	In-dom.	Out-dom.	AbsRel ↓
Dense FS (stage 1)	0.822	0.805	0.874	0.1200
REL + line profile	0.743	0.698	0.879	0.1225
REL + FS profile	0.883	0.886	0.873	0.1220
REL + FS profile + C/H	0.933	0.941	0.909	0.1222

4.2 Component Analysis

Table 2 starts from the stage-1 model, trained with dense FS depth on all pixels and no CAS, which initializes three stage-2 variants that differ in how the instrument is supervised (the caption notes one further difference). Dense supervision alone reaches an AUC of 0.822 but separates contact from hover by less than half as much as FS (2.0 vs. 5.4 mm, mean over the stereo clips). With only a self-referential profile, the released instrument reaches an AUC of only 0.743, as the tip sinks onto the tissue (hover median −0.6 vs. −6.0 mm for FS). With the stereo profile as anchor the AUC

Table 3: **Depth, pose, and reconstruction.** Depth: median-scaled single-image AbsRel (↓) and δ_1 (↑). Pose: AUC@30 (%, ↑) and Sim(3)-aligned ATE (mm, ↓). Reconstruction: per-sequence medians of accuracy and completion (mm, ↓) and of normal consistency (N.C., ↑) on C3VD. Colors mark the top three as in Table 1.

	Depth						Pose				Reconstruction		
	C3VD		SCARED		SERV-CT		C3VD		SCARED		C3VD		
Method	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AUC	ATE	AUC	ATE	Acc.	Comp.	N.C.
VGGT	0.617	0.267	0.064	0.970	0.076	0.966	7.6	6.47	62.7	2.51	4.10	11.11	0.569
Fast3R	0.645	0.271	0.125	0.840	0.138	0.822	23.5	5.54	40.1	5.17	2.42	8.19	0.575
DUSt3R+GA	0.600	0.302	0.117	0.853	0.113	0.881	20.3	4.27	37.7	9.71	2.08	5.93	0.592
Endo3R	0.220	0.711	0.047	0.993	0.119	0.849	6.2	8.29	38.5	9.33	1.85	2.13	0.649
VGGT-Ω	0.607	0.285	0.060	0.971	0.052	0.997	33.9	2.36	75.8	1.17	2.02	6.15	0.607
DAv3 (FT)	0.089	0.938	0.097	0.904	0.081	0.942	17.7	6.30	60.1	1.77	1.85	2.09	0.709
DAv3	0.633	0.270	0.107	0.884	0.091	0.928	9.1	4.81	68.1	1.22	3.24	9.47	0.639
Pi3	0.580	0.292	0.062	0.979	0.063	0.984	45.3	2.43	75.0	1.30	1.84	5.41	0.579
CARE3R (Ours)	0.070	0.961	0.048	0.992	0.060	0.984	63.1	0.61	68.8	1.99	1.86	4.96	0.726

is 0.883, and the contact and hover terms add 0.050, positive on every clip (+0.023 to +0.083), while widening Δ from 2.8 to 6.7 mm. The tissue depth error stays at 0.120–0.123 in all rows, and the stage-2 variants have near-identical conventional metrics (C3VD AbsRel 0.0698–0.0699, pose AUC@30 62.1–63.1, reconstruction N.C. 0.726): contact discrimination varies while per-pixel accuracy does not. FLOOR, REL, and NORM, ablated in Appendix E, keep the AUC within 0.926–0.928 and mainly change the gap magnitude and surface smoothness.

4.3 Depth, Pose, and Reconstruction

Table 3 reports median-scaled single-image depth, relative pose, and C3VD reconstruction after Sim(3) (Umeyama, 1991) and ICP (Besl & McKay, 1992) alignment (Appendix B.4). C3VD and SCARED use held-out sequences of training datasets; SERV-CT (Edwards et al., 2021), an ex-vivo porcine dataset with CT-derived reference depth, is outside our training data.

CARE3R is best on C3VD depth and pose and on reconstruction normal consistency, and competitive elsewhere: Endo3R is slightly better on SCARED depth, VGGT-Ω on SERV-CT depth, VGGT-Ω and Pi3 on SCARED pose (both DAv3 variants also in ATE), Pi3, Endo3R, and DAv3 (FT) marginally on C3VD accuracy, and DAv3 (FT) and Endo3R much more on C3VD completion. Without per-image scaling, its metric AbsRel is 0.127, 0.131, and 0.295 on C3VD, SCARED, and SERV-CT, so the table measures shape, not calibrated depth. A natural alternative to contact supervision is to adapt a strong geometry model to the surgical domain with standard losses on the same data. DAv3 (FT) shows the dissociation between depth accuracy and clearance: fine-tuning its heads cuts C3VD AbsRel from 0.633 to 0.089 and improves every reconstruction measure, yet its contact AUC drops from 0.800 to 0.777. This head-only adaptation converges within 16k steps and over-fits with longer training; it does not reproduce the official DA3 recipe (Appendix B.2). The control shows that this standard adaptation does not recover the gap, not that no fine-tuning strategy could. On instrument segmentation, CARE3R is best on ESDSEG and zero-shot Kvasir-Instrument (Jha et al., 2021) and second to SAM2.1 (Ravi et al., 2025) elsewhere (Appendix D).

5 Conclusion

Endoscopic reconstruction should preserve the spatial relation between instruments and tissue as well as overall scene geometry. We introduced ESD-Contact and its contact benchmark and proposed CARE3R, a feed-forward reconstruction model trained with contact-aware supervision, which improves contact discrimination while remaining competitive on conventional benchmarks. Our evaluation uses four validation clips without a separate test split and single training runs; future work will add a held-out test set, improve blade-shape robustness and clearance calibration, and read contact without annotated tips or masks.

AI USE STATEMENT

Generative AI tools were used to assist with early-stage manuscript organization and language editing. All scientific claims, mathematical derivations, citations, code, figures, and reported results are reviewed and verified by the authors, who take responsibility for the final content of the work.

REFERENCES

Max Allan, Satoshi Kondo, Sebastian Bodenstedt, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

Max Allan et al. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133*, 2021.

Aleks Attanasio, Bruno Scaglioni, Elena De Momi, Paolo Fiorini, and Pietro Valdastri. Autonomy in surgical robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 4:651–679, 2021. doi: 10.1146/annurev-control-062420-090543.

Paul J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

Taylor L. Bobrow, Mayank Golhar, Rohan C. Vijayan, Venkata S. Akshintala, Juan R. Garcia, and Nicholas J. Durr. Colonoscopy 3d video dataset with paired depth from 2d–3d registration. *Medical Image Analysis*, 2023. doi: 10.1016/j.media.2023.102956.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022.

Beilei Cui, Mobarakol Islam, Long Bai, An Wang, and Hongliang Ren. EndoDAC: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera. In *Medical Image Computing and Computer Assisted Intervention*, 2024. doi: 10.1007/978-3-031-72089-5_20.

Beilei Cui, Long Bai, Mobarakol Islam, An Wang, Zhiqi Ma, Yiming Huang, Feng Li, Zhen Chen, Zhongliang Jiang, Nassir Navab, and Hongliang Ren. Learning to efficiently adapt foundation models for self-supervised endoscopic 3d scene reconstruction from any cameras. *arXiv preprint arXiv:2503.15917*, 2025.

Philip J. E. Edwards, Dimitrios Psychogyios, Stefanie Speidel, Lena Maier-Hein, and Danail Stoyanov. SERV-CT: A disparity dataset from cone-beam CT for validation of endoscopic 3d reconstruction. *Medical Image Analysis*, 2021. doi: 10.1016/j.media.2021.102302.

Mayank V. Golhar, Lucas Sebastian Galeano Fretes, Loren Ayers, Venkata S. Akshintala, Taylor L. Bobrow, and Nicholas J. Durr. C3VDv2 – colonoscopy 3D video dataset with enhanced realism. *Scientific Data*, 13:1273, 2026.

Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. doi: 10.1109/CVPR46437.2021.00152.

Caspar Gruijthuijsen, Richard Colchester, Alain Devreker, Allan Javaux, Efthymios Maneas, Sacha Noimark, Wenfeng Xia, Danail Stoyanov, Dominiek Reynaerts, Jan Deprest, Sebastien Ourselin, Adrien Desjardins, Tom Vercauteren, and Emmanuel Vander Poorten. Haptic guidance based on all-optical ultrasound distance sensing for safer minimally invasive fetal surgery. *Journal of Medical Robotics Research*, 3:1841001, 2018. doi: 10.1142/S2424905X18410015.

Jiaxin Guo, Wenzhen Dong, Tianyu Huang, Hao Ding, Ziyi Wang, Haomin Kuang, Qi Dou, and Yun-hui Liu. Endo3R: Unified online reconstruction from dynamic monocular endoscopic video. *arXiv preprint arXiv:2504.03198*, 2025.

Ali Hatamizadeh and Jan Kautz. MambaVision: A hybrid Mamba-transformer vision backbone. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Michel Hayoz, Christopher Hahne, Mathias Gallardo, Daniel Candinas, Thomas Kurmann, Maximilian Allan, and Raphael Sznitman. Learning how to robustly estimate camera pose in endoscopic videos. *International Journal of Computer Assisted Radiology and Surgery*, 18(7):1185–1192, 2023. doi: 10.1007/s11548-023-02919-w.

Yiming Huang, Long Bai, Beilei Cui, Kun Yuan, Guankun Wang, Mobarak I. Hoque, Nicolas Padoy, Nassir Navab, and Hongliang Ren. SurgTPGS: Semantic 3d surgical scene understanding with text promptable gaussian splatting. In *Medical Image Computing and Computer Assisted Intervention*, 2025. doi: 10.1007/978-3-032-05114-1_56.

Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A. Hicks, Vajira Thambawita, Enrique Garcia-Ceja, Michael A. Riegler, Thomas de Lange, Peter T. Schmidt, Håvard D. Johansen, Dag Johansen, and Pål Halvorsen. Kvasir-Instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In *MultiMedia Modeling*, pp. 218–229, 2021. doi: 10.1007/978-3-030-67835-7_19.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. In *European Conference on Computer Vision*, 2024.

Tommie Kerssies, Niccolò Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your ViT is secretly an image segmentation model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with MASt3R. In *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-73220-1_5.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Ruyu Liu, Lin Wang, Mingming Zhou, Jianhua Zhang, Haoyu Zhang, et al. PolypSense3D: A multi-source benchmark dataset for depth-aware polyp size measurement in endoscopy. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Yifan Liu, Chenxin Li, Chen Yang, and Yixuan Yuan. EndoGaussian: Real-time gaussian splatting for dynamic endoscopic scene reconstruction. *arXiv preprint arXiv:2401.12561*, 2024.

Peter Mountney, Danail Stoyanov, and Guang-Zhong Yang. Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Processing Magazine*, 27(4):14–24, 2010. doi: 10.1109/MSP.2010.936728.

Laura Oliva-Maza, Florian Steidle, Julian Klodmann, Klaus Strobl, Arkadiusz Miernik, and Rudolph Triebel. KidneyDepth: A synthetic kidney dataset for metric depth estimation in ureteroscopy. In *Medical Image Computing and Computer Assisted Intervention*, pp. 331–340, 2025. doi: 10.1007/978-3-032-05114-1_32.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Anita Rau, Sophia Bano, Yueming Jin, Pablo Azagra, Javier Morlana, Edward Sanderson, Bogdan J. Matuszewski, Jae Young Lee, Dong-Jae Lee, Erez Posner, et al. SimCol3D–3d reconstruction during colonoscopy challenge. *arXiv preprint arXiv:2307.11261*, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

David Recasens, Jose Lamarca, Jose M. Facil, J. M. M. Montiel, and Javier Civera. Endo-Depth-and-Motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *arXiv preprint arXiv:2103.16525*, 2021.

Javier Rodriguez-Puigvert, Victor M. Batlle, J. M. M. Montiel, Ruben Martinez Cantin, Pascal Fua, Juan D. Tardos, and Javier Civera. LightDepth: Single-view depth self-supervision from illumination decline. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21273–21283, 2023. doi: 10.1109/ICCV51070.2023.01945.

Shuwei Shao, Zhongcai Pei, Weihai Chen, Baochang Zhang, Xingming Wu, Dianmin Sun, and David Doermann. Self-supervised learning for monocular depth estimation on minimally invasive surgery scenes. In *IEEE International Conference on Robotics and Automation*, 2021. doi: 10.1109/ICRA48506.2021.9561508.

Shuwei Shao, Kejin Zhu, Shixing Ma, Xinzhe Du, Baochang Zhang, and Zhe Min. Depth any endoscopy: Towards self-supervised generalizable depth estimation in monocular endoscopy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34126–34137, 2026.

Qingyao Tian, Zhen Chen, Huai Liao, Xinyan Huang, Lujie Li, Sebastien Ourselin, and Hongbin Liu. EndoOmni: Zero-shot cross-dataset depth estimation in endoscopy by robust self-learning from noisy labels. *arXiv preprint arXiv:2409.05442*, 2024.

Nicolas Toussaint, Emanuele Colleoni, Ricardo Sanchez-Matilla, Joshua Sutcliffe, Vanessa Thompson, Muhammad Asad, Imanol Luengo, and Danail Stoyanov. Zero-shot monocular metric depth for endoscopic images. In *MICCAI Workshop on Data Engineering in Medical Imaging*, 2025. doi: 10.1007/978-3-032-08009-7_12.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a. doi: 10.1109/CVPR52734.2025.00499.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.01956.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025b.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-shot stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/CVPR52734.2025.00495.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision*, 2018.

Weixing Xie, Junfeng Yao, Xianpeng Cao, Qiqin Lin, Zerui Tang, Dong Xiao, and Xiaohu Guo. SurgicalGaussian: Deformable 3d gaussians for high-fidelity surgical scene reconstruction. In *Medical Image Computing and Computer Assisted Intervention*, 2024. doi: 10.1007/978-3-031-72089-5_58.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21924–21935, 2025.

Zanwei Zhou, Chen Yang, Piao Yang, Xiaokang Yang, and Wei Shen. EndoDAV: Depth any video in endoscopy with spatiotemporal accuracy. In *Medical Image Computing and Computer Assisted Intervention*, 2025. doi: 10.1007/978-3-032-05114-1_19.

APPENDIX CONTENTS

A Implementation and Evaluation Details **14**

A.1 Feedforward Architecture and Base Objective 14

A.2 ESD-CONTACT Dataset Details 15

A.3 Local Surface Estimation 16

A.4 Triangle Construction and Feasibility Gates 17

A.5 Loss Normalization and State Masks 18

A.6 Hover Floor, Profile Term, and Recipe Settings 19

A.7 Contact Readout and Evaluation Safeguards 20

B Experimental Setup and Evaluation Protocols **21**

B.1 Training Datasets and Supervision 21

B.2 Baselines and Comparison Settings 22

B.3 Contact Operating Points and Aggregation 22

B.4 Task-specific Evaluation Details 23

B.5 Runtime and Memory 23

C Additional Contact Visualizations **24**

D Instrument Segmentation **25**

E Additional Ablations **26**

A IMPLEMENTATION AND EVALUATION DETAILS

A.1 FEEDFORWARD ARCHITECTURE AND BASE OBJECTIVE

Shared representation and outputs. The input frames are patchified with patch size 16 and processed jointly by a 24-level VGGT-Ω aggregator whose within-frame and inter-frame blocks alternate; selected inter-frame blocks restrict cross-frame attention to the camera and register tokens, and the outputs of four levels are concatenated for the heads. A dense head predicts a positive canonical-scale depth $\widehat{D}_i$ and confidence $\widehat{C}_i$, while a camera head predicts a camera-to-world transform $\widehat{T}_i$ and intrinsics $\widehat{K}_i$. The implementation unprojects each depth map as

$$\widehat{P}_i^{\ell}(u) = \widehat{D}_i(u)\,\widehat{K}_i^{-1}\widetilde{u}, \tag{8}$$

where $\widetilde{u}$ is the homogeneous image coordinate. The corresponding global pointmap is

$$\widehat{P}_i^{g}(u) = \left[\widehat{T}_i \begin{bmatrix} \widehat{P}_i^{\ell}(u) \\ 1 \end{bmatrix}\right]_{1:3}. \tag{9}$$

A learned scale query cross-attends to the output register and dense patch tokens. A scalar log-scale is produced for each frame, averaged across the sequence, and exponentiated to give one positive metric scale $\widehat{s}$ per sequence. Accordingly, the metric local pointmap and depth used at inference are

$$\widehat{P}_i^{\ell,m}(u) = \widehat{s}\,\widehat{P}_i^{\ell}(u), \qquad \widehat{D}_i^{m}(u) = \widehat{s}\,\widehat{D}_i(u). \tag{10}$$

A separate EoMT-style query head predicts instrument mask and class logits from the dense patch tokens.

The model therefore returns local and global pointmaps, camera poses, metric scale, instrument segmentation, and confidence in one forward pass. It does not predict contact through a separate classifier. Instead, contact and clearance are read from the relative geometry of the instrument tip and adjacent tissue. In the controlled benchmark, annotated tips and masks isolate geometric error from localization error; the predicted segmentation remains available for end-to-end deployment.

Base objective. The base objective is

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{point}} + 0.1\,\mathcal{L}_{\text{cam}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}. \tag{11}$$

The component losses in the implementation are

$$\mathcal{L}_{\text{point}} = \mathcal{L}^{\text{REL,trim}}_{\text{aleatoric}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} + 0.1\mathcal{L}_{\text{scale}}, \tag{12}$$

$$\mathcal{L}_{\text{cam}} = 100\mathcal{L}_{\text{trans}} + \mathcal{L}_{\text{rot}}, \tag{13}$$

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{class}}. \tag{14}$$

Here $\lambda_{\text{seg}} = 1$ and $\lambda_{\text{normal}} = 20$ in the reported configuration. For a valid pixel with inverse-depth-weighted point error e and predicted confidence $c \in [1, 10^3]$, the aleatoric point term is $ce - \alpha \log c$ with $\alpha = 0.02$, the confidence-aware form of DUSt3R (Wang et al., 2024). $\mathcal{L}_{\text{normal}}$ compares predicted and ground-truth local surface normals, $\mathcal{L}_{\text{scale}}$ is an L_1 loss between predicted and target log-scales, and the camera terms use Huber translation error over all ordered frame pairs and rotation-angle error. Segmentation follows Mask2Former (Cheng et al., 2022): queries are assigned to ground-truth masks by Hungarian matching before the mask, Dice, and class terms are evaluated. The contact objective is detailed in Appendix A.5.

Trimmed point regression. On Hamlyn, StereoMIS, and ESD-CONTACT, whose depth is derived from stereo, the point loss retains the 90% of valid pixels with the smallest confidence-free point errors, keeps ties, and renormalizes by the number retained. The normal term uses its own valid neighborhoods without trimming, so a high retention keeps the depth and normal supervision on nearly the same pixels.

A.2 ESD-CONTACT DATASET DETAILS

Acquisition. All clips were recorded during ex-vivo ESD on porcine stomachs purchased from a local food market; no live animals were involved. Thirteen clips use a stereo endoscope (1280×720 per view, stored in portrait orientation, i.e. 1280 pixels high), and one clip uses a different, monocular endoscope (992×992). A side-view camera (1920×1080, landscape), calibrated with a checkerboard and placed roughly orthogonal to the endoscope axis, is temporally synchronized with the primary views. On the three clips of the last session, an LCR sensor records the knife–tissue electrical resistance, which falls from the megaohm range during hover to tens of kilohms during contact.

Annotation. Annotators label sparse frames of each clip. In the primary view they draw semantic masks (knife, forceps, and lesion) and mark the knife tip and tissue landmarks; in the side view they mark the knife tip, three shaft points, and the same tissue landmarks. Masks are propagated through the video with SAM2 (Ravi et al., 2025) and keypoints with CoTracker2 (Karaev et al., 2024), and all propagated labels are manually reviewed. Gaps of at most 15 frames in the primary-view tip track are filled by linear interpolation and flagged. Each frame carries one of four interaction states, judged on the side view: hover, near, contact, or unknown, where unknown mostly marks contacts whose tip is occluded. On the three sensor clips, an automatic procedure first derives the states from the resistance trace and an annotator then verifies every frame; residual mucus can bridge the electrodes after the tip lifts off, so the raw trace can over-report contact, and such frames are corrected during verification. On the other clips, the states are annotated manually. Near and unknown frames are excluded from the contact supervision and the benchmark.

Depth. Primary-view depth is computed from the rectified stereo pairs with Foundation-Stereo (Wen et al., 2025) and stored in millimetres up to 100 mm; the monocular clip has no depth. FS is reliable near the instrument but noisier on distant, weakly textured tissue. It provides the dense depth and normal supervision on the training clips and the reference depth of the benchmark.

Table 4: **ESD-CONTACT clips**, named by recording date and index. Near and unknown frames are excluded from contact supervision and evaluation. LCR: resistance recorded. †Worn knife tip; excluded from contact supervision.

Clip	Frames	Hover	Contact	Near	Unknown	Endoscope	LCR	Split
0630-1	645	516	115	14	0	stereo		validation
0702-1	306	172	121	7	6	stereo		training
0708-1	177	27	146	4	0	stereo		training
0714-1	572	361	205	6	0	stereo		training
0714-2	718	577	130	7	4	stereo		validation
0714-3	186	96	87	3	0	stereo		training
0722-2†	399	188	117	0	94	stereo		training
0724-1	709	419	290	0	0	stereo		training
0728-2†	598	479	114	0	5	stereo		training
0731-1	885	677	208	0	0	stereo		training
0731-3	712	302	388	0	22	monocular		validation
0828-1	781	569	212	0	0	stereo	✓	validation
0828-2	1,062	811	251	0	0	stereo	✓	training
0828-3	491	184	307	0	0	stereo	✓	training
Total	8,241	5,378	2,691	41	131			

Statistics and split. Table 4 lists the clips. The four validation clips (0630-1, 0714-2, 0828-1, and the monocular 0731-3) are excluded from training. Because the contact-loss settings were chosen on these clips, we call them validation rather than test clips. The benchmark scores 843 contact and 1,963 hover frames of them, 457 and 1,662 on the stereo clips; a few monocular frames lack a valid readout. The remaining ten clips (5,385 frames) are used for training; clips recorded in the same session cover different tissue regions. In 0722-2 and 0728-2 the knife tip was worn and its annotations are unreliable, so these clips are excluded from the contact supervision but kept for dense supervision.

Side-view distances. Side-view relative depth from Depth Anything 3 (Lin et al., 2025) is undistorted and scaled with the known 9.20 mm distance between two marked shaft points, and distances are rejected when the reconstructed tip is inconsistent with the shaft points. For a hover frame, matched tip and tissue observations then provide one point-to-point distance per valid tissue anchor j, computed with the calibrated side-view back-projection X_i^s as

$$d_{ij}^{\text{side}} = \|X_i^s(v_i^{\text{tip}}) - X_i^s(v_{ij}^s)\|_2. \tag{15}$$

For a contact frame, the discrete state supplies the zero-clearance target. Figure 4 shows how these measurements interact with the primary view. The triangle target never samples primary-view tip depth; it uses the primary-view tip ray, the stereo-derived tissue geometry, and the side-view distance. All side-view processing is performed offline, the resulting targets stay fixed during training, and inference requires no auxiliary camera.

A.3 LOCAL SURFACE ESTIMATION

A pointmap provides 3D samples but no explicit topology or signed distance field. The implemented contact loss constructs an oriented local surface around the annotated instrument-tip pixel u_i^{tip}. It uses two rings, each with 16 uniformly spaced samples. For $r \in \{0.03, 0.08\}$ and $k \in \{0, \dots, 15\}$, their image coordinates are

$$u_{irk} = u_i^{\text{tip}} + r \min(H, W) \left[\cos(2\pi k/16) \quad \sin(2\pi k/16)\right]^\top. \tag{16}$$

Let M_i^{inst} be the ground-truth instrument mask and let $\mathcal{B}(X, u)$ denote bilinear sampling. The binary sampling weight is

$$w_{irk} = \mathbb{1}[u_{irk} \text{ is in bounds}]\mathbb{1}[\mathcal{B}(M_i^{\text{inst}}, u_{irk}) < 10^{-3}], \tag{17}$$

so a location is rejected even when only its bilinear interpolation neighborhood touches the instrument. The same bilinear kernel and coordinate convention are used to sample the pointmap.

Let $q_{irk} = \mathrm{sg}[\mathrm{nan2num}(\mathcal{B}(\widehat{P}_i^{\ell}, u_{irk}))]$, where sg is stop-gradient, and flatten (r, k) to a single sample index a. With $S_i = \max(\sum_a w_{ia}, 10^{-8})$, the implementation fits the plane by

$$c_i = \frac{\sum_a w_{ia} q_{ia}}{S_i}, \tag{18}$$

$$\Sigma_i = \frac{\sum_a w_{ia}(q_{ia} - c_i)(q_{ia} - c_i)^\top}{S_i} + 10^{-8} I_3, \tag{19}$$

$$n_i = \mathrm{eigmin}(\Sigma_i), \tag{20}$$

The code then performs one robust refit. It computes absolute point-to-plane residuals, fills invalid positions with one plus the maximum residual, and removes valid samples exceeding three times the resulting median. If fewer than six samples remain, it falls back to the initial weights. The final plane is accepted only when it contains at least six tissue samples and $\sqrt{\lambda_{i,2}} > 10^{-4}$, where $\lambda_{i,2}$ is the second-smallest covariance eigenvalue. Finally, the normal is flipped whenever its z component is positive, giving $n_{i,z} \leq 0$ in the camera convention used by the code. Bilinear sampling at the tip gives $p_i^{\mathrm{tip}} = \mathcal{B}(\widehat{P}_i^{\ell}, u_i^{\mathrm{tip}})$, and the predicted signed clearance is

$$\widehat{d}_i^{\perp} = (p_i^{\mathrm{tip}} - c_i)^\top n_i. \tag{21}$$

With this orientation, positive values place the tip on the camera-facing side of the surface, zero denotes contact with the fitted plane, and negative values indicate penetration or an ordering error.

A.4 Triangle Construction and Feasibility Gates

The side camera directly measures a point-to-point distance, whereas Equation 21 is a point-to-plane distance. Regressing one quantity onto the other introduces a slope- and anchor-dependent bias. CAS removes this mismatch by converting the side-view measurement into a perpendicular clearance target before training.

Figure 4 illustrates the construction, which is computed offline. Let A_i be the unknown 3D tip in the primary camera frame. The annotated tip pixel defines the unit ray

$$\hat{\mathbf{r}}_i = \frac{K^{-1}\widetilde{u}_i^{\mathrm{tip}}}{\|K^{-1}\widetilde{u}_i^{\mathrm{tip}}\|_2}, \tag{22}$$

which does not require tip depth. An offline reference plane $\Pi_i^\star = (c_i^\star, n_i^\star)$ is fitted by SVD to positive FoundationStereo depth samples on the two-ring pattern in Equation 16, after applying the same bilinear instrument-mask rejection and requiring at least six samples. The ray–plane intersection is

$$B_i = \frac{(c_i^\star)^\top n_i^\star}{\hat{\mathbf{r}}_i^\top n_i^\star}\, \hat{\mathbf{r}}_i. \tag{23}$$

For a valid annotated tissue anchor j, the same FoundationStereo map gives C_{ij} by bilinear depth sampling and unprojection. Let $L_{ij} = |B_i C_{ij}|$ and let H_{ij} be the perpendicular foot of A_i on $\Pi_i^\star$. The angle at B_i is $\theta_{ij} = \angle A_i B_i C_{ij}$, encoded as

$$\cos\theta_{ij} = \frac{(-\hat{\mathbf{r}}_i)^\top (C_{ij} - B_i)}{L_{ij}}. \tag{24}$$

The side view supplies $d_{ij}^{\mathrm{side}} = |A_i C_{ij}|$ (Equation 15). The implementation places A_i on the camera-facing half of the tip ray through B_i and keeps the larger root of the quadratic implied by the law of cosines:

$$s_{ij} \equiv |A_i B_i| = L_{ij}\cos\theta_{ij} + \sqrt{(d_{ij}^{\mathrm{side}})^2 - L_{ij}^2(1 - \cos^2\theta_{ij})}. \tag{25}$$

When $d_{ij}^{\mathrm{side}} < L_{ij}$ and $\cos\theta_{ij} > 0$, both roots are positive, i.e. both place the tip on the camera-facing side, and the side-view distance alone cannot separate them; the implementation keeps the larger one. Over all hover frames of the 13 stereo clips, this occurs for about 38% of the accepted anchors and 45% of the fixed anchors j_i used by the hover term, with the discarded root a median 0.40 times the kept one. The target $d_{ij}^{\perp}$ is therefore the larger clearance consistent with the side view, not a uniquely determined one. Projecting this displacement onto the plane normal gives the target

$$d_{ij}^{\perp} = s_{ij}\,|\hat{\mathbf{r}}_i^\top n_i^\star| = |A_i H_{ij}|. \tag{26}$$

An anchor is accepted only if its side-view label and tissue depth are valid and

$$\Delta_{ij} = (d_{ij}^{\text{side}})^2 - L_{ij}^2(1 - \cos^2\theta_{ij}) > 0, \quad s_{ij} > 0, \quad L_{ij} \geq 10^{-6}, \quad |\hat{\mathbf{r}}_i^\top n_i^\star| \geq 10^{-6}. \tag{27}$$

Geometrically inconsistent anchors are rejected rather than projected into a plausible target. Each hover frame uses a fixed anchor j_i, the tissue landmark closest to the tip in the side view, rather than one re-selected from the computed perpendicular distances; if j_i fails these checks, the frame is masked from the hover term.

Three properties are central. First, θ_{ij} depends on the camera ray and tissue geometry but not on unknown tip depth, so the predicted tip cannot feed back into its own target. Second, $d_{ij}^{\perp} = |A_i H_{ij}|$ has the same point-to-plane meaning as $\widehat{d}_i^{\perp}$ in Equation 21. Third, the accepted targets are stored constants: side-view and stereo processing are outside the training graph. The anchor C_{ij} need not lie on $\Pi_i^\star$.

A.5 Loss Normalization and State Masks

The point loss compares predictions to ground-truth points expressed in a unit scene coordinate system. For sequence b, the code computes its millimetre normalization factor after transforming ground-truth points to the first-camera frame:

$$\eta_b = \frac{\sum_{i,u} v_{bi}(u)\, \|P_{bi}^{\text{gt}}(u)\|_2}{\sum_{i,u} v_{bi}(u) + 10^{-8}}, \tag{28}$$

where v_{bi} is the valid-depth mask. The selected triangle target is converted to pointmap units as $t_i = d_{ij_i}^{\perp}/\eta_b$, without further scaling.

To reproduce the implementation's handling of empty state subsets, define

$$\text{WM}(x, m) = \frac{M^{-1}\sum_{i=1}^{M} m_i x_i}{M^{-1}\sum_{i=1}^{M} m_i + 10^{-7}}, \tag{29}$$

where M is the number of flattened frames in the batch. Let m_i^c select contact frames with a valid tip and accepted local plane, and let m_i^h additionally require the hover state, $j_i \geq 0$, and a valid triangle target for anchor j_i. In the reported recipe, the hover mask also includes a non-differentiable stability test: refitting at 0.95 and 1.05 times the ring radii must not change the detached tip-to-plane distance by more than $0.2|t_i|$. A probe whose refit fails is treated as stable. Both the metric hover term and FLOOR share this mask; contact frames do not use this target-relative test. Contact frames impose

$$\mathcal{L}_{\text{contact}} = \text{WM}\Big(\mathcal{H}_\delta(\widehat{d}^{\perp}), m^c\Big), \tag{30}$$

where $\mathcal{H}_\delta$ is PyTorch's Huber loss with $\delta = 0.1$; unlike the millimetre targets and tolerances, δ is set directly in pointmap units. For hover frames, the code applies a relative dead band with $\epsilon_h = 0.2$:

$$e_i = \widehat{d}_i^{\perp} - t_i, \tag{31}$$

$$\widetilde{e}_i = \text{sign}(e_i)\,[|e_i| - 0.2|t_i|]_+, \tag{32}$$

$$\mathcal{L}_{\text{hover}} = \text{WM}\big(\mathcal{H}_\delta(\widetilde{e}), m^h\big). \tag{33}$$

Near, unknown, occluded, and invalid frames contribute neither contact nor hover loss. The two-sided residual corrects both under- and over-separation and has zero gradient inside the tolerance band.

These are the contact and hover terms of Equation 4. With the hover floor and the profile term of Appendix A.6, the full objective is

$$\mathcal{L} = \mathcal{L}_{\text{base}} + 3\,(\mathcal{L}_{\text{contact}} + \mathcal{L}_{\text{hover}} + \mathcal{L}_{\text{floor}}) + 0.3\mathcal{L}_{\text{profile}}. \tag{34}$$

Because the plane samples in Equation 20 are detached, the direct pointmap gradients of the contact, hover, and floor terms pass through p_i^{tip} and not the fitted tissue plane. The formulation adds no inference-time branch.

A.6 HOVER FLOOR, PROFILE TERM, AND RECIPE SETTINGS

Additive hover floor. The side-view target and primary stereo geometry can disagree, so every eligible hover frame retains its triangle-target term and also receives a one-sided clearance floor:

$$\mathcal{L}_{\text{floor}} = \text{WM}\Big(\mathcal{H}_{0.1}\Big([\phi/\bar{\eta}_{b(i)} - \widehat{d}_i^{\perp}]_+\Big), m^h\Big), \qquad \phi = 1.04\,\text{mm}, \tag{35}$$

where $\bar{\eta}_b = \max(\eta_b, 10^{-6})$ is the clamped sequence normalization factor. ϕ is the 90th percentile of the training-set contact distances; the Youden-optimal threshold (1.18 mm) and the 95th percentile (1.24 mm) give nearly the same value. The loss and gradient vanish once the floor is satisfied, including for distant targets.

Normal-weight calibration. The unweighted per-pixel normal gradient was about 61 and 69 times weaker than the depth gradient on tissue and instruments, respectively, in the calibration diagnostic. A weight of 20 brings it to approximately 0.3 times that reference. This calibration sets the relative gradient scale; it is not an evaluation of the complete recipe.

REL and trimmed point regression. Let V_i be the valid depth pixels in frame i, $N_i = |V_i|$, and N_i^{inst} their instrument-pixel count. The per-frame weight normalizer used by the code is

$$w_i(u) = 1 + (\kappa - 1)M_i^{\text{inst}}(u), \qquad a_i = \max\left(1 + (\kappa - 1)\frac{N_i^{\text{inst}}}{\max(N_i, 1)}, 10^{-3}\right), \quad \kappa = 0. \tag{36}$$

Let $h_i(u)$ mark pixels retained by confidence-free error trimming and $k_i = \sum_{u \in V_i} h_i(u)$. On the trimmed datasets, the cutoff is the $\lceil 0.9N_i \rceil$-th smallest error among valid pixels, with ties retained; frames outside this list retain all valid pixels. Writing $\ell_i(u) = \widehat{C}_i(u)\,\varepsilon_i(u) - 0.02 \log \widehat{C}_i(u)$ for the already composed aleatoric term, with $\varepsilon_i(u)$ the inverse-depth-weighted point error averaged over its three coordinates, the actual valid-pixel reduction is

$$\mathcal{L}_{\text{aleatoric}}^{\text{REL,trim}} = \frac{\displaystyle\sum_i \sum_{u \in V_i} \frac{w_i(u)}{a_i} \frac{n_i h_i(u)}{\max(k_i, 1)} \ell_i(u)}{\displaystyle\sum_i n_i}. \tag{37}$$

The normalizer is computed before trimming; the final mean is over all valid pixels in the batch, not an unweighted mean of frames. Empty valid batches contribute zero. The 10^{-3} clamp handles all-instrument frames. Normal supervision retains its separate depth-edge and neighborhood masks, so removing instrument pixels from this regression term does not remove their normal supervision.

Profile term. The knife mask comes from the semantic annotation and is augmented together with the image and the other masks. The principal axis of the knife pixels, oriented towards the annotated tip, is partitioned into $K = 48$ bins that become denser towards the tip. For bin k of frame i, $\bar{z}_{ik}$ is the mean predicted camera depth of its knife pixels and $\bar{z}_{ik}^{\text{FS}}$ the mean FS depth of the same pixels; knife pixels without valid FS depth are left out of the FS mean but still define the bins, so both profiles are binned alike. The first 80% of the bins form the shaft and the remaining bins the tip. With $\bar{z}_i^{\text{shaft}}$ and $\bar{z}_i^{\text{FS,shaft}}$ the means over the shaft bins valid in both profiles,

$$\begin{aligned} \mathcal{L}_{\text{profile}} &= \text{WM}_{i,k}\Big(\big[\,|\delta_{ik}| - \epsilon_{\text{p}}/\bar{\eta}_{b(i)}\,\big]_+,\ \nu_{ik}\Big), \qquad \epsilon_{\text{p}} = 0.2\,\text{mm}, \\ \delta_{ik} &= (\bar{z}_{ik}^{\text{FS}} - \bar{z}_i^{\text{FS,shaft}}) - (\bar{z}_{ik} - \bar{z}_i^{\text{shaft}}), \end{aligned} \tag{38}$$

where ν_{ik} marks tip bins valid in both profiles, in frames with at least 400 knife pixels and at least five valid shaft bins; a bin is valid when it holds at least 15 knife pixels. The FS profile is detached, since it is data rather than a quantity to optimize, and gradients reach the prediction only through the per-bin means, so each correction is shared by all pixels of a bin. Because each profile is centred on its own shaft, a constant depth offset of the instrument leaves the term unchanged: it constrains shape and leaves placement to the contact and hover terms. The term applies to every ESD-CONTACT training frame with enough knife pixels, whatever its interaction state. FLOOR and the profile term convert their millimetre values with $\bar{\eta}_b$.

Exclusions and ablation settings. Contact and hover supervision is disabled on 0722-2 and 0728-2, whose worn knife tips make the tip annotations unreliable, by invalidating their tip and state fields; their dense supervision and the profile term remain, and without a valid tip the profile term orients the knife axis towards the upper-left image corner. In the ablations, removing NORM restores the normal weight to one while normal supervision on ESD-CONTACT stays enabled, and removing REL restores $\kappa = 1$.

A.7 CONTACT READOUT AND EVALUATION SAFEGUARDS

We evaluate whether the predicted pointmap preserves the contact/hover relation without scoring it against the triangle target used for training. The evaluator multiplies the local pointmap by $\widehat{s}$, takes its z coordinate, and bilinearly resizes this metric depth to the native image resolution. It then computes

$$g_i = \mathrm{med}_{u \in \mathcal{T}_i} \widehat{D}_i^m(u) - \mathrm{med}_{u \in \mathcal{R}_i} \widehat{D}_i^m(u), \tag{39}$$

Here $\mathcal{T}_i$ is a 7×7 box centered at the rounded annotated tip, intersected with valid instrument pixels; if fewer than three such pixels remain, the valid box is used without the instrument intersection. $\mathcal{R}_i$ is the valid boundary obtained by a 7×7 dilation of the instrument mask, subtraction of the original mask, and restriction to a 40-pixel radius around the tip; at least five ring pixels are required. Hover generally gives $g_i < 0$ because the tip is closer to the camera, while contact moves the distributions together; note that g_i has the opposite sign of the clearance $\widehat{d}_i^{\perp}$ of Equation 21, which is positive for hover. ROC-AUC computed from g_i and the contact/hover labels is the primary threshold-free metric. Annotated masks and tips are used so the benchmark isolates clearance geometry from segmentation and keypoint error.

Scale alignment and operating points. On the three stereo validation clips, the depth of every method, including CARE3R, is scaled per frame so that its median over tissue pixels (outside the instrument mask, with valid FS depth) matches that of FS. This places all methods on a common millimetre scale; for CARE3R it changes the four-clip AUC only from 0.9332 to 0.9329. The monocular clip has no FS reference, so its AUC uses the unaligned gap and its thresholded decisions use the gap divided by the median ring depth. ROC-AUC is computed within each clip and averaged with equal clip weight. Fixed-threshold balanced accuracy is secondary: at -0.5 mm, FS itself reaches only 0.650 on the stereo clips, because its contact gaps have a median of -0.68 mm. Appendix B.3 specifies the thresholds and their aggregation.

We also report the contact–hover separation

$$\Delta = \mathrm{med}(g_i \mid s_i = \mathrm{contact}) - \mathrm{med}(g_i \mid s_i = \mathrm{hover}), \tag{40}$$

pooled over the FS-aligned frames of the three stereo clips (Figure 7 and Table 8). The same statistic computed from FS depth gives the stereo reference. Because FS also supplies the tissue geometry of the triangle construction, this reference checks scale rather than providing independent ground-truth clearance; the discrete contact labels provide the independent evidence for ordering. Finally, the evaluator reports the median-scale-aligned tissue AbsRel against FS over the full frame and within 120 pixels of the tip, which would reveal a degraded surrounding surface. Conventional depth, pose, reconstruction, and segmentation evaluations test whether the broader geometry is preserved.

Blade profile consistency. A clearance loss could in principle be satisfied by bending the tip towards the camera, which would also enlarge the hover gap read by the benchmark. We therefore compare the predicted blade profile with FS. Along the tip-oriented principal axis of the knife mask, the evaluator takes the median depth in 30 bins at native image resolution and reads the tip dip

$$\mathrm{dip}_i = \min_{k \in \{24,\ldots,29\}} z_{ik}^{\mathrm{med}} - \max_{k \in \{0,\ldots,23\}} z_{ik}^{\mathrm{med}}, \tag{41}$$

which is negative when the tip comes closer to the camera than the shaft. It uses the tip-associated connected component of the knife mask and requires at least 400 knife pixels, a projected extent of 20 pixels, and 15 pixels per bin. On the hover frames of the three stereo clips, the median dip of CARE3R is positive, as for FS (1.24, 1.43, and 1.31 mm vs. 1.13, 0.93, and 0.83 mm), so the tip continues away from the camera; on two clips it exceeds that of FS, the opposite direction to a tip bent towards the camera. The dips of our models are read from their native metric depth, without FS

Table 5: **Training mixture.** Reference information lists the annotations each dataset provides; not all of them are supervised. Weights are relative sampling weights, not frame counts.

Dataset	Reference information	Weight
C3VD (Bobrow et al., 2023)	Registered phantom depth and camera poses	10,000
C3VDv2 (Golhar et al., 2026)	Phantom depth and camera poses	10,000
SimCol (Rau et al., 2023)	Synthetic depth and camera poses	8,000
SCARED (Allan et al., 2021)	Structured-light depth and camera poses	8,000
StereoMIS (Hayoz et al., 2023)	Camera poses; FoundationStereo stereo depth	8,000
Hamlyn (Mountney et al., 2010; Recasens et al., 2021)	FoundationStereo stereo depth	10,000
PolypVirtual (Liu et al., 2025)	Synthetic depth from the virtual subset (recorded poses are inconsistent and unused)	6,000
EndoSynth (Toussaint et al., 2025)	Depth and instrument masks; replicated single images	5,000
EndoVis18 (Allan et al., 2020)	Instrument segmentation annotations	5,000
ESDSEG	In-house instrument segmentation annotations	10,000
KidneyDepth (Oliva-Maza et al., 2025)	Synthetic depth; trajectory metadata	5,000
ESD-CONTACT	FoundationStereo stereo depth, masks, tips, states, side-view targets	8,000

alignment. The lower tail differs on one clip: on 0630-1, 4.8% of the hover frames of CARE3R dip below -1 mm, against 1.0% for FS and none for our model without the contact and hover terms, and the fifth percentile is -0.98 mm (FS $+0.61$ mm). The contact terms therefore add a small, localized shape deviation. Because changing k of the n_h hover frames of a clip can change its AUC by at most k/n_h, the frames that dip below -1 mm account for at most 0.048 of the AUC on 0630-1, where the paired gain of the contact terms is 0.083, and for at most 0.012 of the four-clip mean.

B EXPERIMENTAL SETUP AND EVALUATION PROTOCOLS

B.1 TRAINING DATASETS AND SUPERVISION

Table 5 lists the twelve training datasets. Available annotations differ across datasets, and dense depth includes rendered, registered, and stereo-derived sources. Sampling weights are configuration weights rather than dataset sizes.

Training schedule. Training starts from the first public 1B checkpoint of VGGT-Ω and has three stages. The base stage trains geometry, pose, and segmentation jointly on the eleven datasets other than ESD-CONTACT for 40 epochs of 1,000 iterations. Stage 1 adds the ESD-CONTACT training clips with FS depth supervision, still without CAS, and trains for 20 epochs of 1,000 iterations with a one-cycle schedule (5% warm-up; peak learning rate 3.6×10^{-5} for the heads and 3.6×10^{-6} for the encoder). Stage 2 starts from the last stage-1 checkpoint, enables CAS together with REL, FLOOR, and NORM, and trains for another 20 epochs of 1,000 iterations with peak learning rates of 10^{-5} and 10^{-6}. We report the last stage-2 checkpoint; no checkpoint is selected on the validation clips. The variant without contact supervision in Section 4 follows the same stage-2 recipe with the contact, hover, and floor terms disabled. The shaft-line variant of Table 2 uses the same stages but was trained before three corrections to the camera-pose supervision: SimCol translations are now converted from centimetres, PolypVirtual's inconsistent poses are excluded from the camera loss, and identity pseudo-poses (Hamlyn, ESD-CONTACT) are no longer supervised as static cameras. Re-training the full model with the corrected pipeline changed its four-clip contact AUC by 0.003 and each per-clip AUC by at most 0.009. The patch encoder, the camera and register tokens, and the first 12 aggregator levels stay frozen (621.3 M trainable parameters). Each sample holds 1 to 12 views of 100,000 to 255,000 pixels (about 316^2 to 505^2) with aspect ratios between 0.5 and 2. Stages 1 and 2 use random seed 1111 and four NVIDIA H20 GPUs per run.

B.2 Baselines and Comparison Settings

The geometry comparison uses seven released checkpoints without fine-tuning on our data: Pi3, DAv3 (DA3-LARGE-1.1), VGGT-Ω (1B), VGGT (1B), Fast3R, DUSt3R with global alignment, and Endo3R. For VGGT-Ω we use its first public 1B checkpoint, which also initializes our backbone; the re-trained checkpoint that its authors released later was not evaluated. The released Fast3R inference samples random image-index embeddings in every forward pass, so repeated evaluations differ slightly (by up to 0.013 in per-clip contact AUC); all Fast3R contact numbers come from one run with a per-frame record. These models have no instrument-segmentation head, so segmentation is not applicable to them. The segmentation comparison instead uses an automatically fine-tuned SAM2.1 Hiera-Tiny adaptation (Ravi et al., 2025); at evaluation it receives RGB without ground-truth prompts. A second segmentation baseline, MambaVision-T (Hatamizadeh & Kautz, 2025), combines the official ImageNet-1K backbone with a UPerNet decoder (Xiao et al., 2018) and a binary instrument output. It uses the same segmentation training data, split, and 10:5:5:8 dataset mixture as SAM2.1, selects its epoch on training-split development data only, and is refit on the full training split for 3 epochs (774 optimizer steps). Inference uses one 512×512 RGB frame without temporal context or prompts and the same evaluator, whereas SAM2.1 uses at most eight causal frames and a 2-epoch (390-step) refit. The two segmentation baselines therefore share data, split, and metrics but not temporal context or compute. The MambaVision code and pretrained weights carry non-commercial licenses.

The domain fine-tuning control DAv3 (FT) (Section 4.3) is not one of these seven checkpoints. It starts from DA3-LARGE-1.1 and trains only the depth/confidence head and camera decoder (41,065,707 of 410,941,767 parameters); the backbone, camera encoder, and unused ray branches stay frozen. Training uses our geometry and pose losses on the same geometry splits as CARE3R, FS depth on ESD-CONTACT, and it excludes the validation clips. The contact and hover terms, the profile term, FLOOR, REL, and the increased normal weight are disabled while the base normal term is kept, and segmentation-only datasets are omitted. The run takes 16k optimizer steps (20 epochs of 800 iterations) with seed 42. Its validation point error falls from 0.084 after the first epoch to 0.045 after about 9k steps and stays between 0.044 and 0.047 afterwards (0.045 at the end): this head-only adaptation converges within that budget on the endoscopic data and over-fits with longer training. Its final checkpoint was fixed before evaluation. It does not reproduce the official DA3 training recipe. SERV-CT is not used in this fine-tuning, and the model has no segmentation head.

The C3VD and SCARED evaluations use held-out test sequences, but both datasets occur in our training mixture, so these results are not zero-shot tests for CARE3R. SERV-CT is absent from our mixture and provides a zero-shot depth evaluation for CARE3R. For the external models, we cannot verify that these datasets are absent from their original training data.

For contact evaluation, all models receive RGB images, but annotated instrument masks and tip locations select the score regions. Stereo predictions are aligned per frame by the tissue median to evaluation FS reference depth before the aligned contact, and fixed millimetre-threshold comparisons. This reference-assisted alignment permits a common stereo scale but is not uncalibrated metric prediction or a fully automatic contact detector. The monocular clip has no FS reference, so it is scored only with within-clip ranking metrics and the normalized threshold. A single positive rescaling over a clip preserves score ordering for AUC; frame-varying alignment need not do so. Endo3R's contact AUC is below 0.5 because it places hovering tips slightly behind the tissue (median hover gap $+0.04$ to $+0.17$ mm on the stereo clips); we report the score as measured rather than flipping its sign. Figure 7 omits Fast3R and DUSt3R+GA, whose separation Δ is 0.03 and 0.02 mm. Fast3R, DUSt3R, and Endo3R receive frames in landscape orientation (Table 6); with native portrait input, the four-clip contact AUC is 0.488 for Endo3R and 0.536 for Fast3R.

B.3 Contact Operating Points and Aggregation

The score in Equation 39 treats contact as positive and hover as negative; near, unknown, and occluded-contact frames are excluded. The fixed decisions of Table 1 predict contact when $g_i \geq -0.5$ mm on the FS-aligned stereo clips (457 contact and 1,662 hover frames), or when $g_i / \operatorname{med}(\widehat{D}_i \mid \mathcal{R}_i) \geq -0.01$, a scale-free rule used for all four clips (843 contact and 1,963 hover frames) and for the monocular clip alone. Both thresholds are fixed physical priors, not fitted to the evaluation labels, and neither measures true perpendicular clearance. Contact AUC is computed

within each clip and averaged with equal clip weight, whereas bAcc pools the confusion counts of all frames in a scope. AP is the average precision of contact, computed and averaged like AUC; ρ is the Spearman correlation between the predicted and FS gaps over all frames of a stereo clip, averaged over the three clips.

B.4 Task-specific Evaluation Details

Pose. The angular evaluator forms relative poses, measures rotation and translation-direction errors, and reports AUC at 5, 15, and 30 degrees. The AUC implementation integrates a cumulative one-degree-bin histogram of the maximum of those errors over pooled image pairs and reports percentages; Table 3 reports AUC@30. The trajectory evaluator uses Sim(3) scale-corrected alignment, stride 10, ATE RMSE, and one-frame relative translation/rotation RMSE. The C3VD and SCARED trajectory evaluations each use eight sequences. ATE and relative translation errors are in dataset millimetres; relative rotation error is in degrees. Aligned trajectory errors do not establish native translation scale.

Depth. AbsRel measures valid-pixel absolute relative depth error, and δ_1 is the fraction within a predicted/reference ratio of 1.25. Single-image C3VD, SCARED, and SERV-CT shape comparisons use separate ground-truth median scaling for each image. Reported dataset summaries are valid-pixel-weighted means of image metrics, including RMSE, rather than RMSE recomputed over pooled pixels. Unaligned absolute-metric depth is reported separately for CARE3R. On C3VD video, CARE3R uses its native metric output without test-scale fitting, whereas external models receive one ground-truth median scale per sequence; these are different scale protocols. SCARED uses dataset-8/9 test sequences. SERV-CT CT reference depth is in millimetres and excludes occluded pixels.

Reconstruction. C3VD reference points are reconstructed from depth and reference cameras. Predicted points use the model's global geometry. Eight sequences undergo Umeyama Sim(3) alignment (Umeyama, 1991) followed by ICP (Besl & McKay, 1992). Predicted-to-reference nearest-neighbor distance is accuracy; reference-to-predicted distance is completeness, both in millimetres. Matched normals give normal consistency. Dataset values average per-sequence statistics, so the reported median columns are means of sequence medians rather than pooled-point medians. All methods use the same view selection, valid masks, and alignment.

Segmentation. The evaluator resizes predictions to native ground-truth resolution with nearest interpolation and excludes every ground-truth-255 pixel from class true positives, false positives, and false negatives. It pools class counts across images before computing background/instrument IoU and Dice; mIoU and mDice average the two class scores, whereas instrument IoU/Dice refer only to the foreground. ESDSEG tissue labels map to background; EndoVis18 tissue labels map to ignore.

B.5 Runtime and Memory

Table 6 measures each method's N-frame forward pass on one NVIDIA RTX A6000 using ESD-CONTACT frames, after loading and transfer; checkpoint loading is excluded. Two warmups precede five synchronized timings. Autocast is bfloat16 except DA3, which runs in fp32; DUSt3R includes global alignment and Fast3R/Endo3R include CPU PnP. Peak memory is allocated GPU GiB including the loaded model. Different native input sizes and post-processing prevent a matched-resolution speed ranking.

At the accuracy-evaluated resolution, CARE3R processes eight frames in 0.9529 s (8.40 frames/s). Its separate 512-long-side timing reaches 27.66 frames/s at $N = 8$, but that setting has no measured accuracy point. The 0.1252-s $N = 1$ timing and the eight-frame throughput answer different latency questions; neither establishes real-time clinical deployment.

Table 6: **Runtime on one RTX A6000.** Latency is seconds for one complete N-frame forward, and FPS=N/latency is batch throughput, not single-frame online latency for $N > 1$. Inputs are backbone $H \times W$. Peak allocated memory includes model weights. The long-side-512 CARE3R row is a *timing-only* configuration with no corresponding accuracy evaluation; accuracy tables use 912×512.

Method	Input	Precision	$N = 1$ latency (s)	$N = 8$ latency (s) / FPS	Peak GiB ($N = 8$)
VGGT	518×294	bf16	0.1141	0.4555 / 17.56	7.21
Fast3R	288×512	bf16	0.0848	0.6681 / 11.97	5.19
DUSt3R+GA	288×512	bf16	0.1584	12.5139 / 0.64	3.59
Endo3R	288×512	bf16	0.1298	1.4005 / 5.71	2.95
VGGT-Ω	688×384	bf16	0.0761	0.4662 / 17.16	5.62
DAv3	504×280	fp32	0.0519	0.2275 / 35.17	3.80
Pi3	910×518	bf16	0.1488	1.0926 / 7.32	4.37
CARE3R	912×512	bf16	0.1252	0.9529 / 8.40	9.02
CARE3R @ long side 512	512×288	bf16	0.0786	0.2892 / 27.66	6.11

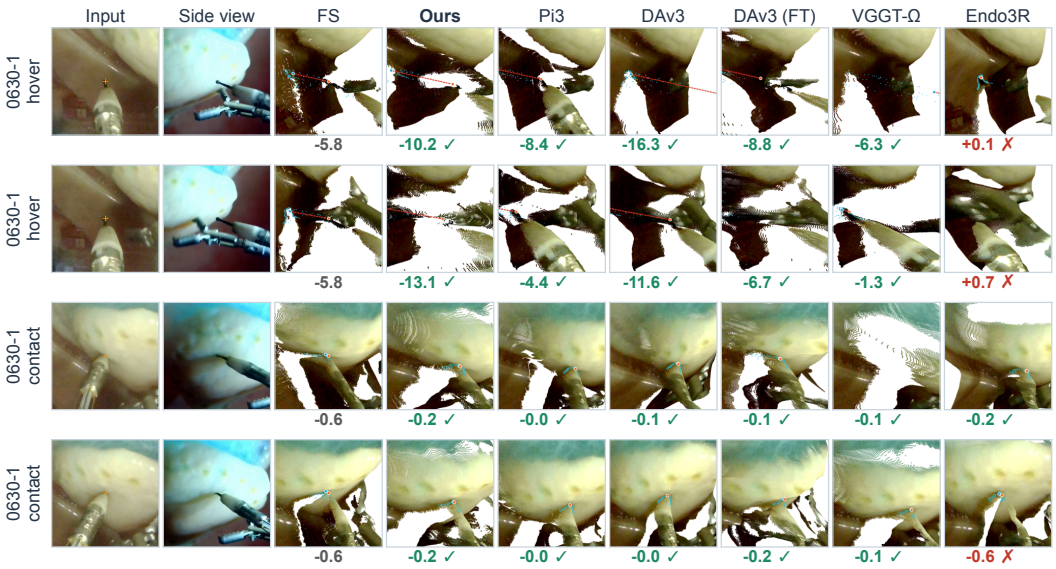


Figure 8: **Remaining model-blind candidates on validation clip 0630-1.** Layout and notation follow Figure 5. CARE3R recovers all four states; on the two hover frames every baseline except Endo3R also recovers hover, and CARE3R overestimates the FS gap (−10.2 and −13.1 vs. −5.8 mm).

C Additional Contact Visualizations

Figures 8–10 show the 12 model-blind candidate frames that do not appear in Figure 5. Together with the three rows of Figure 5, which were picked from the rendered candidates for clarity, they complete the 15 candidates of the selection rule, three hover and two contact frames per stereo clip, ranked by how close their FS gap is to the clip and state median before any model output was inspected. Rendering, gap readout, and the $\tau = -0.5$ mm decision follow Figure 5. CARE3R recovers all seven remaining hover frames and misclassifies two of the five remaining contact frames: −1.5 mm on 0714-2, where FS itself measures −1.0 mm, and −1.0 mm on 0828-1, where FS measures −0.55 mm. On several hover frames, CARE3R places the tip farther from the tissue than FS does (for example, −13.1 vs. −5.8 mm on 0630-1), consistent with its separation exceeding the stereo reference in Figure 7. The baselines also recover several of these hover frames, including both remaining hover frames of 0630-1 for every baseline except Endo3R. The collapsed hover gaps in Figure 5 are therefore frequent rather than universal, and Figure 7 provides the distribution-level comparison. Figure 11 shows the per-frame gap on the third stereo clip.

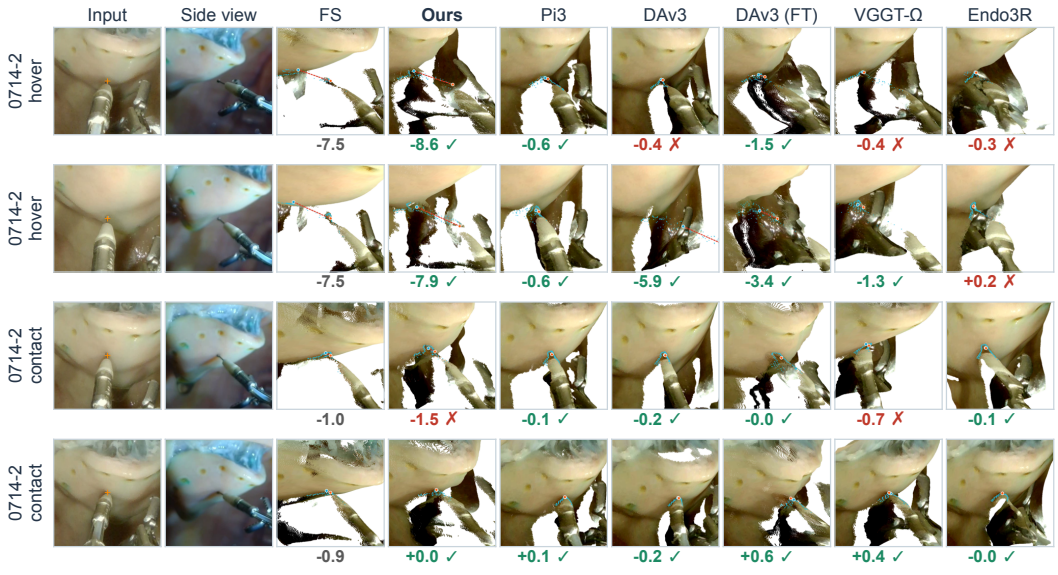


Figure 9: **Remaining model-blind candidates on validation clip 0714-2.** Layout and notation follow Figure 5. CARE3R recovers both hover frames but misclassifies the first contact frame (-1.5 mm), on which FS itself measures -1.0 mm.

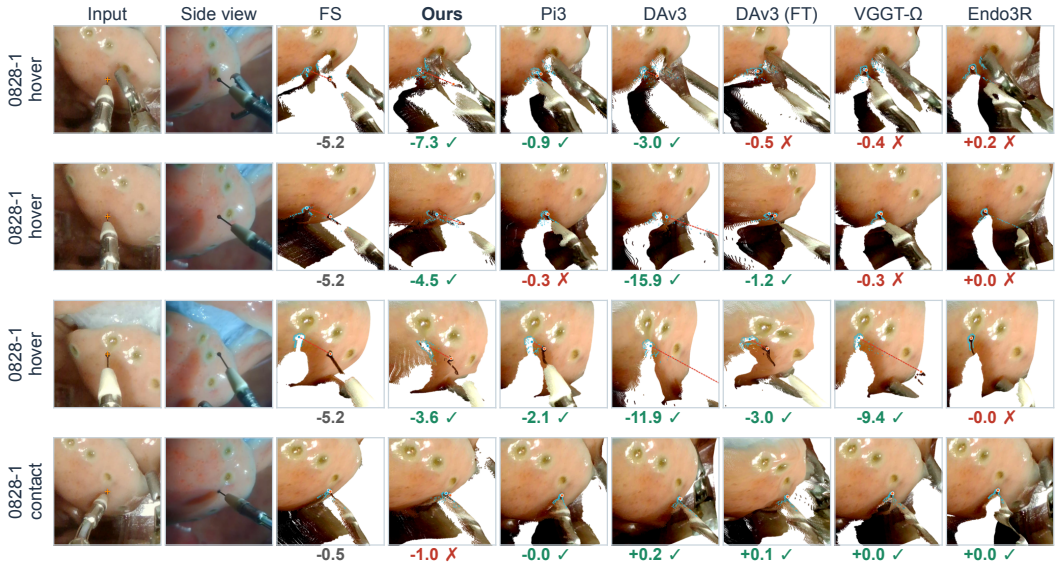


Figure 10: **Remaining model-blind candidates on validation clip 0828-1.** Layout and notation follow Figure 5. CARE3R recovers all three hover frames and misclassifies the contact frame (-1.0 mm), on which FS measures -0.55 mm.

D Instrument Segmentation

Segmentation is an auxiliary output of CARE3R. We compare it with two dedicated segmenters trained on the same segmentation data and split: SAM2.1 (Ravi et al., 2025), adapted automatically and run on up to eight causal frames without ground-truth prompts, and MambaVision-T (Hatamizadeh & Kautz, 2025) with a UPerNet decoder (Xiao et al., 2018), which segments single frames (Appendix B.2). Table 7 reports dataset-pooled mIoU and instrument Dice with the evaluator

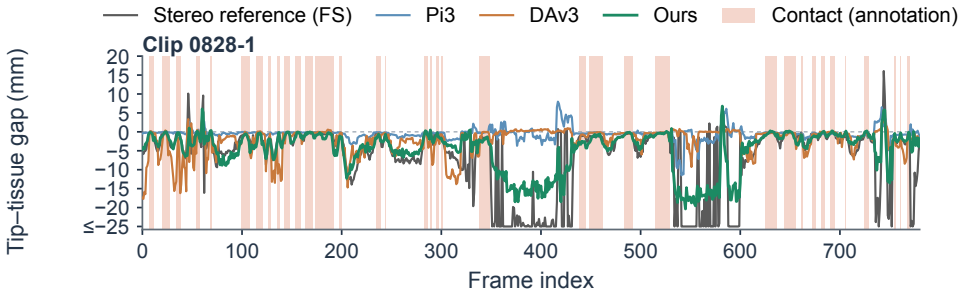


Figure 11: **Per-frame tip–tissue gap on validation clip 0828-1.** Layout and alignment follow Figure 6; gaps below -25 mm are drawn at -25 mm, which affects only FS. In the segments where FS falls below the display floor, the tool hovers high above the tissue, as after frame 525 of 0714-2, and CARE3R again reports a smaller clearance magnitude than FS.

Table 7: **Binary instrument segmentation.** Dataset-pooled mIoU and instrument Dice (↑). SAM2.1 and MambaVision-T are trained on the same segmentation data and split as CARE3R; none of the three is trained on Kvasir, whose columns use the 118-image official test subset. ESD-CONTACT covers the four validation clips. Top-3 results are highlighted as first, second, and third.

	ESDSEG		EndoVis18		ESD-Contact		Kvasir (official test)	
Method	mIoU ↑	Inst. Dice ↑	mIoU ↑	Inst. Dice ↑	mIoU ↑	Inst. Dice ↑	mIoU ↑	Inst. Dice ↑
MambaVision-T	0.8967	0.9090	0.8971	0.9141	0.9664	0.9695	0.6022	0.4516
SAM2.1	0.9308	0.9402	0.9385	0.9504	0.9799	0.9820	0.7473	0.7077
CARE3R (Ours)	0.9557	0.9626	0.9078	0.9252	0.9777	0.9800	0.7568	0.7238

of Appendix B.4. CARE3R is best on ESDSEG and on the zero-shot Kvasir test subset and second to SAM2.1 on EndoVis18 and ESD-CONTACT. On all 590 Kvasir images, which include the 118 test images, mIoU / instrument Dice are 0.696 / 0.637 for CARE3R, 0.694 / 0.626 for SAM2.1, and 0.570 / 0.388 for MambaVision-T. The comparison shares data, split, label mapping, and evaluator, but not temporal context or training compute. Figure 12 shows two frames per dataset, selected from the ground truth alone; the second Kvasir frame, with a dark instrument, is a failure case for CARE3R and MambaVision-T, whereas SAM2.1 recovers most of the instrument.

E ADDITIONAL ABLATIONS

FLOOR, REL, and NORM. Table 8 adds the three recipe settings of Section 3.3 cumulatively on top of the contact and hover terms and the profile term. Every arm starts from the same contact-free checkpoint with a fresh optimizer; this checkpoint comes from a separate, shorter FS-supervised run (8k steps, seed 42, two GPUs), not from the stage-1 run that initializes CARE3R. All arms share the code, data mixture, validation split, seed, and a 10×800-step schedule, and each is scored at its final checkpoint. The contact AUC stays within 0.926–0.928, with per-clip changes of at most 0.009 that change sign across clips, so these settings do not drive contact discrimination. They govern the gap magnitude and surface quality instead. FLOOR leaves Δ and the hover median essentially unchanged, because its 1.04 mm bound on the clearance lies far below the typical hover clearance. REL brings them closest to FS (5.05 and -5.85 mm vs. 5.27 and -5.95 mm) and lowers the instrument's normal dispersion. NORM gives the smoothest surfaces while pushing Δ beyond FS. Arm D re-trains the complete recipe with a shorter schedule, a different initialization, and a different seed than the reported model, so its values (AUC 0.928 vs. 0.933; Δ 5.71 vs. 6.72 mm) are a reproduction check rather than a controlled comparison.

Other tasks. Table 9 reports the non-contact metrics of the four arms, scored with the protocols of Section 4. Differences across arms are small on every row, so none of FLOOR, REL, or NORM trades contact geometry for conventional depth, pose, reconstruction, or segmentation quality. Be-

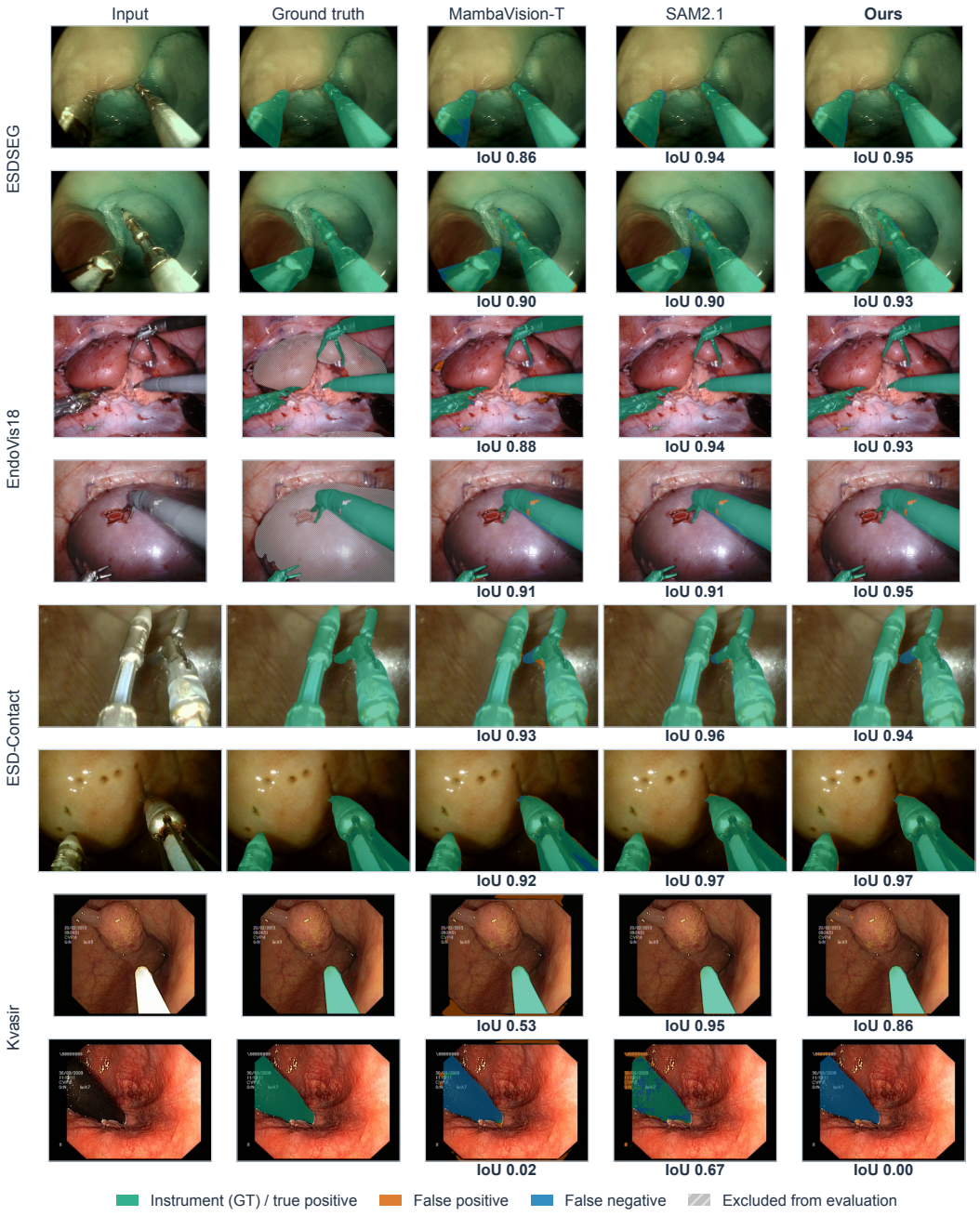


Figure 12: **Qualitative instrument segmentation.** Each pair of rows shows two frames from ESDSEG, EndoVis18, the ESD-CONTACT validation clips, and the official test split of Kvasir-Instrument, which none of the three methods saw in training. The ground-truth column marks instrument pixels in green and pixels excluded from evaluation with grey hatching; in the method columns, green, orange, and blue mark true-positive, false-positive, and false-negative instrument pixels relative to the ground truth, and excluded pixels are left untinted. Frames were selected from the ground truth alone, as the images whose instrument area fraction is closest to the 1/3 and 2/3 quantiles over the frames of each evaluated set that contain an instrument, and the number under each method panel is that image's instrument IoU over the full frame (ESD-CONTACT panels are 3:2 crops centred on the ground-truth instruments).

Table 8: **Cumulative ablation of FLOOR, REL, and NORM.** Every arm keeps the contact and hover terms and the profile term (PROFILE) and adds one setting to the arm above, so each step measures a conditional increment. AUC is the mean over the four validation clips; Δ and the median hover gap pool the three stereo clips after per-frame FS alignment. The FS row is the stereo reference, and its AUC (†) covers only the three stereo clips because the monocular clip has no stereo depth. Normal dispersion is the median angle between each predicted normal and its 5×5-smoothed version on tissue / instrument pixels; it measures roughness, not error against reference normals. Single seed.

	Modules					Contact	Gap (mm)		Normal disp. (°) ↓
Arm	Contact/hover	PROFILE	FLOOR	REL	NORM	AUC ↑	Δ	Hover med.	Tissue / instr.
FS reference						0.892†	5.27	−5.95	4.6 / 5.0
A	✓	✓				0.9257	4.72	−5.50	21.3 / 28.1
B	✓	✓	✓			0.9264	4.65	−5.45	21.3 / 28.0
C	✓	✓	✓	✓		0.9269	5.05	−5.85	21.3 / 25.6
D	✓	✓	✓	✓	✓	0.9281	5.71	−6.61	17.3 / 16.2

Table 9: **Non-contact metrics of the cumulative ablation.** Arms A–D as in Table 8. Depth AbsRel is median-scaled per image unless marked absolute; video AbsRel uses metric output. SERV-CT is zero-shot. Pose reports AUC@30 (%) and Sim(3)-aligned ATE (mm); reconstruction reports per-sequence C3VD medians. Segmentation mIoU uses the evaluator of Appendix B.4; Kvasir uses all 590 images. Absolute values are not comparable with Tables 3 and 7 (different initialization and schedule). Single seed.

Metric	A	B	C	D
C3VD single-image AbsRel ↓	0.0761	0.0767	0.0770	0.0760
SCARED single-image AbsRel ↓	0.0459	0.0462	0.0459	0.0455
C3VD single-image AbsRel, absolute ↓	0.1446	0.1480	0.1487	0.1462
C3VD video AbsRel, metric ↓	0.1544	0.1574	0.1569	0.1579
SERV-CT AbsRel (zero-shot) ↓	0.0545	0.0543	0.0540	0.0530
Pose AUC@30, C3VD / SCARED ↑	62.69 / 65.65	63.05 / 65.68	62.86 / 65.81	62.81 / 65.92
ATE, C3VD / SCARED (mm) ↓	1.012 / 2.329	1.011 / 2.335	1.020 / 2.327	1.016 / 2.314
C3VD Acc.-med. / Comp.-med. (mm) ↓	1.812 / 4.752	1.808 / 4.735	1.808 / 4.759	1.800 / 4.733
C3VD N.C.-med. ↑	0.713	0.714	0.714	0.719
Seg. mIoU, ESDSEG / EndoVis18 ↑	0.919 / 0.944	0.918 / 0.944	0.919 / 0.944	0.915 / 0.944
Seg. mIoU, ESD-Contact / Kvasir ↑	0.978 / 0.706	0.978 / 0.708	0.978 / 0.707	0.978 / 0.704

cause the initialization, schedule, and seed differ from the reported model, only differences within Table 9 are meaningful; its absolute values are not comparable with Tables 3 and 7 (for example, EndoVis18 mIoU is 0.944 here and 0.908 for CARE3R, and C3VD ATE is about 1.02 vs. 0.61 mm).