

Recent advances in synthetic virtual 3D endoscopy to assist in skull base surgery

Victor E. Staartjes , Olivier Zanier , Kenneth Arockia , Stefanos Voglis , Giovanni Muscas , Andrea Boschi , Alessandro Della Puppa , Adrian Elmi-Terander , Luca Regli & Carlo Serra

To cite this article: Victor E. Staartjes , Olivier Zanier , Kenneth Arockia , Stefanos Voglis , Giovanni Muscas , Andrea Boschi , Alessandro Della Puppa , Adrian Elmi-Terander , Luca Regli & Carlo Serra (2026) Recent advances in synthetic virtual 3D endoscopy to assist in skull base surgery, Expert Review of Neurotherapeutics, 26:4, 371-380, DOI: [10.1080/14737175.2026.2642838](https://doi.org/10.1080/14737175.2026.2642838)

To link to this article: <https://doi.org/10.1080/14737175.2026.2642838>



© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



View supplementary material [↗](#)



Published online: 15 Mar 2026.



Submit your article to this journal [↗](#)



Article views: 574



View related articles [↗](#)



View Crossmark data [↗](#)

REVIEW



Recent advances in synthetic virtual 3D endoscopy to assist in skull base surgery

Victor E. Staartjes^{a,b}, Olivier Zanier^a, Kenneth Arockia^a, Stefanos Voglis^a, Giovanni Muscas^c, Andrea Boschi^c, Alessandro Della Puppa^c, Adrian Elmi-Terander^{b,d}, Luca Regli^a and Carlo Serra^a

^aMachine Intelligence in Clinical Neuroscience (MICN) & Microsurgical Neuroanatomy Laboratory, Department of Neurosurgery, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Zurich, Switzerland; ^bDepartment of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden; ^cDepartment of Neurosurgery, Careggi University Hospital, Florence, Italy; ^dDepartment of Neurosurgery, Region Örebro County, Örebro University Hospital, Örebro, Sweden

ABSTRACT

Introduction: The advent of the operating microscope revolutionized neurosurgery by introducing stereoscopic visualization and enabling microneurosurgery. Endoscopy, in turn, brought a compact, wide-angle view through minimal access corridors but sacrificed depth perception. Although 3D endoscopes partially restore stereopsis, they remain costly and hardware dependent. Synthetic or virtual 3D endoscopy aims to recreate depth perception algorithmically from standard 2D endoscopic footage using computer vision and deep learning.

Areas covered: This expert review delineates the three main methodological pillars of depth estimation in endoscopy (motion-based, shading-based, and learning-based approaches) and outlines how recent advances such as MAST3R (Matching and Stereo 3D Reconstruction) and DINOv2/3 foundation models have transformed dense depth prediction. The authors highlight datasets, algorithmic trade-offs, and challenges specific to skull base surgery, including reflective surfaces, deformation, and limited texture. The review synthesizes evidence showing that software-based virtual 3D reconstruction can achieve relatively low-latency performance and generate geometrically consistent, visually realistic views without dedicated hardware.

Expert opinion: Virtual 3D endoscopy may enhance hand-eye coordination, depth judgment, and training for junior surgeons. While measurable clinical gains may remain modest, software-based, hardware-agnostic 3D visualization offers a pragmatic, scalable alternative to expensive optical 3D systems, potentially democratizing stereoscopic vision in skull base and endoscopic neurosurgery.

ARTICLE HISTORY

Received 15 October 2025
Accepted 5 March 2026

KEYWORDS

Skull base; depth prediction; virtual 3D; synthetic imaging; microsurgery; endoscopy; neurosurgery

1. Introduction

The introduction of the operating microscope into surgery revealed a completely new world of possibilities within surgery. First conceived by Carl Zeiss in 1921, it was first used within otolaryngology and has since seen standard adoption in this field, particularly for otolaryngological skull base surgery [1]. Within neurosurgery, it was first applied by Theodore Kurze around 1957 for a patient with trigeminal neuralgia [2]. The real breakthrough and democratization of microsurgery, however, appeared after M. Gazi Yasargil's work including but not limited to his seminal book series 'Microneurosurgery,' which established working with the operating microscope not only as a pure technical adjunct but as a novel philosophy utilizing the added visualization provided by the microscope [3]. He describes in his books that the key advantages of the operating microscope are not only magnification and illumination at depth – those already enabling a view into a novel world of detail hitherto unknown and enabling suddenly the treatment of complex diseases with good clinical outcomes and patient safety – but emphasizes first and foremost the

ability of binocular microscopes to maintain depth of field even when peering through a small 'keyhole.' This ability to see depth is lost on the human eye without a stereoscopic microscope – the human eye cannot perceive depth information in a deep hole. This revolution enabled the true benefit of microsurgery, which has since then been the greatest advancement in neurosurgery.

In a parallel vein, neuro-endoscopy was developed to provide an alternate, more 'close-up' or 'blown-up' view through natural cavities or minimal-invasive burr holes for various, particularly skull base and intraventricular, surgical procedures. While, Philipp Bozzini introduced the idea of endoscopy first around 1806 [4], this began in neurosurgery in 1910 with Victor Lespinasse, who performed the first neuro-endoscopic procedure, namely a coagulation of the choroid plexus in hydrocephalus using a rigid cystoscope [5]. Later, in 1923, William Mixer executed a third ventriculostomy using an endoscope [5]. Since then, endoscopy has – particularly in recent decades – seen a massive pickup in popularity, and fields such as spinal endoscopy are being rediscovered as truly minimally invasive

CONTACT Victor E. Staartjes ✉ victoregon.staartjes@usz.ch Machine Intelligence in Clinical Neuroscience (MICN) Laboratory, Department of Neurosurgery, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Frauenklinikstrasse 10, Zurich 8091, Switzerland

Supplemental data for this article can be accessed online at <https://doi.org/10.1080/14737175.2026.2642838>

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Article highlights

- The operating microscope introduced stereoscopic depth perception to neurosurgery, fundamentally transforming surgical precision and safety.
- Conventional 2D endoscopy lacks depth information, forcing surgeons to rely on indirect visual cues such as motion parallax and anatomical landmarks.
- 3D endoscopes restore stereopsis but remain limited by cost, weight, and lack of universality across systems.
- Software-based virtual 3D endoscopy aims to infer depth from standard 2D video, creating a stereoscopic experience without dedicated hardware.
- Three primary methodological paths enable virtual 3D reconstruction: 1) Depth from motion (structure-from-motion/SLAM) – exploits endoscope movement to infer 3D geometry. 2) Depth from shading – models near-field illumination to derive local surface topography. 3) Learning-based methods – use self-supervised or foundation model features (e.g. DINOv2/3, MAST3R) to predict dense depth.
- Emerging generative 3D priors, such as MAST3R, regularize both geometry and appearance, stabilizing reconstructions in texture-poor or reflective surgical fields.
- Virtual 3D endoscopy is particularly valuable for junior surgeons transitioning from microscopy to endoscopy, enhancing spatial awareness and confidence.
- While measurable clinical outcome differences may be small, the approach offers a scalable, cost-effective means of democratizing stereoscopic visualization in skull base and minimally invasive neurosurgery.

options. Of course, in other fields such as laparoscopy, endoscopic visualization has become the unquestionable standard. Even in neurosurgery, many procedures such as ventriculotomy or transsphenoidal surgery for pituitary tumors are most commonly carried out endoscopically nowadays [6].

A conventional operating endoscope, however – or any conventional endoscope including in e.g. colonoscopy – does not visualize depth stereoscopically. Thus, the endoscope, while offering several other potential situative benefits compared to microscopy such as a smaller footprint and different angle of view as well as excellent illumination, does not provide this crucial information for microsurgical techniques. However, experienced neurosurgeons are able to provide very safe endoscopic skull base surgery even without the added benefit of depth perception, which is explained by our ability to take into account subtle visual cues that provide indirect depth information (such as shadows, color gradients, reflections) as well as by the fact that – having studied microsurgical anatomy (e.g. that of the endonasal transsphenoidal approach), we can often infer the depth of structures relatively to each other by recognizing anatomical key points and landmarks with known and consistent relative positions. Other phenomena such as motion parallax due to camera movement (relative motion) provide dynamic depth cues, too. It has been shown, though, that only experienced surgeons can get 3D depth perception through such clues without using 3D visualization and that novices instead profit more from 3D visualization [7–11].

From microsurgery, however, we know that adding depth information – stereoscopic vision – is invaluable, and this is particularly true in skull base microsurgery. Even a millimeter of error can cause injury to vital structures such as venous sinuses and the carotid arteries in the cavernous sinus. Depth perception is necessary for precise micromovements in the

third axis and thus allows the microsurgeon to judge distances when dissecting tumors off critical neurovascular tissue – in the end, surgery happens in three dimensions, not two. But while standard endoscopes lack 3D visualization capabilities, even in endoscopy the importance of adding depth perception has now been realized: 3D endoscopes do exist, and some centers regularly employ such devices. It has been shown that adding depth cues via employment of a 3D endoscope improves surgical dexterity and particularly for junior neurosurgeons improved the learning curve markedly [12]. 3D endoscopes employ dual cameras or special stereo optic tips which enable generating a 3D view on a 3D screen, requiring the surgeon to wear 3D glasses. In practice, surgeons report a feeling of greater hand-eye coordination and precision [8,12]. Does 3D endoscopy lead to measurably better clinical outcomes or reduced complications? This will be, in our opinion, very hard to demonstrate, since severe complications due to misjudgment of anatomy do happen but are rare. This makes it difficult to study or to show superiority in terms of complications in a trial. When these events happen, though, they are potentially devastating for the patient, and even avoiding such a misjudgment in 1 in 1000 patients due to improved visualization would – in our view – be worthwhile. Some studies have been performed: in a literature study, no significant difference in extent of resection, operating time, or complications could be shown between 3D and 2D groups in pituitary endoscopic surgery. There are signs toward, as mentioned, improved subjective fine motor skills and learning curves, which are certainly relevant [13,14]. This could potentially lead to improved surgical safety particularly for more complex endoscopic skull base procedures, such as trans nasal resection of trigeminal schwannoma [15] or trans-eustachian tube cerebrospinal fluid leak repair [16,17].

Nonetheless, even though 3D endoscopes already exist as hardware solutions, they are not without their limitations. Initial 3D endoscopes were bulkier with thicker tip diameters. They are also heavier, the higher weight leading to potentially more fatigue in the surgeons' hands. Then, it is also true that cost is a significant barrier here: 3D endoscopes and their monitors are significantly more expensive. Here, it must be said that a lot of neurosurgical centers have some form of 3D monitor, because these are nowadays often used to enable 3D viewing of microsurgical procedures (the microscope always being a stereoscopic, 3D microscope). While some of these limitations have become a bit less pronounced, with standard 4 mm 3D endoscopes being developed and costs going down a bit, it still is unrealistic for most centers worldwide to change their endoscopes to a 3D endoscope – the benefit is there for sure, but it is subtle and in many situations does not justify entirely replacing a perfectly good 2D system.

Nowadays, with the graphical processing power available to most clinicians in the 21st century, interest has moved greatly toward software-based solutions to formerly hardware problems. The presented dilemma is exactly one such case: 3D endoscopy is desired but hitherto requires a dedicated, expensive 3D endoscope. Software-based solutions are always very difficult to engineer and validate – but once they have been developed and rigorously validated, they can in principle be rolled out quite easily and broadly in a variety of settings, even

importantly in low-middle income countries and rural settings. Particular interest is vested in hardware-agnostic software solutions (meaning that the software solution can be applied to any kind of hardware/endoscope of conventional 2D type, regardless of model and manufacturer). This requires a robust technique for depth prediction from 2D images which in turn allows creation of virtual 3D images and – in the optimal use-case – real-time conversion into a synthetic 3D video feed that is displayed live to the surgeon. This is an exquisitely difficult machine learning problem (although it can also be tackled with non-learning-based methods, as will be elucidated herein) that has therefore received a lot of interest from biomedical engineers and data scientists in recent years. In this expert review article, we aim to elucidate not only the use-case of virtual 3D endoscopy as we already have done above but also to provide an in-depth overview of the state-of-the-art in methodologies and applications of the depth prediction methods that underly virtual 3D endoscopy.

2. Methodological foundations: the three basic paths toward depth estimation

2.1. No. 1. – Depth from motion: parallax, pose, and reconstruction when the endoscope is moving

If one accepts the endoscope's constant micro-movements as signal rather than nuisance, then depth naturally falls out of parallax. The key intuition here is the one that every surgeon already inherently has with a gentle 'waggle' or micromovement of the scope: nearer anatomical structures sweep across the field of view faster than further away tissue. This approach can be called structure-from-motion (SfM) or, more contemporaneously, visual SLAM (Simultaneous Localization and Mapping).

Mathematically, the pipeline alternates between estimating the endoscope's six-degree-of-freedom motion between frames and triangulating points that remain visible across those frames; a mapping process then densifies those points into a surface and renders a depth map back into the current view. In endoscopy, this must all be done under difficult optics: very short working distances, wide fields of view, radial distortion, and near-point lighting. Datasets such as EndoSLAM [18] (porcine gastrointestinal endoscopy with synthetic and real depth) and EndoMapper [19] (24+ hours of calibrated clinical video) catalyze this progress by providing sequences on which to measure pose error and reconstruction fidelity. Newer endoscopy-oriented frameworks (e.g. OneSLAM, CudaSIFT-SLAM, and BodySLAM) [20–22] show how these ideas port across gastroscopy, colonoscopy, and laparoscopy while keeping near real-time speed – in practice, this gives a coherent 3D scaffold of the scene from nothing more than the 2D feed [18–23]. Motion-based pipelines produce metrically consistent geometry once the scale is fixed, which makes measurement, AR overlays, and trajectory guidance feasible. They are also lighting-robust compared with shading-based methods, because geometry is inferred from correspondence, not brightness. Where they struggle is exactly where skull base surgery is hard for humans too: low texture (smooth bone or blood film making key point detection difficult), deformation (instrument movement such as suction/forceps, pulsation, soft tissue manipulation), and self-

occlusion of the view from instruments. These violate assumptions of feature richness, rigidity, and stable visibility that triangulation prefers. The literature's answer is twofold: tracking is decoupled from mapping with short, reliable keyframe windows for pose estimation, while mapping integrates over longer timespans with outlier detection and rejection. Then, dense reconstructions are stabilized by photometric bundle adjustment or by fusing short-range depth surrogates (optical flow, learned depth) where features are sparse. Still, the problem of scale realism maintains, because monocular SfM determines only relative depth. Surgical systems therefore fix scales with either an external tracker on the scope (as is common in neuronavigation), or a tool prior (a suction tip or curette of known diameter visible in-frame), or with registration to pre-op CT/MR of the sinonasal corridor. With any one of these anchors, low-millimetric surface error is reachable in phantoms and clinical video, which is why motion-based reconstructions frequently underpin AR experiments and are thus potentially attractive for skull base navigation as well [19–22].

Two newer developments are worth calling out because they directly address endoscopy's specific challenges. First, topological or multi-map SLAM splits a long procedure into locally consistent submaps and then stitches them together, improving robustness to large, repetitive anatomy and enabling re-localization after occlusions of the field of view. This has been shown on EndoMapper colonoscopies, even showing efficient real-time use based on GPUs. Second, learned priors for matching and reconstruction (e.g. MAST3R) [24,25] can replace sometimes unreliable key point matching with dense, 3D-aware correspondences, which increases pose stability in low-texture or specular scenes and provides better initial maps for later fusion. Intuitively, these priors give the SLAM system a 'sense' of plausible 3D even when classic parallax cues are weak for a few frames. This corresponds exactly to those moments that challenge endoscopic surgeons and algorithms alike, and where such assistive measures might be most powerfully implemented [20,21,24,25].

2.2. No 2. – Depth from shading: near-field constant illumination encoding depth information

Endoscopic optics offer a second path to depth that does not rely on motion: near-light photometry. Because the light source on a rigid endoscope sits only millimeters from the lens, the brightness of a patch of tissue depends strongly on its tilt and distance from the light source. The classical shape-from-shading idea, namely, to recover surface information from intensity and integrate into depth, must be reformulated for this setting: instead of a distant, uniform light source, there needs to be modeling that assumes a near point light source offset slightly, depending on model, from the optical axis. It has to include perspective projection and calibrates illumination fall-off over distance. The result is surprisingly powerful for fine relief: especially bony ridges and mucosal folds can be shown even if the scope barely moves. Some state-of-the-art demonstrations take the next step toward something called photometric stereo endoscopy (PSE). By multiplexing the light

source (LEDs) so that the same view is captured under a handful of distinct lighting directions in rapid succession, a calculation can be made per-pixel, and then this information can be incorporated into a topographic map with high density. In benchtop and in-vivo GI endoscopy work, PSE has accurately estimated sub-millimetric features and done so within an endoscope-form-factor system, making it a realistic surgical technology [26,27]. The intuition is appealingly surgical: even when you hold the scope steady, light and shadow play across tissue in ways that somehow model information about local shape. But this also exposes the method's weaknesses, which might be particularly relevant for endoscopic skull base surgery. Specular highlights and reflections violate Lambertian reflectance and, if not accounted for, can confuse the surface rendering. Shadows and albedo (light reflectance) changes (blood, wet tissue and mucosa) confound normal estimation. As a potential solution, the endoscopic literature on this topic therefore has several mitigating ideas. Light position and intensity are calibrated quantitatively for each scope, and specular detection and down-weighting are applied. Where possible, multi-illumination frames are captured in a time-multiplexed burst that is short enough not to induce motion blur but long enough to separate lighting directions. When motion is available, small viewpoint changes are explicitly exploited to integrate normals into a globally consistent depth field and to regularize albedo. More recent works apply depth – from-shading or photometry as ground truths for deep learning-based solutions, which is the topic we will come to next. In this context, they are either used as near-light shading losses to ensure physical or absolute plausibility or as a refinement stage that sharpens a network's coarse depth along high-curvature surfaces [26–30]. The general 'rule of thumb' emerging from these papers is that photometry is most effective for bone and textured dry surfaces, but barely okay on minimally wet mucosa with good Lambertian reflectance assumptions, and becomes fragile and not useful in pools of blood or smoke or wet tissue, where depth-from-motion or deep learning-based approaches should be applied.

2.3. No. 3 – Learning-based and generative depth: from supervised to foundation model applications

A third avenue to approach depth prediction dispenses with attempts at explicit geometric inversion and instead learns to predict depth directly from endoscopic images. Two training regimes dominate: in the supervised case, a depth network (usually a compact encoder–decoder) is trained on frames with labeled depth from phantoms or cadavers (laser scans such as LiDAR (Light Detection and Ranging)) or from registered pre-operative computed tomography (CT) scans. The losses (the way that error is calculated during training) combine per-pixel error with edge-aware smoothness, so that small e.g. bony steps do not so easily wash out. More commonly, especially in surgical applications to date, research groups such as our own have adopted self-supervision: namely, for example, a depth network and a pose estimation method trained so that depth at time t and the estimated relative pose to frame $t \pm 1$ can re-synthesize the neighboring frame. In this case, the photometric reprojection error

becomes the learning signal. To make this work in the busy and moving, unclean operative endoscopic view, losses can mask out instruments and specularities, and add geometry-aware terms: for example, penalties on depth gradients at edges. Several recent endoscopy-specific models in the literature have demonstrated that with these ingredients, single-frame dense depth that is useful for navigation and 3D mapping can be learned without requiring absolute depth information such as from LiDAR, CT or stereo supervision at training and that accuracy now approaches what e.g. applications like video-to-CT registration needs [31–33]. The intuition behind these methods is that these networks internalize the visual regularities of the skull base and find clues, such as how bone brightens under near light, how mucosa reflects when wet, how instruments desaturate, et cetera. These clues are then translated into depth in a single forward prediction. For practical purposes, however, some considerations must be made. Namely, first, monocular depth is scale-ambiguous and relative: without an anchor of absolute depth, in the eyes of the model, 10 and 20 mm differ only by a relative multiplicative constant. Deployed systems therefore introduce metricization/calibration at test time via an external scope tracker, a known instrument in view, or fast registration to pre-op CT – just in any way shape of form taking absolute depth information to calibrate the relative multiplicative constant mentioned above. Perhaps, this is slightly too involved for trans nasal endoscopy, but there have been efforts with robot-assisted studies to go further by including the robot's kinematics, allowing also for some absolute depth information. Second, frame-by-frame predictions can produce flicker of depth between images: to avoid this, the literature consistently uses temporal fusion, which simply means warping the previous depth into the current frame via the estimated pose and then confidence-weighting the fusion of the two. Confidence (in the depth prediction) itself can be predicted (using heteroscedastic regression) or inferred from photometric residuals, and visualized, so stereopsis can be disabled when the estimator is unsure – after all, as with any clinical technical adjunct, it would be unwise to rely too much on its information if the quality of that information is knowingly poor (the same counts for using neuronavigation when the known registration accuracy is poor). These are thus not simply cosmetic details, but instead these are the differences between an in-silico demo and a trustworthy virtual allied technology that is safe in the operating room [31,34–37].

Over roughly the last 2 years, three additional trends have made learning depth materially better for the operating room. First, video-aware self-supervision has matured: longer-range cycle-consistency, latent priors for pose, and better view occlusion handling make learned depth less finicky to scope motion, transient smoke, and instrument sweeps, which are problems in this specific skull base application. Second, there is a growing recognition that surgical images are not always strictly Lambertian. Namely, decoupling shading and albedo even within a self-supervised depth network has been shown to improve robustness in highly specular images such as often is the case in trans nasal endoscopy. Third, the community has adopted general-purpose depth-related features from foundation models for videos to the endoscopic domain. This is

achieved either by fine-tuning with adapter layers to endoscopic footage or by coupling these models to somewhat surgical or benchmarked losses (e.g. EndoDepth benchmark). This last development is where the DINO-of video foundation models become relevant, which is what we want to touch upon in the following subsection [38,39].

DINOv2 is a self-supervised vision transformer (VT) trained at scale on a curated, diverse image corpus (LVD-142 M) to produce all-purpose dense features that transfer across datasets and tasks with little or no fine-tuning – thus, a type of foundation model in its true conceptual sense [40]. For depth, the important property is that DINOv2's intermediate layers encode geometry-aware correspondences remarkably well: if one freezes the backbone and trains a light decoder on top, dense matching and monocular depth improve compared with training from scratch. Some research groups thus keep DINOv2 frozen and train adapter layers with self-supervised endoscopic losses, achieving strong depth performance on SCARED and Hamlyn datasets with few trainable parameters (especially useful for applications with small datasets, such as skull base endoscopy with rare disease entities). Others fine-tune the backbone on endoscopy to align features with near-light appearance. Either way, the intuition is that DINOv2 supplies stable predictions of depth from generalized foundation model features that correlate with depth [40–42].

In a newer version, DINOv3 [43] pushes the same idea to larger scales and with architectural refinements specifically aimed at dense prediction. Whereas DINOv2 already delivered excellent pixel-level features, DINOv3 addresses known failure modes, namely that e.g. feature collapse at fine resolution, that stability is upheld over very large batches, leading to an even more stable 'out of the box' dense features. Early reports emphasize better dense correspondence and stronger downstream monocular depth when the backbone is swapped in place of prior VTs, and the community is beginning to explore DINOv3 backbones for surgical depth and SLAM in the same parameter-efficient regimes used with DINOv2. For the proposed application in skull base endoscopy, DINOv2 made it easier to get good dense features with relatively little amounts of surgical data, whereas DINOv3's improvements make those features sharper and more stable [41–43].

Between these general-purpose methods, MAST3R (Matching and Stereo 3D Reconstruction) 3D priors offer a complementary lever. Unlike generic VT backbones, MAST3R is trained to output point-map correspondences that are explicitly 3D consistent. Plugging such a prior into a surgical SLAM or depth-from-video pipeline has as a consequence that pose estimation remains more stable even in difficult corridors and that dense fusion produces cleaner meshes with fewer outliers and artifacts. Both of these effects can be particularly advantageous for surgical virtual 3D endoscopy, as images need to be truly smooth and realistic in order not to be obtrusive to the surgeon. Conceptually, MAST3R shows what a 'generative prior' for 3D can achieve by regularizing not just for appearance but also takes into account the geometry and leads to more realistic 3D surfaces [24,25].

It is worth ending this section with an honest inventory of trade-offs versus the previously discussed methods. Learned

depth is fast (single-pass inference at video rate on a relatively standard GPU is possible), dense (no more holes where features fail in other non-learning-based methods), and increasingly robust to lighting and deformation changes. However, such approaches will sometimes produce confidently wrong depth outside its training distribution, which is why demonstrating depth prediction confidence is so crucial for realistic application (again, this is akin to neuronavigation, where the surgeon wants to see a measure of registration accuracy before commencing using is). In contrast, motion-based pipelines are more interpretable and less of a 'black box,' can be metrically/absolutely sound once calibrated. In a similar vein, depth from shading can recover relatively high performance and resolution detail without motion, but it demands disciplined optics, relatively stable lighting and struggles in the wet tissues and challenges specific to endoscopy. Also, these techniques are more difficult to execute in real time compared to learning-based approaches, which demonstrably can be rolled out with shorter latencies [19,26,38–40].

3. Expert opinion

In our view, and as supported by the relevant literature, 3D visualization can be a powerful tool especially for junior surgeons who perhaps have limited experience with endoscopy and/or who are transitioning from the stereoscopic view in microscopy to a 2D endoscopic view. Nonetheless, it is obvious that the benefits are small and barely measurable by conventional means (for example, it is unlikely that adding in depth information will lead to a measurably lower complication rate or to more gross total resections – this would also not be expected). The benefits are more 'soft' and can perhaps be seen in surrogate measures such as surgeon confidence. Perhaps, rare instances of miscalculated depth and anatomical misjudgment can be avoided. In our opinion, consequently, adding depth information can contribute, but it is probably overkill to expect every neurosurgical/skull base center to buy an expensive 3D endoscope, especially because 2D endoscopes are already so widely distributed and available. This brings us to one of our specialties within our laboratory: striving to develop software-based solutions to hardware problems. If an open-source software-based method for 3D visualization of endoscopic footage can be provided, running in real-time and on relatively standard GPU hardware, and being visualized to the surgeon in an intuitive way, this will perhaps be enough to cross the barrier of investment/implementation into clinical practice. On the other hand, hardware solutions to hardware problems (such as robots and intraoperative dedicated imaging machines, or in this case: a dedicated 3D endoscope) often work excellently but are limited to those that have the surplus purchasing power to invest in these technologies – which certainly have only marginal day-to-day measurable benefits.

Consequently, we see the development of such a solution as a noble goal and are actively researching optimizations. As mentioned, several problems remain.

Our experience and that of the literature powerfully show that synthesis of virtual 3D endoscopic footage from 2D footage is feasible with a relatively low latency. From

a translational standpoint, latency is at least as critical as geometric accuracy. Human-factors and surgical computer-vision studies typically regard a latency of up to around 320 ms as feasible for surgery [44], and a framerate of ≥ 30 fps (corresponding to about 30–33 ms per frame) as a practical threshold for ‘real-time’ visual feedback in minimally invasive surgery, with lower frame rates (10–15 fps) being usable but noticeably sluggish for continuous hand – eye coordination [45]. Most endoscopic depth and SLAM methods, including monocular systems such as BodySLAM and MAST3R-SLAM, report near-real-time performance while requiring high-end workstation GPUs under benchmark conditions but have not yet been systematically evaluated in the visually adverse setting of skull base surgery [20]. We therefore treat real-time virtual 3D endoscopy as a plausible but still emerging capability, rather than a solved problem.

Still, solving the problem especially in endoscopic skull base surgery, as for example in the resection of pituitary tumors through an endoscopic trans nasal transsphenoidal approach, is an exquisitely difficult machine learning problem. This can be appreciated when viewing some sample endoscopic skull base surgery footage provided in Video 1. Here, a short video from endoscopic pituitary surgery is shown, in conventional 2D. At this stage in the video, the dura of the sella turcica, bulging out due to tumor tissue within, can be seen, as well as the bulges of the carotid arteries bilaterally and the opticocarotid recesses superiorly. At this stage, the

video shows incision of the dura in a cruciate fashion. Especially at these stages of the surgery, added depth perception may be useful especially for junior surgeons, to avoid complications due to anatomical/depth misjudgment, such as incision into the carotid artery. In the video, it can also be appreciated that many factors such as glares, blood pooling, camera blur due to blood, and wet tissue make depth prediction exquisitely difficult in this application. In addition, modern technical adjuncts such as video angiography or other artificially induced colors [46] could require extrapolation from pre-trained models if not included within the training dataset.

To illustrate the results of some single-frame virtual 3D reconstructions from 2D endoscopy, we further provide several examples from different surgical phases of pituitary endoscopic surgery. First, Figure 1 includes an illustration of simple predicted depth color maps using DINOv2 from endoscopic images of the nasal (top) and sphenoidal (bottom) stages of the surgery. Of course, these depth maps are not of much use until converted into textured and realistic 3D views of the actual endoscopic image. Such textured reconstructions are shown in Figures 2 and 3, one from the nasal phase and one from the sellar phase of the surgery. For the sake of demonstrating a 3D result in 2D, we opted here to show a polyaxial view. For the reader, we recommend checking out Video 2 and Video 3, which demonstrate the same but using a dynamic image, so that the plastic and realistic depth effect that is achieved can be optimally demonstrated.

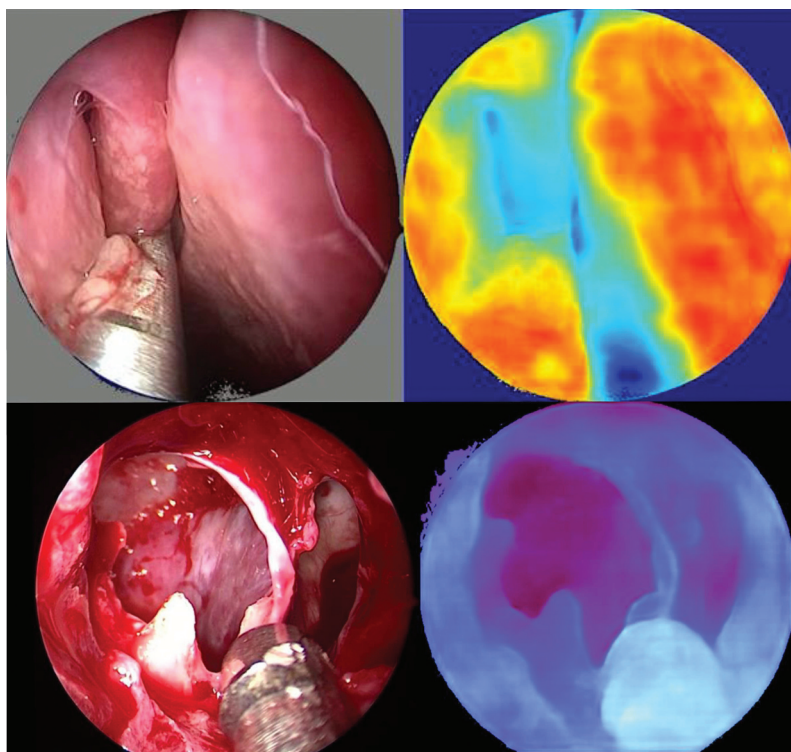


Figure 1. Illustration of predicted depth color maps using DINOv2 from endoscopic images of the nasal (top) and sphenoidal (bottom) stages of the surgery. On the top, dissection of the nasal turbinates is seen. On the bottom, dissection and removal of parts of the sphenoid mucosa and rostrum is seen to gain access to the sphenoid sinus. On the left, conventional endoscopic footage is seen, while on the right, depth maps with different color codes are shown. On the top depth map, red and yellow signify closer structures while darker blue signifies structures further away. On the bottom, darker blue signifies structures further away, while lighter color means closer. It can also be appreciated that many factors such as glares (wet mucosa of the nasal septum as well as bright overexposed bone such as at the anterior sphenoid sinus wall), blood pooling (at the floor of the rostrum), camera blur due to blood (mostly at the top of the screen), and wet tissue make depth prediction exquisitely difficult in this application.

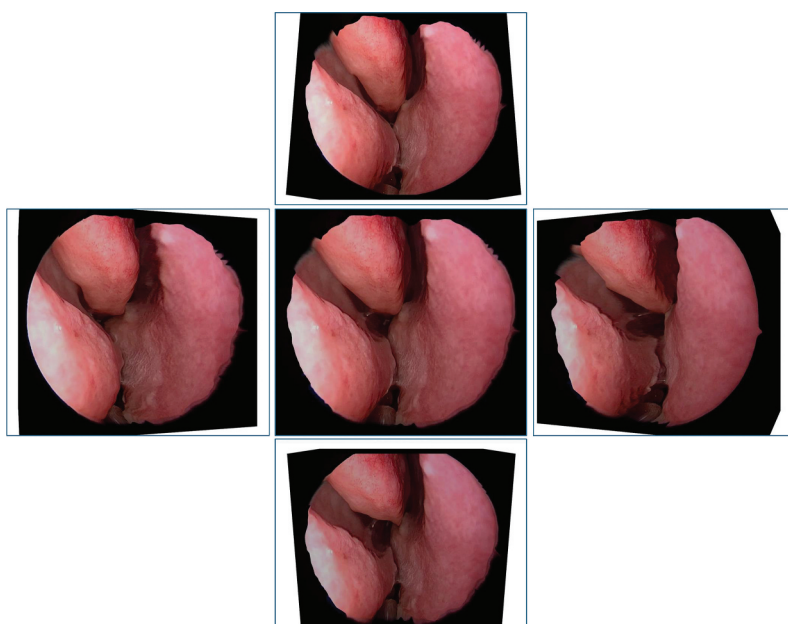


Figure 2. Polyaxial 2D view to illustrate a 3D-reconstructed image frame of the nasal turbinates/nasal cavity predicted from conventional 2D endoscopy. In the middle, the neutral view as seen endoscopically in 2D is shown, whereas tilted views are provided to be able to show the plastic and realistic feeling of depth achieved through depth prediction. This becomes even more apparent in Video 2, which shows the same, but in motion. This figure as well as its associated video show the nasal phase of the surgery, going through the right nostril, where the inferior and middle turbinates of the nasal cavity can be seen on the left in the image, on the right the nasal septum is seen.

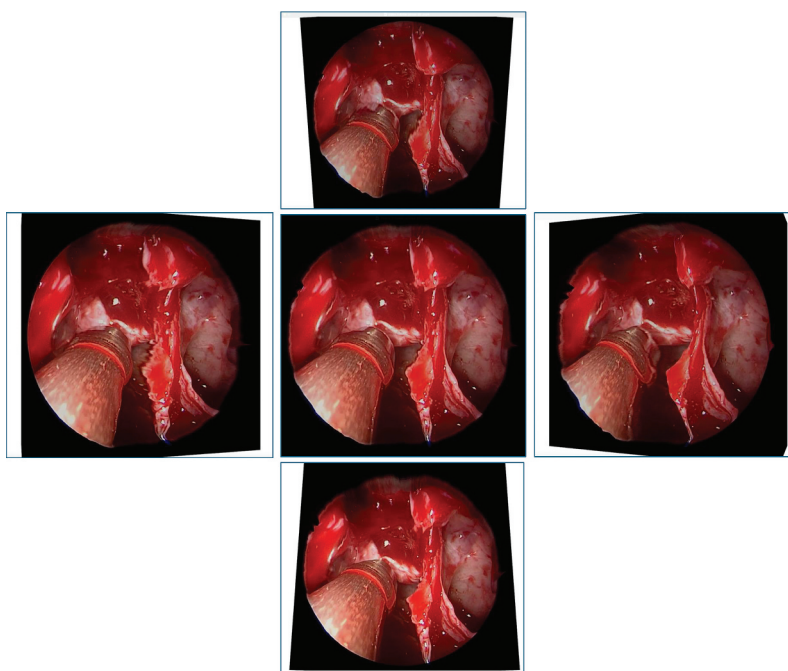


Figure 3. Polyaxial 2D view to illustrate a 3D-reconstructed image frame of the sellar phase of the surgery predicted from conventional 2D endoscopy. In the middle, the neutral view as seen endoscopically in 2D is shown, whereas tilted views are provided to be able to show the plastic and realistic feeling of depth achieved through depth prediction. This becomes even more apparent in Video 3, which shows the same, but in motion. This image shows a suction instrument sucking on tumor tissue within the sella turcica on the left, next to a bony septum of the sphenoid sinus to the right in the image.

Then, even if depth prediction and texturing are optimized, one of the main barriers to implementation will be offering the 3D view to the endoscopic surgeon in a low-latency, naturally integrated and non-obtrusive manner. This would be either as a stereoscopic image from a high-definition 3D

screen with the surgeon wearing 3D glasses or as virtual reality (VR) within a headset. The latter method is less useful in actual surgery, as the surgeon needs to be able to look away from the endoscopic footage itself to e.g. fetch an instrument, look at preoperative imaging, or make

adjustments/technical salvaging on instruments (e.g. wiping blur/condensation on the endoscope with an anti-condensation agent regularly, or adjusting the degree of bending on a flexible skull base suction tip). The VR approach would thus be less suitable. Instead, the already proven method of 3D glasses and 3D screen can be applied. This is already regularly applied in exoscopic surgery, for example. Still, even generating a near-real-time framerate and latency will always require top-grade GPU hardware and, even in their presence, inference may be sub optimally slow with current technologies: as an example, we attach exemplary reconstructed 3D videos of transsphenoidal surgeries generated using a DINOv3 approach, rendered on a RTX 3090 GPU. Video 4 shows a pre-rendered video, while Video 5 shows a real-time rendered video on this hardware. The real-time rendering achieved a frame rate of 2–3 fps and a latency of 300–400 ms. In addition, in complex surgical scenes, blood, smoke, specular highlights, and non-rigid tissue deformation reduce feature stability and temporal consistency, which in turn increases computational demand through repeated feature re-detection, matching, and re-initialization. Prior work in surgical computer vision has shown that robustness to these intraoperative disturbances typically requires additional processing stages (e.g. filtering, temporal modeling, or redundancy), thereby increasing inference time and computational load relative to controlled conditions [47,48]. This illustrates that computing power is still a relative barrier to eventually implementing this technology in actual clinical practice.

We are however confident, considering the amount of technical research and improvement in pipelines that is ongoing especially in the GI endoscopy as well as in the laparoscopy world, that we will see a clinically viable product that runs on relatively standard GPUs with a sufficiently low latency not to be disturbing, within the next decade. Especially in skull base surgery, where the use of endoscopes versus microscopes is on the rise (endoscope assistance in microscopic surgery, extended endonasal approaches, transorbital surgery) as well as in other neurosurgical fields (ventriculotomy, spinal endoscopy) this technique could have an impact – especially for junior surgeons globally, and especially in combination with other software-based assistive technologies such as anatomical structure detection and prediction.

Acknowledgments

We thank the patients whose anonymized data were used for this research.

Funding

V Staartjes is supported by the Prof. Dr Max Cloetta Foundation.

Author contributions

CRedit: **Victor E. Staartjes**: Conceptualization, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing; **Olivier Zanier**: Conceptualization, Data curation, Formal analysis, Project administration, Writing – original draft, Writing – review & editing; **Kenneth Arockia**: Data curation, Formal analysis, Project administration, Writing – original draft, Writing – review & editing; **Stefanos Voglis**: Data

curation, Project administration, Writing – original draft, Writing – review & editing; **Giovanni Muscas**: Data curation, Project administration, Writing – original draft, Writing – review & editing; **Andrea Boschi**: Data curation, Project administration, Writing – original draft, Writing – review & editing; **Alessandro Della Puppa**: Data curation, Project administration, Writing – original draft, Writing – review & editing; **Adrian Elmi-Terander**: Data curation, Project administration, Writing – original draft, Writing – review & editing; **Luca Regli**: Data curation, Project administration, Writing – original draft, Writing – review & editing; **Carlo Serra**: Conceptualization, Data curation, Methodology, Project administration, Writing – original draft, Writing – review & editing.

Declaration of interest

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

References

Papers of special note have been highlighted as either of interest (*) or of considerable interest () to readers**

1. Falkenberg KJ, Matta II. The use of the operating microscope in otolaryngology. *Guthrie J.* 1985;55(2):91–98. doi: [10.3138/guthrie.55.2.091](#)
2. Uluç K, Kujoth GC, Başkaya MK. Operating microscopes: past, present, and future. *Neurosurg Focus.* 2009;27(3):E4. doi: [10.3171/2009.6.FOCUS09120](#)
3. Yasargil MG. *Microneurosurgery, volume I: microsurgical anatomy of the basal cisterns and vessels of the brain, diagnostic studies, general operative techniques and pathological considerations of the intracranial aneurysms.* Thieme; 1984.
4. Reuter M. [Philipp Bozzini (1773–1809): the endoscopic idealist]. *Urol Ausg A.* 2006;45(9):1084–1088, 1090–1091.
5. Stachura K, Libionka W. An outline of the history of neuroendoscopy. *Przegl Lek.* 2007;64(2):118–120.
6. Cappabianca P, Cavallo LM, de Divitiis E. Endoscopic endonasal transsphenoidal surgery. *Neurosurgery.* 2004;55(4):933–940; discussion 940–941. doi: [10.1227/01.NEU.0000137330.02549.0D](#)
7. Xin G, Liu Y, Xiong Y, et al. The use of three-dimensional endoscope in transnasal skull base surgery: a single-center experience from China. *Front Surg [Internet].* 2022 [cited 2025 Oct 13];9. doi: [10.3389/fsurg.2022.996290](#)
8. Inoue D, Yoshimoto K, Uemura M, et al. Three-dimensional high-definition neuroendoscopic surgery: a controlled comparative laboratory study with two-dimensional endoscopy and clinical application. *J Neurol Surg Part Cent Eur Neurosurg.* 2013;74(6):357–365. doi: [10.1055/s-0033-1345100](#)
9. Nebor I, Anderson Z, Mejia-Munne JC, et al. 2D versus 3D endoscopy: head-to-head comparison in a simulated model of endoscopic endonasal dural suturing. *J Neurol Surg Part B Skull Base.* 2022;83(4):423–429. doi: [10.1055/s-0041-1736635](#)
10. ten Dam E, Helder HM, van der Laan BFAM, et al. The effect of three-dimensional visualisation on performance in endoscopic sinus surgery: a clinical training study using surgical navigation for movement analysis in a randomised crossover design. *Clin Otolaryngol.* 2020;45(2):211–220. doi: [10.1111/coa.13494](#)
- **Clinical study critically discussing the actual relevance of 3D visualization in endoscopic cranial surgery.**
11. Lin X, Ono K, Ohata K, et al. Improving operator depth vision in gastrointestinal endoscopy: a pilot study of a new three-

- dimensional imaging system. *DEN Open*. 2025;6(1):e70174. doi: 10.1002/deo2.70174
12. Altieri R, Tardivo V, Pacca P, et al. 3D HD endoscopy in skull base surgery: from darkness to light. *Surg Technol Int*. 2016;29:359–365.
 13. Tabaei A, Anand VK, Fraser JF, et al. Three-dimensional endoscopic pituitary surgery. *Neurosurgery*. 2009;64(5 Suppl 2):288–293; discussion 294–295.
 - **One of the first neurosurgical contributions to 3D endoscopic surgery by early adopters.**
 14. Uvelius E, Siesjö P. 3-D endoscopy in surgery of pituitary adenomas, prospective evaluation of patient gain using basic outcome parameters. *J Clin Neurosci Off J Neurosurg Soc Australas*. 2020;76:166–170. doi: 10.1016/j.jocn.2020.03.026
 15. Kiyofuji S, Shin M, Kondo K, et al. Endoscopic transnasal resection of trigeminal schwannoma. *J Neurol Surg Part B Skull Base*. 2022;83 (Suppl 3):e637–e638.
 16. Brown EC, Lucke-Wold B, Cetas JS, et al. Surgical parameters for minimally invasive trans-eustachian tube CSF leak repair: a cadaveric study and literature review. *World Neurosurg*. 2019;122:e121–e129.
 17. Lucke-Wold B, Brown EC, Cetas JS, et al. Minimally invasive endoscopic repair of refractory lateral skull base cerebrospinal fluid rhinorrhea: case report and review of the literature. *Neurosurg Focus*. 2018;44(3):E8. doi: 10.3171/2017.12.FOCUS17664
 18. Ozyoruk KB, Gokceler GI, Coskun G, et al. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos: endo-SfMLearner [Internet]. arXiv.org. 2020 [cited 2025 Oct 12]. Available from: <https://arxiv.org/abs/2006.16670v3>
 19. Azagra P, Sostres C, Ferrandez Á, et al. EndoMapper dataset of complete calibrated endoscopy procedures [Internet]. arXiv.org. 2022 [cited 2025 Oct 12]. Available from: <https://arxiv.org/abs/2204.14240v2>
 20. Manni G, Lauretti C, Prata F, et al. BodySLAM: a generalized monocular visual SLAM framework for surgical applications [Internet]. arXiv.org. 2024 [cited 2025 Oct 12]. Available from: <https://arxiv.org/abs/2408.03078v2>
 21. Elvira R, Tardós JD, Montiel JMM. CudaSIFT-SLAM: multiple-map visual SLAM for full procedure mapping in real human endoscopy [Internet]. arXiv.org. 2024 [cited 2025 Oct 12]. Available from: <https://arxiv.org/abs/2405.16932v1>
 22. Teufel T, Shu H, Soberanis-Mukul RD, et al. OneSLAM to map them all: a generalized approach to SLAM for monocular endoscopic imaging based on tracking any point. *Int J Comput Assist Radiol Surg*. 2024;19(7):1259–1266. doi: 10.1007/s11548-024-03171-6
 23. Herrera-Granda EP, Torres-Cantero JC, Peluffo-Ordóñez DH. Monocular visual SLAM, visual odometry, and structure from motion methods applied to 3D reconstruction: a comprehensive survey. *Heliyon*. 2024;10(18):e37356. doi: 10.1016/j.heliyon.2024.e37356
 24. Leroy V, Cabon Y, Revaud J. Grounding image matching in 3D with MAST3R [Internet]. arXiv.org. 2024 [cited 2025 Oct 12]. Available from: <https://arxiv.org/abs/2406.09756v1>
 25. Murai R, Dexheimer E, Davison AJ. MAST3R-SLAM: real-time dense SLAM with 3D reconstruction priors [Internet]. arXiv.org. 2024 [cited 2025 Oct 12]. Available from: <https://arxiv.org/abs/2412.12392v2>
 - **This study demonstrates that integrating learned 3D reconstruction priors (MAST3R) into SLAM enables real-time, dense, and globally consistent 3D reconstruction, directly supporting low-latency, robust scene understanding.**
 26. Parot V, Lim D, González G, et al. Photometric stereo endoscopy. *J Biomed Opt*. 2013;18(7):076017. doi: 10.1117/1.JBO.18.7.076017
 27. Kim J, Al Faruque H, Kim S, et al. Multimodal endoscopic system based on multispectral and photometric stereo imaging and analysis. *Biomed Opt Express*. 2019;10(5):2289–2302. doi: 10.1364/BOE.10.002289
 28. Lei L, Li J, Liu M, et al. Shape from shading and optical flow used for 3-dimensional reconstruction of endoscope image. *Acta Otolaryngol (Stockh)*. 2016;136(11):1190–1192. doi: 10.1080/00016489.2016.1193895
 29. Wu C, Narasimhan SG, Jaramaz B. Shape-from-shading under near point lighting and partial views for orthopedic endoscopy. 2007.
 30. Bobrow TL, Golhar M, Arayakarnkul S, et al. Multi-contrast laser endoscopy for in vivo gastrointestinal imaging. *ArXiv*. 2025. p. arXiv:2505.10492v2.
 31. Liu X, Sinha A, Ishii M, et al. Dense depth estimation in monocular endoscopy with self-supervised learning methods. *IEEE Trans Med Imag*. 2020;39(5):1438–1447. doi: 10.1109/TMI.2019.2950936
 32. Yang Y, Shao S, Yang T, et al. A geometry-aware deep network for depth estimation in monocular endoscopy [Internet]. arXiv.org. 2023 [cited 2025 Oct 12]. Available from: <https://arxiv.org/abs/2304.10241v1>
 33. Xu Z, Li B, Hu Y, et al. Self-supervised monocular depth and pose estimation for endoscopy with generative latent priors [Internet]. arXiv.org. 2024 [cited 2025 Oct 12]. Available from: <https://arxiv.org/abs/2411.17790v2>
 34. Li W, Hayashi Y, Oda M, et al. Enhanced self-supervised monocular depth estimation with self-attention and joint depth-pose loss for laparoscopic images. *Int J Comput Assist Radiol Surg*. 2025;20 (4):775–785. doi: 10.1007/s11548-025-03332-1
 35. Li W, Hayashi Y, Oda M, et al. Multi-view guidance for self-supervised monocular depth estimation on laparoscopic images via spatio-temporal correspondence. In: *Med Image Comput Comput Assist Interv – MICCAI 2023 26th Int Conf Vanc BC Can Oct 8–12 2023 Proc Part IX* [Internet]; Berlin, Heidelberg. Springer-Verlag; 2023. p. 429–439. [cited 2025 Oct 12]. doi: 10.1007/978-3-031-43996-4_41
 36. Huang B, Zheng J-Q, Nguyen A, et al. Self-supervised depth estimation in laparoscopic image using 3D geometric consistency. In: Wang L, Dou Q, Fletcher PT, et al., editors. *Med Image Comput Comput Assist Interv – MICCAI 2022* [Internet]. Cham: Springer Nature Switzerland; 2022 [cited 2025 Oct 13]. p. 13–22. Available from: https://link.springer.com/10.1007/978-3-031-16449-1_2
 37. Hwang S-J, Park S-J, Kim G-M, et al. Unsupervised monocular depth estimation for colonoscope system using feedback network. *Sensors*. 2021;21(8):2691. doi: 10.3390/s21082691
 38. Wang X, Yang B, Wei M, et al. Deep learning for endoscopic depth estimation: a review. *Displays*. 2025;90:103086. doi: 10.1016/j.displa.2025.103086
 39. Reyes-Amezcuca I, Espinosa R, Daul C, et al. EndoDepth: a benchmark for assessing robustness in endoscopic depth prediction [Internet]. arXiv. 2024 [cited 2025 Oct 12]. Available from: <http://arxiv.org/abs/2409.19930>
 40. Oquab M, Darcet T, Moutakanni T, et al. DINOv2: learning robust visual features without supervision [Internet]. arXiv; 2024 [cited 2025 Oct 12]. Available from: <http://arxiv.org/abs/2304.07193>
 41. Cui B, Islam M, Bai L, et al. Surgical-DINO: adapter learning of foundation models for depth estimation in endoscopic surgery. *Int J Comput Assist Radiol Surg*. 2024;19(6):1013–1020. doi: 10.1007/s11548-024-03083-5
 - **This study shows that adapter-based fine-tuning of a foundation model (DINO) enables accurate depth estimation in endoscopic surgery with limited task-specific data, supporting efficient domain adaptation of large vision models for intraoperative 3D perception.**
 42. Cui B, Islam M, Bai L, et al. EndoDAC: efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera. In: Linguraru MG, Dou Q, Feragen A, et al., editors. *Med Image Comput Comput Assist Interv – MICCAI 2024* [Internet]. Cham: Springer Nature Switzerland; 2024 [cited 2025 Oct 12]. p. 208–218. Available from: https://link.springer.com/10.1007/978-3-031-72089-5_20
 43. Siméoni O, Vo HV, Seitzer M, et al. DINOv3 [Internet]. arXiv.org. 2025 [cited 2025 Oct 12]. Available from: <https://arxiv.org/abs/2508.10104v1>
 44. Wang Y, Ai Q, Shi T, et al. Influence of network latency and bandwidth on robot-assisted laparoscopic telesurgery: a pre-clinical experiment. *Chin Med J (Engl)*. 2025;138(3):325–331. doi: 10.1097/CM9.0000000000003257

45. Kamtam DN, Shrager JB, Malla SD, et al. Deep learning approaches to surgical video segmentation and object detection: a scoping review [Internet]. arXiv. 2025 [cited 2025 Dec 7]. Available from: <http://arxiv.org/abs/2502.16459>
46. Schmidt A, Mohareri O, DiMaio S, et al. Tracking and mapping in medical computer vision: a review. Med Image Anal. 2024;94:103131. doi: 10.1016/j.media.2024.103131
47. Richter F, Shen S, Liu F, et al. Autonomous robotic suction to clear the surgical field for hemostasis using image-based blood flow detection. IEEE Robot Autom Lett. 2021;6(2):1383–1390. doi: 10.1109/LRA.2021.3056057
48. Liu Z, Gao W, Zhu J, et al. Surface deformation tracking in monocular laparoscopic video. Med Image Anal. 2023;86:102775. doi: 10.1016/j.media.2023.102775