INVITED PAPER

# CAI4CAI: The Rise of Contextual Artificial Intelligence in Computer-Assisted Interventions

*This article overviews ideas as to how to incorporate the range of prior knowledge and instantaneous sensory information from experts, sensors and actuators for use in computer-assisted interventions, as well as learning how to develop a representation of the surgery or intervention among a mixed human-AI team of actors. In addition, the design of interventional systems and associated cognitive shared control schemes for online uncertainty awareness when making decisions in the OR or the IR suite is discussed, and it is noted how this is critical for producing precise and reliable interventions.*

By Tom Vercauteren, Mathias Unberath, Nicolas Padoy, and Nassir Navab

**ABSTRACT** | Data-driven computational approaches have evolved to enable extraction of information from medical images with reliability, accuracy, and speed, which is already transforming their interpretation and exploitation in clinical practice. While similar benefits are longed for in the field of interventional imaging, this ambition is challenged by a much higher heterogeneity. Clinical workflows within interventional suites and operating theaters are extremely complex and typically rely on poorly integrated intraoperative devices, sensors, and support infrastructures. Taking stock of some of the most exciting developments in machine learning and artificial intelligence for computer-assisted interventions, we highlight the crucial need to take the context and human factors into account in order to address these challenges. Contextual artificial intelligence for computer-assisted intervention (CAI4CAI) arises as an emerging opportunity feeding into the broader field of surgical data science. Central challenges being addressed in CAI4CAI include how to integrate the ensemble of prior knowledge and instantaneous sensory information from experts, sensors, and actuators; how to create and communicate a faithful and actionable shared representation of the surgery among a mixed human–AI actor team; and how to design interventional systems and associated cognitive shared control schemes for online uncertainty-aware collaborative decision-making ultimately producing more precise and reliable interventions.

Manuscript received August 1, 2019; revised September 12, 2019; accepted October 4, 2019. Date of publication October 23, 2019; date of current version December 26, 2019. The work of T. Vercauteren was supported in part by the Medtronic/Royal Academy of Engineering Research Chair under Grant RCSRF1819\7\34, in part by the Wellcome Trust under Grant 203148/Z/16/Z and Grant WT101957, and in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant NS/A000049/1 and Grant NS/A000027/1. The work of M. Unberath was supported in part by NIH/NIBIB under Grant R01 EB0223939, in part by Johns Hopkins University, and in part by the Fellowship of the Malone Center for Engineering in Healthcare, Johns Hopkins University. The work of N. Padoy was supported in part by the French State Funds managed by the Agence Nationale de la Recherche (ANR) through the Investissements d'Avenir Program under Grant ANR-16-CE33-0009 (DeepSurg), Grant ANR-18-CE45-0011-03 (OptimiX), and Grant ANR-11-LABX-0004 (Labex CAMI) and in part by BPI France through Project CONDOR. *(Corresponding author: Tom Vercauteren.)*

**T. Vercauteren** is with the School of Biomedical Engineering & Imaging Sciences, King's College London, London WC2R 2LS, U.K. (e-mail: tom.vercauteren@kcl.ac.uk).

**M. Unberath** is with the Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218 USA.

**N. Padoy** is with the ICube institute, CNRS, IHU Strasbourg, University of Strasbourg, 67081 Strasbourg, France.

**N. Navab** is with the Fakultät für Informatik, Technische Universität München, 80333 Munich, Germany.

Digital Object Identifier 10.1109/JPROC.2019.2946993

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

**KEYWORDS** | Artificial intelligence; computer-assisted interventions; context-aware user interface; data fusion; interventional workflow; intraoperative imaging; machine and deep learning; surgical data science; surgical planning; surgical scene understanding.

## I. INTRODUCTION

Contemporary progresses in machine learning and artificial intelligence have permitted the development of tools that can assist clinicians in exploiting and quantifying clinical data including images, textual reports, and genetic information. State-of-the-art algorithms are becoming mature enough to provide automated analysis when applied to well-controlled clinical studies and trials [1], [2], but adapting these tools for patient-specific management remains an active research area, with the bulk of the research community having focused on fully automated machine learning tools. These considerations become especially critical in the highly heterogeneous context of surgery and interventional procedures that require patient- and team-specific decision support tools being able to draw information from nonstandardized interventional devices integrated into diverse interventional suites. Compared to computational tasks in radiology, the domain of computer-assisted intervention further creates unique methodological challenges, such as imposing stringent time constraints in the interventional suite, requiring knowledge of procedural data, and needing methods that deal with dynamic environments.

In this article, keeping a focus on imaging data, we review existing work and share insights on future developments of machine learning strategies that decipher, support, augment, and integrate into various surgical and interventional workflows while providing the flexibility required by clinical management. Flexibility is, for example, mandated to be able to deal with missing input sources, react to real-time user feedback, adapt to the patient risk aversion and preferences, handle uncertain or contradictory information, learn from potentially small and heterogeneous data, and so on. All of them are common in computer-assisted interventions. Imaging sources of particular interest for surgery and intervention include a wide range of well-known interventional modalities, such as surgical microscopy, video endoscopy, X-ray fluoroscopy, and ultrasound, more emerging biophotonics imaging modalities, such as hyperspectral imaging, endomicroscopy, and photoacoustic imaging, and also span classical radiology modalities, such as MRI and CT, that remain the main sources of imaging data for preoperative intervention planning and postoperative assessment. We argue that the stringent need to consider the context when analyzing surgical and interventional data coupled with the heterogeneity of information sources and domain knowledge in computer-assisted intervention applications calls for the development of novel domain-specific contextual artificial intelligence solutions, a domain that we coin as the contextual artificial intelligence for computer-assisted intervention (CAI4CAI). Feeding into the broader field of surgical data science [3]–[5], CAI4CAI will focus on the underpinning machine learning methodology exploiting contextual information and human interaction to enable the required responsiveness to deliver the clinical impact on surgery and interventional sciences.

To support our claim, we highlight some of the transformative machine learning research results and methodologies currently being developed across the spectrum of tasks in computer-assisted interventions. The impact of machine learning in intervention planning is discussed in Section II, intraoperative data fusion in Section III, intelligent intraoperative imaging in Section IV, surgical and endoscopic vision in Section V, and clinical workflow monitoring and support in Section VI. In these sections, we will highlight how flexible deep learning-based tools are becoming critical for the design of effective and efficient intervention planning solutions. During surgery, navigation solutions are often used to map preoperative information in the context of the intervention. However, navigation does not account for intraoperative changes. Learning how to coregister images is now leading to intraoperative registration solutions that are able to cope with the highly challenging task of aligning preoperative to intraoperative images coming from different imaging modalities. Concurrently, AI methodology is advancing to go beyond traditional navigation-based data fusion and image overlay to exploit information coming from complex or synergistic data sources. This is giving rise to what we refer to as intelligent intraoperative imaging. Data-driven modeling strategies coming from the computer vision community are acting as instrumental starting points to achieve semantic information extraction from interventional data sources, including endoscopic videos, with applications ranging from automated polyp detection to surgical activity recognition. To deliver improved clinical outcomes through AI, all these building blocks are increasingly being integrated at the level of the complete surgical workflow with applications spanning the full breadth of surgical data science. In this area, starting from the data-driven mapping of clinical workflow and skills assessment, AI is now helping make contextual decision support tools and conditionally autonomous intervention a reality. Finally, closing thoughts are provided and further budding applications of CAI4CAI are discussed in Section VII.

## II. INTERVENTION PLANNING

### A. Clinical Adoption of Intervention Planning Tools

Once a decision is made for a patient to undergo an interventional procedure, for any nontrivial operation, patient-specific planning of the intervention is required. The steps involved usually necessitate the acquisition of reference preoperative imaging data, semantic segmentation of anatomical structures in these images,

determination of the surgical approach, and elaboration of an intraoperative plan leading to optimal outcomes for the patient. Such a plan might encompass establishing a surgical path and target, designing, or selecting a patient-specific implant or assistive adjunct tool such as a drill or saw guide [6]. In the majority of cases, such intervention planning is performed by a team of healthcare professionals, each with their own expertise, known as the multidisciplinary team (MDT). Relatively, little computer assistance is currently available for interventional planning in the clinic. Notable exceptions can be found in the field of neurosurgery, oral and maxillofacial surgery, and orthopedic surgery. What these specialties share is a relatively static surgical scene due to the proximity of rigid bone structures. Computed tomography (CT) provides a rich source of 3-D imaging information in this context. Indeed, due to the quantitative nature of CT images and the good contrast of bone, automated segmentation of bone has proved to be clinically reliable. Because of the seminal work of the Retrospective Registration Evaluation Project (RREP) [7], it is also clear that preoperative rigid registration of different imaging modalities, such as MR and CT, provides a robust means of fusing soft tissue contrast information with accurate bone delineation for neurosurgical planning. Such technical advances have supported the adoption of stereotactic surgery as a means of accurately targeting and guiding instrument toward deep-seated brain structures for procedures, such as brain biopsies for tumor grading and electrode implantation for the treatment of movement disorder or the localization of epileptic seizure onset zones. While computer-assisted surgical planning and subsequent surgical navigation become standard of care in neurosurgery and a few other disciplines, even in these fields, there is major scope to make the workflow more efficient through the development of further machine learning-enabled computer assistance.

### B. Machine Learning in Interventional Planning

Commercial surgical planning products are still limited in the automation they support, with many of the most advanced ones essentially relying on classical image analysis methods, such as atlas-based segmentation [9], to delineate soft-tissue structures of interests for a patient showing no gross pathological brain changes. Clinicians are often left with manual or generic interactive methods to delineate other structures of interest and define their surgical plan. When interventional planning only relies on the clinician getting a volumetric representation of the patient anatomy from preoperative data, advanced visualization techniques, such as cinematic rendering [10], can be considered as alternatives to explicit segmentation of structures. These may produce results that are less sensitive to noise and data variability but do not enable more quantitative planning. Developments of deep machine learning segmentation algorithms dedicated to medical imaging [11], [12] are rapidly changing to a level of accuracy at which automated segmentation of structures of interest can be done in a population of patients even in the presence of gross pathological changes [13]. However, many challenges remain for these tools to become of practical use for intervention planning purposes. Poor generalization, when faced with slight domain changes, is a recognized problem in the entire medical imaging community including on the diagnostic side. Expanding the size of the data sets on which deep learning algorithms are trained would certainly mitigate generalization issues by providing a much larger variety of training cases. Collaborative efforts within the community are notably focusing on providing open-access large annotated data sets for machine learning training purposes in some specific use cases [1]. However, collecting task-specific large annotated databases for medical imaging purposes faces its own challenges, given the time and expertise required to provide detailed annotations as well as the legal, privacy, and storage questions pertaining to sharing large patient data sets across multiple sites. Federated learning for multi-institutional collaboration in medical imaging [14], [15] provides a potential technical solution to this problem. Implementing such solutions at scale will require concerted efforts reaching far beyond the methodological research community. Furthermore, changes such as device upgrades or challenges posed by new clinical indications will not be captured by increasing the pool of retrospective training data. Active research to address such inevitable but unpredictable domain gaps is rooted in domain adaptation techniques [16]. These advances are necessary for automated machine learning tools to make an impact on the clinical setting. Prospective randomized clinical trials (RCTs) are widely seen as the only source of trustworthy clinical evidence, yet studies implementing RCTs with systems relying on deep learning tools for medical imaging currently remain noteworthy exceptions [17].

### C. Importance of Flexible Contextual Machine Learning

What distinguishes segmentation in surgical planning from segmentation in diagnostic imaging is, nonetheless, that the objective is not necessarily always that of reaching the best performance in getting the structures delineated with subvoxel accuracy. Surgical planning needs to respect the patient-specific needs and preferences of the surgeon. This requires putting the clinical team at the center and promoting flexible tools that integrate into the surgical workflow. Interactive deep learning methodologies are emerging to combine rich prior knowledge embedded in retrospective data from previous patients with as-sparse-as-possible annotations provided by clinicians [8], [18]. As illustrated in Fig. 1, deep interactive segmentation allows the clinical expert to refine the results from an initial automated step and, most importantly, to adapt the inferred results on the fly based on contextual information. Furthermore, given the heterogeneity and evolving nature of the surgical practice, additional flexibility is required to handle potentially missing input modalities. Recent work

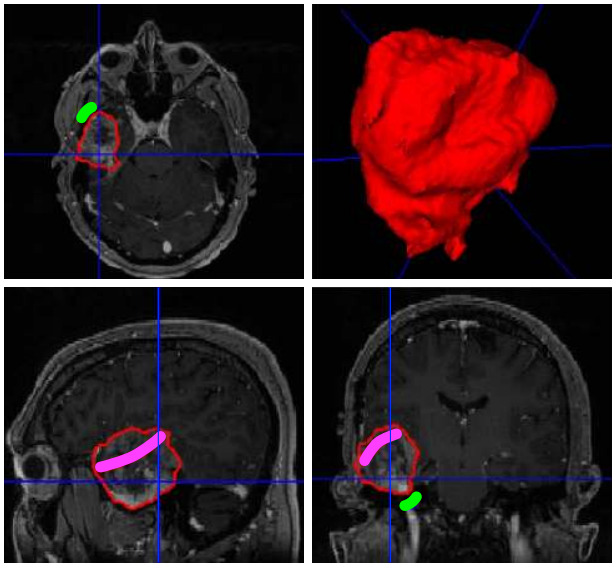

**Fig. 1.** *Interactive algorithms are required to deliver context-aware artificial intelligence. In this example, using the algorithm presented in [8], brain tumor segmentation is initially performed automatically using a pretrained algorithm. As a part of the surgical planning, the user may want to refine the segmentation by providing scribbles to denote areas that should be excluded (green region) or included (pink region) irrespective of the initial segmentation. The algorithm then adapts its output to respect the user input.*

in deep machine learning is focusing on dealing with such dynamic heteromodal context while exploiting heterogeneous sources of data for the training process [19], [20]. Bringing flexible machine learning tools to maturity will certainly play an important role in supporting the clinical adaption of AI in surgery.

As highlighted earlier, segmentation of structures from preoperative images is often the foundation of computer-assisted surgical planning, and this currently remains the state of the art in many commercial solutions. Such static segmentation, when combined with intraoperative registration already, provides useful surgical navigation information for relatively static surgical scenes as is the case in neurosurgery. Nevertheless, computer assistance for intervention planning has the potential to provide impact much beyond the ability to automate the creation of 3-D anatomical models and overlay of functional data. Patient-specific simulation of given surgical plans has, for example, been introduced in orthopedic surgery with a long history in acetabular fracture surgery [21]. State-of-the-art orthopedic surgery planning systems allow to design patient-specific implants and patient-specific surgical guides by enabling the simulation of the effect of different implants and implantation strategy on key outcome-related parameters, such as the range of motion of articulation or the limb length [22]. However, these tools often ignore the effect of soft tissue in the simulation process and still require very labor-intensive work for the surgical team to design patient-specific plans. Expert systems capable of automatically optimizing the surgical plan for a given orthopedic surgery are now being developed [23] and promise to make surgical planning more efficient [24]. In the context of deep brain insertion of instruments, machine learning approaches capable of automatically planning trajectories of multiple instruments, to maximize the efficacy of the surgery while minimizing intraoperative risks and avoiding collisions between instruments, have demonstrated a significant reduction in planning time for the implantation of stereoelectroencephalography electrodes for epilepsy treatment [25] and for laser interstitial thermal therapy [26]. Contextual and flexible machine learning for surgical planning promises to push the boundaries of interventional planning by exploiting data-driven approaches and real-time user feedback to efficiently plan for complex situations. An instrument bending model was, for example, trained in [27] to predict the deviation between an original surgical plan assuming rigid electrodes and the actual electrode paths as measured on a postoperative CT. Provided reliable uncertainty estimates on the prediction can be achieved, embedding such deflection models in the trajectory planning is expected to improve the safety and accuracy of stereoelectroencephalography electrode implantation planning.

Effectively, planning is moving away from the extraction of information captured in existing data and representative of a given (preoperative) time point. Context-aware learning methods are now being developed to also predict therapy-related changes and better inform interventional planning. By exploiting computationally complex noninvasive cardiac electrophysiology modeling coupled with transfer learning approaches, Giffard-Roisin *et al.* [28] notably achieved online personalized predictions of electrophysiology cardiac resynchronization therapy responses, thereby paving the way for better patient selection and patient-specific therapy optimization. In nonquasi-static environments, surgical planning is currently further limited by our capabilities to predict intraoperative anatomical changes. In abdominal surgery, for example, segmentation of structures from preoperative images may inform the clinician about the relative spatial organization of lesions and vascular structures. However, at the onset of a minimally invasive procedure, gas insufflation is typically performed to create the surgical workspace. This has a serious impact on the geometry of the anatomy and challenges any attempt of intraoperative use of a 3-D model of the anatomy generated from preinsufflation images. Current approaches typically rely on focusing on smaller regions where rigidity assumptions between preoperative and intraoperative data may still hold [29], thereby limiting the scope of surgical planning. Data-driven prediction of anatomical changes relating to gas insufflation in laparoscopic surgery was proposed in [30]. Still, in the context of liver surgery, a system able to take into account nonimaging patient data and factual knowledge gathered from quotable sources, such as clinical guidelines, was proposed to support individualized treatment planning [31]. While relying on

handcrafted features and exploiting models with limited expressiveness, this article paved the way for more holistic interventional planning. It is expected that the context-aware interventional planning will be informed by refined prediction models to suggest therapeutic plans cognizant of clinical experience as well as potential intraoperative changes and associated risks but also flexible enough to take into account any further input from the interventional team interacting with a responsive planning system.

## III. INTRA-OPERATIVE DATA FUSION

### A. Navigation and Image Registration Challenges

No matter how refined and capable interventional planning becomes, its full value for procedural guidance and intraoperative decision-making support remains contingent on appropriate geometric alignment with intraoperatively acquired data. This alignment is achieved using registration methods that either rely on dedicated external hardware, such as optical or electromagnetic tracking systems [32], or operate directly on intraoperative images [33].

Image-based registration in the interventional context has received substantial academic attention [34], [35]. This is because external navigation, while improving surgical accuracy, is associated with increased procedural time and complex and manual intraoperative calibration procedures that may lead to a high level of surgeon frustration [36]. It is widely believed that image-based registration will better integrate with procedural workflow, mitigating many negative aspects of external tracking approaches while providing similar accuracy. Furthermore, since no additional hardware is required, there is great potential for widespread adoption and deployment of these purely software-driven methods. This suggests that navigated surgery may also become available in remote and rural hospitals that could not afford dedicated equipment otherwise.

Despite the clear opportunity, image-based registration is not yet widely used in interventional clinical practice. This is because, depending on the clinical context, several challenges of image-based registration have not yet been solved reliably. During surgery, the anatomy undergoes highly complex deformations, including the loss of mass or topological changes during resections. Accurately recovering bio-mechanically plausible transformations that represent an anatomical change from preoperative to intraoperative state that is measured with different imaging modalities is the subject of the ongoing research. Here, we will focus on two of the associated challenges: 1) modeling image similarity between the images of the same anatomy but acquired with different modalities and 2) estimating initial transformation parameters that are good enough for registration algorithms to succeed.

On a high level, image registration seeks to find a transformation that, when applied to the moving image, aligns it with the target image such that the locations in both images are in correspondence. Quantifying *correspondence* is achieved using image similarity metrics that, usually, operate on the image intensity values. A straightforward comparison of intensity values, e.g., using a simple sum of squared differences, is generally unrewarding since the underlying assumption on image formation is prohibitively strong, even when moving and target images are acquired with the same imaging modality. For interventional image fusion, the problem is more challenging since images of different modalities must be aligned. In this case, the additive Gaussian noise assumption underpinning the sum of squared differences is certainly violated. Even worse, due to the different physical processes that govern image formation, there is no guarantee that the same anatomical structures are visible in both images, thereby challenging the adequacy of co-occurrence-based similarity metrics, including correlation and mutual information. Nonetheless, despite these limitations, model-based image similarity criteria currently remain the state-of-the-art performers in many interventional image-registration tasks, including ultrasound to MRI registration for neurosurgical guidance [37], [38].

### B. Contextual Learning for Image Registration

Using deep learning to go past some of the limitations of classical image registration is an active area of research. However, due to the fundamental challenge of gathering ground-truth data for image registration, many of the most successful learning-based registration methods for diagnostic images exploit unsupervised learning and optimize a classical image similarity metric-based loss [39], [40]. This approach remains unsuitable for most interventional purposes where more flexible solutions are required. A prominent example highlighting the need to take the interventional context into account is a transrectal ultrasound (TRUS)-guided prostate biopsy. Conventionally, the biopsy target is segmented on preoperative 3-D MR images, and this must then be registered to intraoperative 3-D TRUS volumes. Since MR and TRUS images exhibit a substantially different image appearance, contrast, and artifact level, this suggests that no good mathematical model exists to describe image similarity between these two modalities. Data-driven approaches that do not explicitly model intensity correlations to test for image correspondence but optimize a surrogate measure thereof now achieve state-of-the-art performance. One candidate surrogate measure can be defined by enforcing segmentations of the same structures to exhibit maximal overlap after registration [41]. Remarkably, learning to optimize for such losses does not require access to ground truth for the spatial transformation and leverages application-specific annotations that are considered as weak annotations. Further contextual information can be captured by learning data-driven spatial transformation models or regularization terms [42]. Related physics-based deformation models have been trained to predict shape changes in segmented organs from sparse annotations, which could

be used for augmented reality purposes [43], [44]. Taking account of the interventional context one step further, Hu *et al.* [45] noticed that in many cases, including MR-TRUS-guided biopsy, the main purpose of interventional data fusion is to propagate a patient-specific target defined on a preoperative image to its interventional counterpart and proposed to replace the registration step by a conditional segmentation one.

Even in scenarios where data-driven similarity metrics may be learned, finding the transformation that optimally aligns a pair of images can remain nontrivial. This is because image similarity is well defined, i.e., informative, only in a narrowly circumscribed vicinity around the true transformation, emphasizing the need for appropriate initialization, such that the initial mismatch falls within the *capture range* of the image similarity metric and optimization algorithm [46]. While adequate initialization is challenging in all registration scenarios, it is considered to be most detrimental in slice-to-volume applications. Such applications are common in image-guided interventions, with the most prominent examples being the bijective alignment of 2-D B-mode ultrasound to 3-D MR or CT volumes or the projective registration of preoperative 3-D MR or CT volumes, or CAD models to intraoperative 2-D X-ray or endoscopy images.

In cases where the 3-D imaging protocol context is well defined, i.e., one is guaranteed to observe the same extent of anatomy, direct approaches to initialization are possible. These methods only accept the 2-D image as input and directly estimate its initial pose relative to a 3-D canonical atlas coordinate system that is implicitly defined by the choice of 3-D image database [47], [48] or tool model [49]. These approaches are attractive, mainly due to two reasons. First, run times are short since only 2-D images must be processed. Second, they lend themselves well for scenarios where 2-D slices are acquired successively to reconstruct a full 3-D volume. However, due to the complexity of the problem and canonical atlas assumption, their performance is often limited in practice.

When a canonical space cannot be defined, alternative approaches typically mimic the external tracking workflow where relative poses are inferred analytically. While external tracking devices require attachment or implantation of artificial fiducial markers to get position information readouts, AI-based approaches seek to establish correspondence directly from the images or from sparse but corresponding image locations. In [50], by learning from a data set of tracked ultrasound, the authors demonstrated that without inference-time reliance on the tracker, deep learning approaches can estimate the 3-D motion occurring in between consecutive 2-D ultrasound images with an accuracy far exceeding that of conventional speckle decorrelation techniques and matching that of the external tracker. This allows for a sensorless 3-D freehand ultrasound and creates new opportunities in computer-assisted interventions. Another complementary powerful concept for trackerless image alignment is the detection and identification of anatomical landmarks. These are particularly appealing since they carry semantic meaning and, consequently, define point correspondence across modalities and domains. Reliably detecting anatomical landmarks is complicated because of changing appearance based on viewpoints but has recently become possible due to powerful convolutional neural network-based image analysis for anatomical landmarks, as shown in the pelvis [46], [51] and knees [52]. The same concept of point correspondence naturally extends to tools and implants where, rather than relying on anatomical landmarks, keypoints on the CAD model are used [53]–[55]. The aforementioned approaches aim at discovering the well-defined points; however, finding the same arbitrary point in multiple images is equally appropriate to establish correspondence. In this formulation of the problem, an AI-based algorithm is trained to produce a pose invariant latent representation of point appearance. Then, query points can be randomly sampled in one image that is then rediscovered in the target image [56], thereby establishing correspondence. This approach is appealing since it does not impose any prior on the imaged object; however, learning a pose invariant latent representation so far has only been demonstrated for comparably small pose differences.

## IV. INTELLIGENT INTRA-OPERATIVE IMAGING

### A. From Data Fusion to Intelligent Imaging

Intelligent intraoperative imaging refers to augmenting the value of intraoperative images for clinical decision-making by providing additional information that is tailored to the context of the intervention. In increasingly granular order, the context here describes the interventional requirements specific to a certain procedure, step in the surgical workflow, decision, or even surgeon's preferences. So far, efforts in this direction are dominated by data fusion methods that seek to enrich intraoperative images with procedural planning information that exists from preoperative data. While this approach, even when relying on classical CAI tools, has been deployed successfully for several types of procedures [33], it is fundamentally limited in its capabilities of fully leveraging all acquired data. This is because the value of intraoperative images is reduced to a proxy to support, e.g., image-based registration or as a means for overlay, while all *intelligent information* that really augments the decision-making is propagated solely from preoperative images. In addition to underexploiting intraoperative images, this strategy only allows for displaying information derived from preoperative data that become outdated as surgery progresses. This calls for the development of intelligent intraoperative imaging that fully leverages the information contained in interventionally acquired data in real-time. Augmenting decision-making in this way offers clear opportunities by: 1) automating quantitative measurements required for precision medicine and 2) extracting information that

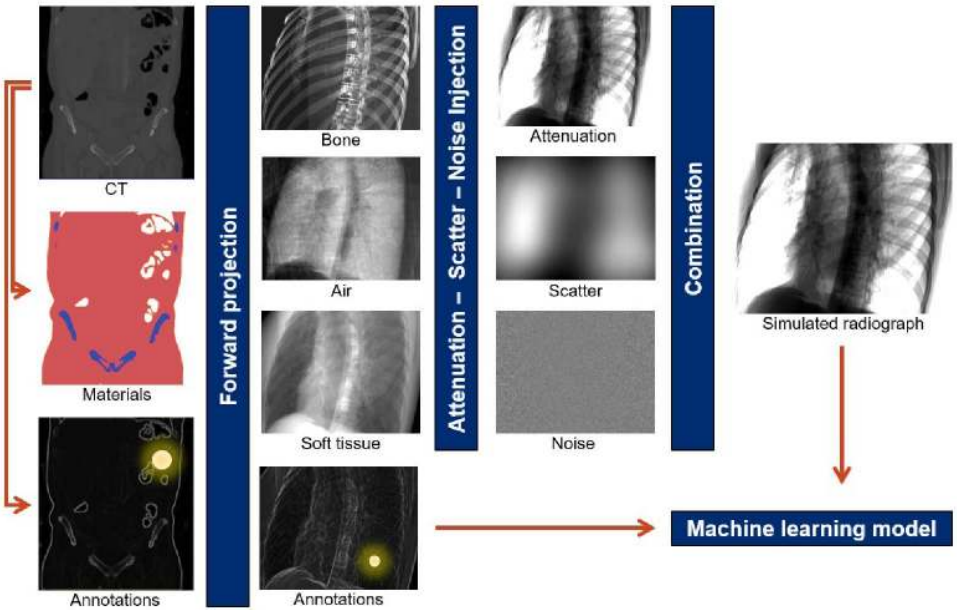


**Fig. 2.** ***Realistic simulation of X-ray image formation from preoperative CT is one possibility to create large quantities of well-annotated images. Pipeline represents the simulation approach described in [57].***

is otherwise not easily accessible, which may allow the development of new surgical techniques. Still, contextual and intelligent interventional image analysis is not yet the mainstream technology because, compared to diagnostic image analysis, the environment for developing AI solutions is even more hostile. From our experience working with clinical collaborators across different sites and specialties, we believe that this is primarily due to three reasons. First, while hundreds of images are acquired for procedural guidance, only very few, if any, are archived [58]–[60], thereby suggesting a severe lack of meaningful data for researchers to work with. Second, learning targets beyond segmentation are not well established or defined. Third, images of the anatomy are acquired from multiple viewpoints, the exact poses of them are not reproduced nor known. Finally, the overall variability in the data is further amplified by surgical modification of anatomy and the presence of tools. Overall, the accessible data are heavily unstructured and exhibits enormous variation, which challenges meaningful data augmentation strategies. As a consequence, in order to train AI algorithms on interventional images, solutions to the data set curation and annotation problem must be found first. Overcoming these hurdles seems challenging and is reflected in the observation that only very little work has considered learning in this context. It is worth mentioning that the lack of annotated and/or paired data equally affects other methods presented in this article.

## B. Simulation-Based Training

Initial steps in addressing the data problem have been taken, serving as a stepping stone for the transformative technology that is *intelligent imaging*. While the large-scale acquisition of highly structured data is tractable for some interventional applications, particularly ultrasound [61], [62], most other approaches rely on synthetic data generation from physical models of the scene. This paradigm is attractive because all quantities of interest are precisely known by design; however, if the simulation is performed naïvely, AI models trained on synthetic data will not generalize to clinically acquired images because of the large domain mismatch paired with poor generalizability of today's models [57]. Three complementary ways have recently been shown to mitigate this problem. First, if the clinically acquired data are available in addition to the well-annotated synthetic data, style transfer algorithms can be trained that alter the appearance of real data to close the domain gap, as shown for the ophthalmic surgical microscopy [63], [64]. Using such enhanced simulated data for training of more complex tasks has been applied successfully to endoscopy [65] and X-ray imaging [66]. Second, if too little clinical data are available, learning a style transfer algorithm is impossible. In these cases, a powerful alternative is increasing the realism of synthetically generated images in a model-based approach. Doing so requires accurate models of all physical principles that govern image formation; however, approximations are usually required to reduce simulation time to acceptable levels. Realistic simulation works well for X-ray-based modalities, as illustrated in Fig. 2 and demonstrated in [57] and [67]. It has also been proposed in endoscopic imaging [68]. However, the level of required realism likely depends on the application and learning target since it has been shown that even less realistic simulations could be adequate, e.g., in some ultrasound applications [69]. The aforementioned approaches aim at reproducing the real data appearance that is very complicated in practice. If closely matching real data appearance is found to be

impossible, domain randomization can be used to improve the robustness of the trained model to partially unseen data. Rather than perfectly matching real data characteristics, the goal of domain randomization is to generate multiple versions of the same sample with all but the important characteristics randomized. When training AI algorithms on such data sets, the models are assumed to become robust to these types of domain changes. Domain randomization can be seen as image formation-based data augmentation and has recently been applied to X-ray imaging [70] as well as colonoscopy [68], where achieving realistic image appearance is very complicated due to fine texture and specular reflectance of the tissue. It is worth mentioning that all the above-mentioned techniques for synthetic data usage are similar in that AI algorithms never process real data during training. This characteristic is associated with a notable drop in performance when applied to real data due to residual domain mismatch. Consequently, assessing algorithmic performance only on a synthetic test set will severely overestimate the AI models accuracy during deployment and quantitative experiments on clinical data are required. Ultimately, training the AI directly on real data is preferable, highlighting the need for further research on unsupervised and self-supervised learning to leverage large quantities of unlabeled data.

### C. Intelligent Imaging in Interventional Biophotonics

Although conventional interventional imaging, such as X-ray fluoroscopy, surgical microscopy, endoscopy, and ultrasound, will benefit from being augmented by contextual AI, another interesting area in which the intelligent imaging paradigm is expected to make an important impact is that of the interventional biophotonics imaging. The initial focus in biophotonics has been on developing optimal, task-specific, contrast agents that would be merely be directly visualized, e.g., in tumor-specific fluorescence imaging. The biophotonics community has, however, faced stringent challenges in identifying versatile contrast agents suitable for use in patients and realized that tissue differentiation would remain challenging with such an approach. Advanced high-dimensional optical imaging techniques are currently seen as promising solutions for intraoperative tissue characterization, with the advantages of being noncontact, nonionizing, and noninvasive or minimally invasive. However, because of the high-dimensional nature of the generated data, direct visualization by the clinical team becomes impractical. This calls for automated learning-based information extraction before display. As in the previous examples of intelligent imaging, many of the most advanced AI-supported interventional biophotonics imaging devices currently exploit model-based learning or unsupervised learning. Point-based measurement devices able to measure the Raman scattering have recently been translated into commercial products [71] with support from supervised classification [72] or unsupervised dimensionality reduction [73]. Addressing the lack of wide-field information in point-based systems, the community has looked into modalities such as hyperspectral imaging [74] with an increasing use of machine learning to solve some of the intrinsic challenges of high-dimensional data. Indeed, while bearing rich information, the raw 2-D -space + wavelength + time data that hyperspectral imaging produce are difficult to interpret for clinicians as it generate a temporal flow of 3-D information that cannot be simply displayed in an intuitive fashion. Innovative use of invertible neural networks in combination with model-driven simulation has been used to train neural network-based regressors that are capable of real-time operation and can provide uncertainty estimates for oxygen saturation measurement from hyperspectral data [75]. Unsupervised deep manifold embedding for hyperspectral imaging was proposed in [76], and deep learning was used for reconstruction from sparse hyperspectral data [77]. Intelligent imaging concept with simulation- or model-based trainings are also being progressed with other emerging biophotonics imaging modalities, such as for superresolution in endomicroscopy [78], [79], and artifact suppression in photoacoustic imaging [80].

### D. Toward Prospectively Planned Intelligent Imaging

With the availability of training data, via either dedicated data collection or synthetic generation, AI algorithms can be developed to analyze intraoperative images in near real time and supply contextual information to improve decision-making. Omitting applications to endoscopic video sources that are discussed in depth in Section V and focusing first on the interventional X-ray imaging, benefits of real-time machine learning range from segmentation of tools [53], [81], [82], anatomical landmark detection [51], [52], anatomy localization [83], and denoising [84], [85], to surgical phase recognition [81]. Corresponding developments can be found for ultrasound imaging [86]–[88].

While the above-mentioned list of applications merely hints at the potential that AI-based analysis of interventional images has to offer, there is an interesting observation: the majority of *intelligent imaging* algorithms, including all the aforementioned methods, try to provide richer information by the automated analysis of traditionally acquired images, with little or no knowledge of the image acquisition workflow. This raises an interesting question: if it is known what information is desired or desirable at any given point during the surgery, is it possible to prospectively acquire an image that is most informative in that particular context? Initial steps in this direction have recently been reported, exploiting ultrasound image formation to suppress scatter [89] or beamforming a B-mode image [90], [91] together with producing its segmentation [69]. Zaech *et al.* [92] use an AI-based algorithm to recommend

task-optimal and patient-specific C-arm X-ray trajectories during cone-beam CT of spinal fusion surgery, and similar ideas arise for ultrasound transducer positioning [93].

The domain of real-time interventional image analysis is fairly untapped as of yet but offers great opportunities for workflow analysis, surgical progress monitoring, including anticipation and adverse event detection, and supplying rich information for human-in-the-loop decision-making. In addition, task-aware and autonomous imaging modalities may benefit interventional imaging already one step before the image is analyzed and may, thus, give rise to disruptive technology and novel surgical approaches.

## V. SURGICAL AND ENDOSCOPIC VISION

### A. Recognizing Endoscopic Activity

Standard endoscopic imaging is certainly the modality most closely relating to natural images. It should, therefore, not be surprising that machine learning tools for interventional images have developed most rapidly in this field. As a proxy for the eyes of the surgeon inside the patient, the endoscopic camera is the privileged source of digital information to understand the activities performed during endoscopic procedures. Endoscopic videos usually capture most of the activities performed within the patient. Recognizing and understanding these activities are essential to develop novel assistance systems that are reactive to the context, e.g., that can provide timely instructions to operating room (OR) staff, enforce safety checkpoints, or log automatically relevant information within the surgical report. Surgical activity recognition from endoscopic videos is, however, a highly challenging task due to the variability existing across patients, surgical treatments, and surgical teams.

In the recent years, a large body of work has focused on recognizing the surgical steps of a procedure directly from the videos [94]–[99]. This has notably been the case in cholecystectomy, a common procedure consisting in removing the gallbladder, which is frequently used in research due to its high frequency of occurrence and well-standardized protocol [100]. There, the steps include, for instance, "the Calot triangle dissection, cystic duct and artery clipping and cutting, gallbladder dissection, and gallbladder packaging." Recognition of these steps allows for the automated understanding of the progress of the surgery. To perform recognition, models of the underlying workflow of the procedure are learned from data sets of exemplary videos, annotated manually with the different steps. In [97], the model consists, for example, of a visual feature extractor relying on a deep neural network that feeds a temporal recognition model, such as a hierarchical hidden Markov model or an LSTM model. Several types of procedures have been successfully studied for step recognition besides cholecystectomy. Examples are cataract surgery [95], [96] and laparoscopic sleeve gastrectomy [98]. As the current recognition methods show very promising results and real-time capabilities, they can potentially be directly embedded in the endoscopic tower to deliver contextual support. Other interesting prediction tasks have been tackled with success using deep learning methods. In [101] and [102], the remaining duration of the procedure is predicted in real time using deep recurrent models trained directly from video data. In [97], [103], and [104], the presence of the instruments in the surgical scene is automatically detected. Additional applications include bleeding and smoke detection [105], [106], as well as surgery type identification at the beginning of the procedure [107].

Beyond the recognition of the surgical steps indicating the progress of the surgery and the recognition of events, such as bleeding, many potential applications, such as safety monitoring and human–robot cooperation, require a finer level of understanding of the surgical activities. Future research, therefore, needs to demonstrate accurate recognition of the detailed interactions between the tools and the anatomy. To have an impact beyond a single OR, recognition methods will also need to scale up to different types of surgeries, ORs, and hospitals without requiring the manual annotations of large data sets for each situation. Recent methods exploiting nonannotated videos through self-supervision or weak-supervision [104], [108]–[111] or exploiting synthetically generated surgeries [64] may prove very useful to train the next generation of surgical recognition systems.

### B. Understanding Image Semantics

Understanding the surgical scene from the endoscopic images is fundamental for context-aware intelligent computer-aided assistance. During augmented reality visualization, precise pixel-based segmentation of the tools is necessary for handling occlusions and providing the user with the correct perception. Implementing safety warnings, such as no-go zones, requires the detection of critical anatomy. When another imaging modality is used, its registration to the endoscopic video may require the localization of anatomical landmarks [113]. Similarly, implementing degrees of autonomy during robotic surgery requires the localization and recognition of the neighboring tools and anatomy.

Recently, a large body of work has targeted the detection and segmentation of surgical instruments [114]. Deep learning methods have been proposed for both bounding box or articulated tool detection [115]–[117] and for pixel-based tool segmentation [118], [119]. Their superiority has been confirmed on laparoscopic and surgical microscopy data sets in two international challenges organized in 2015 and 2017 at the MICCAI conferences [120], [121]. Still, the data sets used for evaluation are limited in size and variability. They are far from representing the diversity of surgical scenes, which can indeed be very challenging due to the presence of occlusions, smoke, bleeding, specularity, motion blur, and deformation. Furthermore, the aforementioned approaches are fully supervised and, therefore, impose

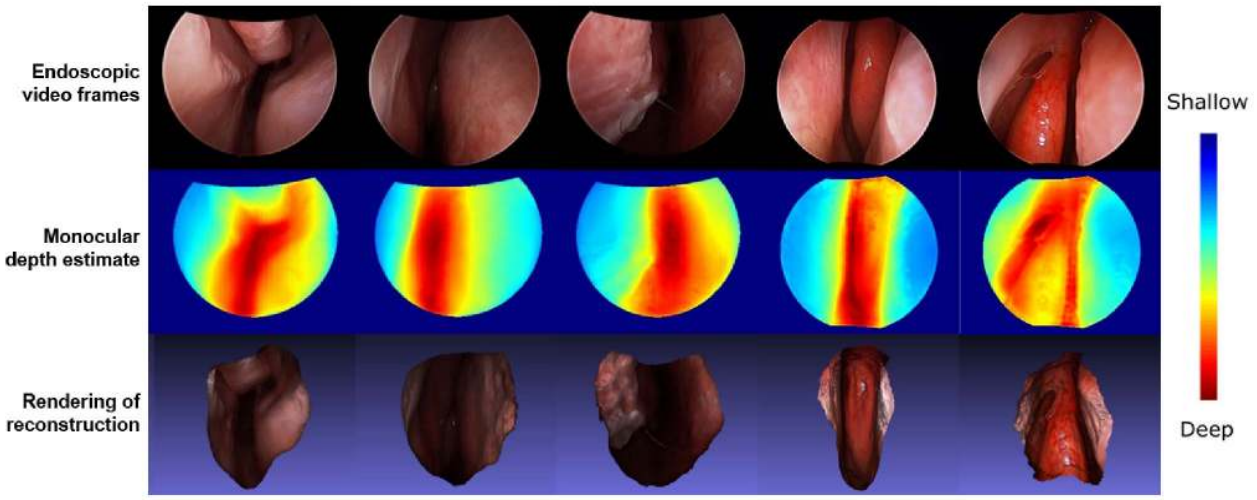


**Fig. 3.** *Endoscopic video (top), monocular depth estimate (middle), and rendering of a photorealistic reconstruction (bottom). Results were achieved using the self-supervised method described in [112].*

an important burden on the collection of representative training data sets. New approaches are needed that can generalize easily to various types of procedures and be trained using weaker information for training, such as image-level tool presence [104], point annotation [122], or scribbles [123].

Far less work has addressed the much needed anatomy detection and segmentation, certainly due to the lack of available public data sets. The community is, however, putting large efforts in this direction, as illustrated by the recent generation of the CaDIS data set [124], which contains pixel-level annotations for 36 semantic classes in cataract surgery videos. Progress has also been achieved in specific areas, such as liver segmentation [125], lesion detection and characterization during gastroscopy [126], or polyp detection during colonoscopy [17], [127]. Here, again, deep learning is the state of the art, as demonstrated for polyp detection in a challenge organized at MICCAI 2015 [128]. Due to the real-time capabilities of deep learning approaches, the intraoperative benefits of such systems already start to be evaluated in RCTs [17].

### C. Reconstructing Anatomic Geometry

Endoscopy mimics the surgeon's eyes within the body, but due to the monocular construction of endoscopes, it lacks one important visual cue: depth. This shortcoming has implications: it has recently been shown that the availability of 3-D anatomic geometry benefits several clinical tasks, including the detection of critical anatomy, such as polyps [129], and the registration of preoperative 3-D data to endoscopy video to enable navigation [130]. In addition, analyzing 3-D representations of anatomy would allow for the introduction of quantitative measurements, enabling the standardization of clinical reporting across sites. Recovering anatomic 3-D geometry, e.g., to augment endoscopic video with depth cues or to provide dense 3-D reconstruction, has gained considerable traction and is now an emerging discipline with developments often orthogonal to those for complementary tasks, e.g., segmentation. This is because deep learning-based algorithms are able to exploit image-level features to provide dense depth estimates even from monocular video, complementing traditional optical endoscopy with depth sensing as "pseudomodality." However, training depth estimation algorithms on endoscopic sequences is complicated in practice because no paired depth measurements exist naturally. While paired data can be generated in silico via simulation from CT [65], [68], [131], the resulting trained models will need to overcome the domain mismatch to real clinical data with methods similar to that presented in Section IV. Recently, self-supervised training paradigms that rely on traditional multiview stereo approaches have received increasing attention as they can be trained directly and solely from the endoscopic video. Multiview stereo algorithms, including structure from motion [112], [130] and simultaneous localization and mapping [132], can be adapted to work with endoscopic video, but they cannot provide dense 3-D reconstructions due to the lack of photometric constancy in endoscopic video and texture scarceness that complicate feature matching across frames. These algorithms do, however, provide a few reconstructed 3-D points and, more importantly, relative camera poses that can be used to supervise monocular depth estimation [112], [132]. A representative photorealistic reconstruction achieved using a structure from motion supervised depth estimation method is shown in Fig. 3. These methods achieve state-of-the-art performance with good generalization ability; however, the resulting reconstructions are only up to scale. Among the biggest premises of video-based reconstruction is the possibility of monitoring anatomical change during surgery. This would require methods to robustly handle various sorts of uncontrollable variation, including bleeding, smoke, or tool pres-

ence. Solutions to these problems are currently unknown. Even in more controlled scenarios, widespread adoption of learning-based reconstruction from the endoscopic video is hindered by the lack of publicly available data sets, making it unclear how well today's algorithms perform on clinical data. This challenge is further aggravated by the lack of direct evaluation targets. When applied to real clinical data, current reconstruction or dense estimation algorithms can only be evaluated via surrogate tasks, such as video-CT registration [112], [133] or polyp classification [129].

## VI. CLINICAL WORKFLOW MONITORING AND SUPPORT

### A. Notion of Surgical Control Tower

While imaging alone provides valuable information, modern procedures rely increasingly on a variety of complex devices and intricate workflows. This limits the knowledge extraction that AI systems can do based on imaging alone and makes it difficult for humans to properly analyze in real time the wealth of available data. Furthermore, even though the quality of care has generally improved with the introduction of new surgical techniques and devices, adverse events still occur, a large part of that are preventable [135], [136]. Humans are prone to fatigue, teams to miscommunications, devices can fail, and for all roles, surgical tasks require an ever-increasing level of specialization. The increased use of digital equipment in the OR, however, opens up new opportunities for support and monitoring, at the level of the whole room, by providing artificial intelligence systems with real-time data that capture a faithful representation of the processes taking place during the surgery. Indeed, most of the activities happening in the room can be captured digitally either through interactions with equipment, such as information systems, room control interfaces, imaging devices and instruments, or through the use of sensors, such as ceiling-mounted cameras, which are now becoming widespread and increasingly used for documentation, teaching, and augmented reality assistance. Consequently, it is highly likely that in the near future, assistance systems will be fully integrated in a digital OR that will monitor surgical processes through AI, akin to a *surgical control tower* [137], [138], that can analyze the whole digital information in real time to provide context-aware support and information within and outside the OR. Applications for such a control tower are, for instance, the transmission of live information about the OR status, the adaptation of user-interfaces to the surrounding context, the display of instructions within the OR, the creation of an automated report, the recording of the activities for archiving and legal purposes, the enforcement of safety checklists, the detection of anomalies with respect to past workflows, and improved scheduling for staff and patients. To perform these tasks, the control tower will have access to and crunch masses of multimodal digital data coming from hundreds of past surgeries.

### B. Endeavor Rooted in Surgical Data Science

An essential component of the control tower is the data-driven modeling and understanding of the clinical activities, an undertaking that taps into the emerging research field of surgical data science [3], [4]. Machine learning has been key to generate models of procedural interventions from data [139], [140], and ontologies have also been developed to standardize the resulting models [141]. Implementations of such AI-based applications start to emerge in various institutions, besides the ones focusing on analyzing endoscopic videos already mentioned in Section V. As video data remain one of the main sources of information, they highly rely on deep learning. Videos captured by the cameras mounted in the room provide indeed a rich source of information about the activities without disrupting the workflow. For instance, a patient and staff radiation exposure monitoring system for hybrid procedures illustrated in Fig. 4 was proposed in [134]. It relies on several RGB-D cameras to estimate the 3-D pose of the persons and room layout, which can then be used to simulate and visualize *in situ* X-ray propagation around the patient table. Haque *et al.* [142] develop a system to monitor hand hygiene in hospital corridors in order to analyze and reduce the hospital-acquired infection. The approach uses a large set of depth cameras installed to observe the hand-soap dispensers. For the intensive care unit, Ma *et al.* [143] and Yeung *et al.* [144] present methods based on color or depth video data for the detection of patient mobilization activities. Key building blocks to the success of these applications are the estimation of clinician and staff poses [145]–[147], as well as the recognition of their activities [148]–[151]. As for traditional visual data, deep learning-based approaches are currently the best-performing methods for these tasks though it should be noted that they do not necessarily perform as well on clinical data yet. This is due to the specificity of clinical videos, where staffs wear gowns and masks, colors are often similar, and cameras observe the room from restricted positions, but also from the fact that there is no clinical COCO or Imagenet data set yet. Srivastav *et al.* [152] evaluate the state-of-the-art human pose estimation approaches, and Issenhuth *et al.* [153] evaluate the state-of-the-art face detection approaches on clinical data. Both studies show a large margin for improvement. Since the development of large annotated data sets of clinical videos may be difficult due to the expertise required and the restrictions on data, other approaches need to be developed, for instance, using the nonannotated data for transfer learning [153].

This will also help deploy the surgical control tower in new clinical environments, as the variability in room layout, camera configuration, and workflow can be high. Retraining the assistance systems using only nonannotated data from the novel environment or a tiny subset of annotated data will be crucial for the adoption of these technologies. As even the collection of nonanno-

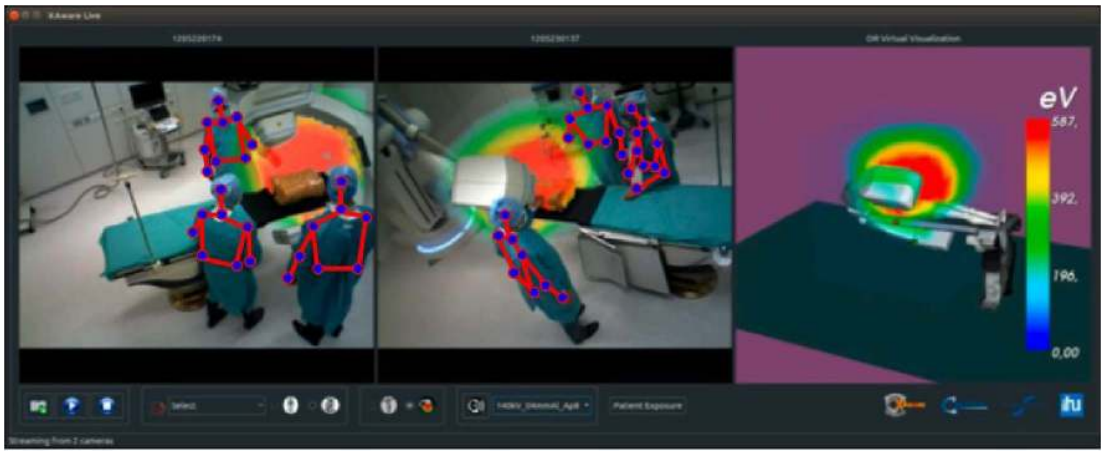


**Fig. 4.** *Capturing the 3-D context of the OR is necessary for providing AI-based decision support and monitoring risk. In this example, the staff radiation exposure during an X-ray-based procedure is computed in situ via simulation and displayed with augmented reality in a training scenario [134].*

tated video data can be challenging due to data and privacy regulations, it may also be required to implement federated learning approaches or develop methods that are able to cope with privacy-preserving data, such as depth-only videos [142] or even low-resolution depth videos [154]. In [154], it is shown that 2-D human pose estimation can be achieved with reasonably high accuracy on depth images downsampled by ten to the resolution of $64 \times 48$. By using other information, such as system events [155] or speech analysis [156], the analysis of clinical activities will be further improved.

## VII. DISCUSSION AND CONCLUSION

While AI is starting to impact CAI, as described in this article, there is a number of challenges that are specific to surgery and intervention to overcome to deliver clinical impact. The leveraging context within learning paradigms will be crucial to address those in a clinically meaningful way. The emerging field of CAI4CAI offers researchers a large set of open problems to tackle. These notably stem from the heterogeneity of surgical procedures and their particular requirements for intraoperative imaging [157], the difficulties in data acquisition, the complexity in modeling and inferring decision-making processes, and the intricacy of the execution of surgical tasks. Over the years, the CAI community has defined increasingly powerful surgical process models [158] to gain an actionable understanding of surgical procedures while describing interventions as a sequence of tasks and activities at different granularity levels. At the finest level, mapping what should be the *Language of Surgery* [159], researchers currently break down surgical gestures into semantically relevant motion units called *surgemes* that are further composed of sequences of motion primitives named *dexemes* [160]–[162]. However, this taxonomy mostly focused on the surgical action and, in particular, on surgical tool manipulation and could, thus, rather be considered as mapping the *Language of Surgical Dexterity*. This is already a laudable achievement and led to scientists and engineers being able to, e.g., quantify the success of a training program for executing different surgical actions [163], [164]. As suggested by the study conducted by Birkmeyer *et al.* [165] for bariatric surgery, surgical skills can be highly correlated with the surgical outcome for certain procedures. AI systems have been shown capable of evaluating technical skills using data from either training scenarios [166] or real procedures [167]. However, by severely underutilizing the rich information contained in other data sources, the *Language of Surgical Dexterity* is still not capturing the most complex aspects of surgical decision-making. To address the need to capture, understand, and support all the cognitive interactions and processes taking place in the OR, the surgical data science community will need to drive the deployment of real-time multimodal data acquisition systems that will be used routinely. At the same time, it will foster the development of new standards and regulations aiming at increasing the interoperability of data, devices, and models. This will directly benefit CAI4CAI by simplifying the implementation and training of learning algorithms involving databases from multiple institutions while maintaining privacy, e.g., through federated learning. CAI4CAI in combination with surgical data science and surgical process modeling could, thus, aim at defining and understanding the ultimate *Language of Surgery* based on a large number of heterogeneous data sources used continuously by surgeons and interventional teams to guarantee the best outcomes for a given procedure. As the field blossoms, CAI4CAI researchers will address some of the most rewarding questions in computer-assisted intervention. Could CAI4CAI allow us to learn how decisions are made, or missed, throughout surgical procedures? Could CAI4CAI support such decision-makings? Instead of going

through the traditional path of segmentation, registration, navigation, and visualization, could contextual machine learning allow us to optimize these steps for each given objective and allow for real-time computation and feedback based on a large amount of heterogeneous data, including preoperative and intraoperative imaging, patient characteristics, and surgeon preferences?

With more capable and flexible learning paradigms, synergistic collaboration is expected to happen between humans and AI-powered actors. The field is already seeing exciting attempts to bring the user and the user experience at the center of our research questions. For example, novel spatially aware visualization beyond traditional user interfaces is explored in [134] and [168]. The challenge of improving human situational awareness in operating with solutions beyond visualization is addressed in [169] with the use of context-specific soundtracks. Introduction of novel multimodal interaction paradigms and technologies within ORs will require extensive use of machine learning to optimize the user interfaces and to provide maximally relevant information and support while preventing inattentional blindness [170]. By developing systems that are able to learn from previous surgeries performed by experts how to provide context-aware support and instructions directly in the OR, in the manner of a virtual coach, as in [171], AI could have a strong impact on improving patient care. This is another aspect of CAI4CAI that needs particular focus from the scientific community and requires MDTs, including clinicians, user experience experts, and machine learning scientists, to work together and come up with intelligent end-to-end CAI solutions.

Finally, in this article, we did not have a particular focus on robotics. However, both surgical robotics and robotic imaging will play increasingly crucial roles in the years to come. Machine learning is demonstrating convincing results in real-time tool tracking [118], [172]–[174]. This, for example, enables automatic positioning of intraoperative OCT imaging planes within surgical microscopy for ophthalmic surgery [119], [175]. Integration of robotics within surgical suites would require them to act intelligently and synergistically with the human team and to be fully context-aware at all moments. The wish to have real-time multimodal imaging requires full intelligence and automation. It also requires direct communication and collaboration between surgical robots, imaging robots, surgeons, and surgical teams. CAI4CAI will have the challenge of enabling such ultimate intelligence, which requires many years of research and development in many disciplines while remembering a past experience with the first generation of context-aware computing [176]. Not only does CAI4CAI offer numerous exciting research directions but it also promises to revolutionize surgery and, therefore, the future of healthcare at a global scale. ■

## REFERENCES


[1] A. L. Simpson *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," Feb. 2019, *arXiv:1902.09063*. [Online]. Available: https://arxiv.org/abs/1902.09063

[2] E. Gibson *et al.*, "NiftyNet: A deep-learning platform for medical imaging," *Comput. Methods Programs Biomed.*, vol. 158, pp. 113–122, May 2018.

[3] L. Maier-Hein *et al.*, "Surgical data science for next-generation interventions," *Nature Biomed. Eng.*, vol. 1, no. 9, pp. 691–696, 2017.

[4] S. S. Vedula and G. D. Hager, "Surgical data science: The new knowledge domain," *Innov. Surgical Sci.*, vol. 2, no. 3, pp. 109–121, 2007.

[5] L. Maier-Hein *et al.*, "Surgical data science: A consensus perspective," 2018, *arXiv:1806.03184*. [Online]. Available: https://arxiv.org/abs/1806.03184

[6] T. Peters and K. Cleary, *Image-Guided Interventions Technology and Applications*. Boston, MA, USA: Springer, 2008.

[7] J. West *et al.*, "Comparison and evaluation of retrospective intermodality brain image registration techniques," *J. Comput. Assist. Tomogr.*, vol. 21, no. 4, pp. 554–568, 1997.

[8] G. Wang *et al.*, "DeepIGeoS: A deep interactive geodesic framework for medical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1559–1572, Jul. 2019.

[9] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, 2015.

[10] D. Comaniciu, K. Engel, B. Georgescu, and T. Mansi, "Shaping the future through innovations: From medical imaging to precision medicine," *Med. Image Anal.*, vol. 33, pp. 19–26, Oct. 2016.

[11] W. Li, G. Wang, L. Fidon, S. Ourselin, M. J. Cardoso, and T. Vercauteren, "On the compactness, efficiency, and representation of 3D convolutional networks: Brain parcellation as a pretext task," in *Information Processing in Medical Imaging* (Lecture Notes in Computer Science), vol. 10265. Cham, Switzerland: Springer-Verlag, Jun. 2017, pp. 348–360.

[12] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[13] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BraTS challenge," 2018, *arXiv:1811.02629*. [Online]. Available: https://arxiv.org/abs/1811.02629

[14] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *Proc. BrainLes Challenge MICCAI*, 2018, pp. 92–104.

[15] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, "BrainTorrent: A peer-to-peer environment for decentralized federated learning," 2019, *arXiv:1905.06731*. [Online]. Available: https://arxiv.org/abs/1905.06731

[16] K. Kamnitsas *et al.*, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Proc. IPMI*, 2017, pp. 597–609.

[17] P. Wang *et al.*, "Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: A prospective randomised controlled study," in *Proc. Gut*, 2019, pp. 1813–1819.

[18] G. Wang *et al.*, "Interactive medical image segmentation using deep learning with image-specific fine-tuning," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1562–1573, Jul. 2018.

[19] R. Dorent, W. Li, J. Ekanayake, S. Ourselin, and T. Vercauteren, "Learning joint lesion and tissue segmentation from task-specific hetero-modal datasets," in *Proceedings of Machine Learning Research*, vol. 102, M. J. Cardoso *et al.* Eds. Cambridge, MA, USA: PMLR, 2019, pp. 164–174.

[20] R. Dorent, S. Joutard, M. Modat, S. Ourselin, and T. Vercauteren, "Hetero-modal variational encoder-decoder for joint modality completion and segmentation," in *Proc. MICCAI*, 2019, pp. 74–82.

[21] M. D. A. M. Digioia, III, B. Jaramaz, C. Nikou, R. S. Labarca, J. E. Moody, and B. D. Colgan, "Surgical navigation for total hip replacement with the use of HipNav," *Operative Techn. Orthopaedics*, vol. 10, no. 1, pp. 3–8, 2000.

[22] M. Boudissa, A. Courvoisier, M. Chabanas, and J. Tonetti, "Computer assisted surgery in preoperative planning of acetabular fracture surgery: State of the art," *Expert Rev. Med. Devices*, vol. 15, no. 1, pp. 81–89, 2018.

[23] P. Kulyk, L. Vlachopoulos, P. Fürnstahl, and G. Zheng, "Fully automatic planning of total shoulder arthroplasty without segmentation: A deep learning based approach," in *Computational Methods and Clinical Applications in Musculoskeletal Imaging*, T. Vrtovec, J. Yao, G. Zheng, and J. M. Pozo, Eds. Cham, Switzerland: Springer, 2019, pp. 22–34.

[24] F. Carrillo, L. Vlachopoulos, A. Schweizer, L. Nagy, J. Snedeker, and P. Fürnstahl, "A time saver: Optimization approach for the fully automatic 3D planning of forearm osteotomies," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham, Switzerland: Springer, 2017, pp. 488–496.

[25] R. Sparks *et al.*, "Automated multiple trajectory planning algorithm for the placement of stereo-electroencephalography (SEEG) electrodes in epilepsy treatment," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 1, pp. 123–136, Jan. 2017.

[26] V. N. Vakharia *et al.*, "Automated trajectory planning for laser interstitial thermal therapy in

mesial temporal lobe epilepsy," *Epilepsia*, vol. 59, no. 4, pp. 814–824, 2018.
[27] A. Granados *et al.*, "A machine learning approach to predict instrument bending in stereotactic neurosurgery," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 238–246.
[28] S. Giffard-Roisin *et al.*, "Transfer learning from simulations on a reference anatomy for ECGI in personalized cardiac resynchronization therapy," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 2, pp. 343–353, Feb. 2018.
[29] J. Ramalhinho *et al.*, "A pre-operative planning framework for global registration of laparoscopic ultrasound to CT images," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 8, pp. 1177–1186, Aug. 2018.
[30] S. F. Johnsen *et al.*, "Database-based estimation of liver deformation under pneumoperitoneum for surgical image-guidance and simulation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 450–458.
[31] K. März *et al.*, "Toward knowledge-based liver surgery: Holistic information processing for surgical decision support," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 10, no. 6, pp. 749–759, Jun. 2015.
[32] T. Koivukangas, J. P. Katisko, and J. P. Koivukangas, "Technical accuracy of optical and the electromagnetic tracking systems," *SpringerPlus*, vol. 2, no. 1, p. 90, 2013.
[33] F. A. Jolesz, *Intraoperative Imaging and Image-Guided Therapy*. New York, NY, USA: Springer, 2014.
[34] J. P. W. Pluim and J. M. Fitzpatrick, "Image registration," *IEEE Trans. Med. Imag.*, vol. 22, no. 11, pp. 1341–1343, Nov. 2003.
[35] R. Liao, L. Zhang, Y. Sun, S. Miao, and C. Chefd, "A review of recent advances in registration techniques applied to minimally invasive therapy," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 983–1000, Aug. 2013.
[36] L. Joskowicz and E. J. Hazan, "Computer aided orthopaedic surgery: Incremental shift or paradigm change?" *Med. Image Anal.*, vol. 33, pp. 84–90, Oct. 2016.
[37] B. Fuerst, W. Wein, M. Müller, and N. Navab, "Automatic ultrasound'MRI registration for neurosurgery using the 2D and 3D LC2 metric," *Med. Image Anal.*, vol. 18, no. 8, pp. 1312–1319, 2014.
[38] W. Wein, "Brain-shift correction with image-based registration and landmark accuracy evaluation," in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*. Cham, Switzerland: Springer, 2018, pp. 146–151.
[39] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Med. Image Anal.*, vol. 52, pp. 128–143, Feb. 2018.
[40] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1788–1800, Aug. 2019.
[41] Y. Hu *et al.*, "Weakly-supervised convolutional neural networks for multimodal image registration," *Med. Image Anal.*, vol. 49, pp. 1–13, Oct. 2018.
[42] Y. Hu *et al.*, "Adversarial deformation regularization for training image registration neural networks," in *Medical Image Computing and Computer Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 11070. Cham, Switzerland: Springer, 2018, pp. 774–782.
[43] M. Pfeiffer, C. Riediger, J. Weitz, and S. Speidel, "Learning soft tissue behavior of organs for surgical navigation with convolutional neural networks," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 7, pp. 1147–1155, Jul. 2019.
[44] J.-N. Brunet, A. Mendizabal, A. Petit, N. Golse, E. Vibert, and S. Cotin, "Physics-based deep neural network for augmented reality during liver surgery," in *Medical Image Computing and Computer Assisted Intervention—MICCAI* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2019.
[45] Y. Hu, E. Gibson, D. C. Barratt, M. Emberton, J. A. Noble, and T. Vercauteren, "Conditional segmentation in lieu of image registration," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2019.
[46] J. Esteban, M. Grimm, M. Unberath, G. Zahnd, and N. Navab, "Towards fully automatic X-ray to CT registration," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2019.
[47] B. Hou *et al.*, "Predicting slice-to-volume transformation in presence of arbitrary subject motion," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2017, pp. 296–304.
[48] M. Bui, S. Albarqouni, M. Schrapp, N. Navab, and S. Ilic, "X-Ray PoseNet: 6 DoF pose estimation for mobile X-Ray devices," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 1036–1044.
[49] S. Miao, Z. J. Wang, and R. Liao, "A CNN regression approach for real-time 2D/3D registration," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1352–1363, May 2016.
[50] R. Prevost *et al.*, "3D freehand ultrasound without external tracking using deep learning," *Med. Image Anal.*, vol. 48, pp. 187–202, Aug. 2018.
[51] B. Bier *et al.*, "X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2018, pp. 55–63.
[52] B. Bier *et al.*, "Detecting anatomical landmarks for motion estimation in weight-bearing imaging of knees," in *Proc. Int. Workshop Mach. Learn. Med. Image Reconstruct.* Cham, Switzerland: Springer, 2018, pp. 83–90.
[53] C. Gao, M. Unberath, R. Taylor, and M. Armand, "Localizing dexterous surgical tools in X-ray for image-based navigation," in *Proc. IPCAI*. Cham, Switzerland: Springer, 2019, pp. 1–4.
[54] D. Kügler, A. Stefanov, and A. Mukhopadhyay, "i3PosNet: Instrument pose estimation from X-ray," 2018, *arXiv:1802.09575*. [Online]. Available: https://arxiv.org/abs/1802.09575
[55] H. Esfandiari, R. Newell, C. Anglin, J. Street, and A. J. Hodgson, "A deep learning framework for segmentation and pose estimation of pedicle screw implants based on C-arm fluoroscopy," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 8, pp. 1269–1282, 2018.
[56] H. Liao, W.-A. Lin, J. Zhang, J. Zhang, J. Luo, and S. K. Zhou, "Multiview 2D/3D rigid registration via a point-of-interest network for tracking and triangulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12638–12647.
[57] M. Unberath *et al.*, "Enabling machine learning in X-ray-based procedures via realistic simulation of image formation," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 9, pp. 1517–1528, 2019.
[58] D. A. Clunie *et al.*, "Technical challenges of enterprise imaging: HIMSS-SIIM collaborative white paper," *J. Digit. Imag.*, vol. 29, no. 5, pp. 583–614, 2016.
[59] F. M. Murad *et al.*, "Image management systems," *Gastrointestinal Endoscopy*, vol. 79, no. 1, pp. 15–22, 2014.
[60] A. Deganello, M. E. Sellars, G. T. Yusuf, and P. S. Sidhu, "How much should i record during a CEUS examination? Practical aspects of the 'real-time' feature of a contrast ultrasound study," *Ultraschall Med.*, vol. 39, no. 5, pp. 484–486, 2018.
[61] F. Milletari, V. Birodkar, and M. Sofka, "Straight to the point: Reinforcement learning for user guidance in ultrasound," 2019, *arXiv:1903.00586*. [Online]. Available: https://arxiv.org/abs/1903.00586
[62] Y. Hu *et al.*, "Freehand ultrasound image simulation with spatially-conditioned generative adversarial networks," in *Proc. RAMBO Workshop MICCAI* (Lecture Notes in Computer Science), vol. 10555. Cham, Switzerland: Springer-Verlag, Sep. 2017, pp. 105–115.
[63] S. Engelhardt, R. De Simone, P. M. Full, M. Karck, and I. Wolf, "Improving surgical training phantoms by hyperrealism: Deep unpaired image-to-image translation from real surgeries," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. 2018, pp. 747–755.
[64] I. Luengo, E. Flouty, P. Giataganas, P. Wisanuvej, J. Nehme, and D. Stoyanov, "SurReal: Enhancing surgical simulation realism using style transfer," in *Proc. Brit. Mach. Vis. Conf.*, Newcastle, U.K., Sep. 2018, p. 116.
[65] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2572–2581, Dec. 2018.
[66] Y. Zhang, S. Miao, T. Mansi, and R. Liao, "Task driven generative modeling for unsupervised domain adaptation: Application to X-ray image segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2018.
[67] M. Unberath *et al.*, "DeepDRR—A catalyst for machine learning in fluoroscopy-guided procedures," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2018.
[68] F. Mahmood, R. Chen, S. Sudarsky, D. Yu, and N. J. Durr, "Deep learning with cinematic rendering: Fine-tuning deep neural networks using photorealistic medical images," 2018, *arXiv:1805.08400*. [Online]. Available: https://arxiv.org/abs/1805.08400
[69] A. A. Nair, T. D. Tran, A. Reiter, and M. A. L. Bell, "A deep learning based alternative to beamforming ultrasound images," in *Proc. ICASSP*, 2018, pp. 3359–3363.
[70] D. Toth, S. Cimen, P. Ceccaldi, T. Kurzendorfer, K. Rhode, and P. Mountney, "Training deep networks on domain randomized synthetic X-ray data for cardiac interventions," in *Proceedings of Machine Learning Research*, vol. 102, M. J. Cardoso *et al.* Eds. Cambridge, MA, USA: PMLR, 2019, pp. 468–482.
[71] M. Jermyn *et al.*, "Intraoperative brain cancer detection with Raman spectroscopy in humans," *Sci. Transl. Med.*, vol. 7, no. 274, p. 274ra19, 2015.
[72] M. Jermyn *et al.*, "Neural networks improve brain cancer detection with Raman spectroscopy in the presence of operating room light artifacts," *J. Biomed. Opt.*, vol. 21, no. 9, p. 094002, 2016.
[73] C. Banbury *et al.*, "Development of the self optimising Kohonen index network (SKiNET) for Raman spectroscopy based detection of anatomical eye tissue," *Sci. Rep.*, vol. 9, no. 1, p. 10812, 2019.
[74] J. Shapey *et al.*, "Intraoperative multispectral and hyperspectral label-free imaging: A systematic review of *in vivo* clinical studies," *J. Biophoton.*, vol. 12, no. 9, p. e201800455, 2019.
[75] T. J. Adler *et al.*, "Uncertainty-aware performance assessment of optical imaging modalities with invertible neural networks," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 6, pp. 997–1007, 2019.
[76] D. Ravì, H. Fabelo, G. M. Callic, and G.-Z. Yang, "Manifold embedding and semantic segmentation for intraoperative guidance with hyperspectral brain imaging," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1845–1857, Sep. 2017.
[77] J. Lin *et al.*, "Dual-modality endoscopic probe for tissue surface shape reconstruction and hyperspectral imaging enabled by deep neural networks," *Med. Image Anal.*, vol. 48, pp. 162–176, Aug. 2018.
[78] D. Ravì, A. B. Szczotka, S. P. Pereira, and T. Vercauteren, "Adversarial training with cycle

consistency for unsupervised super-resolution in endomicroscopy," *Med. Image Anal.*, vol. 53, pp. 123–131, Apr. 2019.

[79] A. B. Szczotka, D. Ravì, D. I. Shakir, S. P. Pereira, and T. Vercauteren, "Effective deep learning training for single-image super-resolution in endomicroscopy exploiting video-registration-based reconstruction," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 6, pp. 917–924, Jun. 2018.

[80] D. Allman, F. Assis, J. Chrispin, and M. A. L. Bell, "Deep neural networks to remove photoacoustic reflection artifacts in *ex vivo* and *in vivo* tissue," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2018, pp. 1–4.

[81] P. Ambrosini, D. Ruijters, W. J. Niessen, A. Moelker, and T. van Walsum, "Fully automatic and real-time catheter segmentation in X-ray fluoroscopy," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2017, pp. 577–585.

[82] K. Breininger *et al.*, "Multiple device segmentation for fluoroscopic imaging using multi-task learning," in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Cham, Switzerland: Springer, 2018, pp. 19–27.

[83] R. Sa *et al.*, "Intervertebral disc detection in X-ray images using faster R-CNN," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 564–567.

[84] Y. Matviychuk *et al.*, "Learning a multiscale patch-based representation for image denoising in X-ray fluoroscopy," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2330–2334.

[85] S. G. Hariharan *et al.*, "Learning-based X-ray image denoising utilizing model-based image simulations," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2019.

[86] E. M. A. Anas, P. Mousavi, and P. Abolmaesumi, "A deep learning approach for real time prostate segmentation in freehand ultrasound guided biopsy," *Med. Image Anal.*, vol. 48, pp. 107–116, Aug. 2018.

[87] C. Mwikirize, J. L. Nosher, and I. Hacihaliloglu, "Convolution neural networks for real-time needle detection and localization in 2D ultrasound," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 5, pp. 647–657, 2018.

[88] M. Pesteie, V. Lessoway, P. Abolmaesumi, and R. N. Rohling, "Automatic localization of the needle target for ultrasound-guided epidural injections," *IEEE Trans. Med. Imag.*, vol. 37, no. 1, pp. 81–92, Jan. 2017.

[89] A. C. Luchies and B. C. Byram, "Deep neural networks for ultrasound beamforming," *IEEE Trans. Med. Imag.*, vol. 37, no. 9, pp. 2010–2021, Sep. 2018.

[90] W. Simson *et al.*, "End-to-end learning-based ultrasound reconstruction," 2019, *arXiv:1904.04696*. [Online]. Available: https://arxiv.org/abs/1904.04696

[91] W. Simson, M. Paschali, N. Navab, and G. Zahnd, "Deep learning beamforming for sub-sampled ultrasound data," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2018, pp. 1–4.

[92] J.-N. Zaech *et al.*, "Learning to avoid poor images: Towards task-aware C-arm cone-beam CT trajectories," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2019.

[93] F. Milletari, N. Rieke, M. Baust, M. Esposito, and N. Navab, "CFCM: Segmentation via coarse to fine context memory," 2018, *arXiv:1806.01413*. [Online]. Available: https://arxiv.org/abs/1806.01413

[94] T. Blum, N. Padoy, H. Feußner, and N. Navab, "Modeling and online recognition of surgical phases using hidden Markov models," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2008, pp. 627–635.

[95] F. Lalys, L. Riffaud, D. Bouget, and P. Jannin, "A framework for the recognition of high-level surgical tasks from video images for cataract surgeries," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 4, pp. 966–976, Dec. 2012.

[96] G. Quellec, M. Lamard, B. Cochener, and G. Cazuguel, "Real-time segmentation and recognition of surgical tasks in cataract surgery videos," *IEEE Trans. Med. Imag.*, vol. 33, no. 12, pp. 2352–2360, Dec. 2014.

[97] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "EndoNet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 86–97, Jan. 2017.

[98] M. Volkov, D. A. Hashimoto, G. Rosman, O. R. Meireles, and D. Rus, "Machine learning and coresets for automated real-time video segmentation of laparoscopic and robot-assisted surgery," in *Proc. ICRA*, 2017, pp. 754–759.

[99] S. Bodenstedt *et al.*, "Active learning using deep Bayesian networks for surgical workflow analysis," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 6, pp. 1079–1087, 2019.

[100] N. Padoy, T. Blum, S.-A. Ahmadi, H. Feussner, M.-O. Berger, and N. Navab, "Statistical modeling and recognition of surgical workflow," *Med. Image Anal.*, vol. 16, no. 3, pp. 632–641, 2012.

[101] A. P. Twinanda, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "RSDNet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations," *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 1069–1078, Apr. 2019.

[102] S. Bodenstedt *et al.*, "Prediction of laparoscopic procedure duration using unlabeled, multimodal sensor data," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 6, pp. 1089–1095, 2019.

[103] H. Al Hajj, M. Lamard, P.-H. Conze, B. Cochener, and G. Quellec, "Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks," *Med. Image Anal.*, vol. 47, pp. 203–218, Jul. 2018.

[104] C. I. Nwoye, D. Mutter, J. Marescaux, and N. Padoy, "Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 6, pp. 1059–1067, 2019.

[105] A. Leibetseder, M. J. Primus, and K. Schoeffmann, "Automatic smoke classification in endoscopic video," in *Proc. 24th Int. Conf. MultiMedia Modeling (MMM)*, Bangkok, Thailand, Feb. 2018, pp. 362–366.

[106] T. Okamoto, T. Ohnishi, H. Kawahira, O. Dergachyava, P. Jannin, and H. Haneishi, "Real-time identification of blood regions for hemostasis support in laparoscopic surgery," *Signal, Image Video Process.*, vol. 13, no. 2, pp. 405–412, 2019.

[107] S. Kannan, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "Future-state predicting LSTM for early surgery type recognition," *IEEE Trans. Med. Imag.*, to be published.

[108] I. Funke, A. Jenke, S. T. Mees, J. Weitz, S. Speidel, and S. Bodenstedt, "Temporal coherence-based self-supervised learning for laparoscopic workflow analysis," in *Proc. 1st Int. Workshop, OR 2.0 5th Int. Workshop (CARE), 7th Int. Workshop (CLIP), 3rd Int. Workshop (ISIC)*, Granada, Spain, Sep. 2018, pp. 85–93.

[109] G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of CNN-LSTM networks," 2018, *arXiv:1805.08569*. [Online]. Available: https://arxiv.org/abs/1805.08569

[110] T. Roß *et al.*, "Exploiting the potential of unlabeled endoscopic video data with self-supervised learning," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 6, pp. 925–933, 2018.

[111] T. Yu, D. Mutter, J. Marescaux, and N. Padoy, "Learning from a tiny dataset of manual annotations: A teacher/student approach for surgical phase recognition," in *Proc. IPCAI*, 2019, pp. 1–13.

[112] X. Liu *et al.*, "Self-supervised learning for dense depth estimation in monocular endoscopy," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Cham, Switzerland: Springer, 2018, pp. 128–138.

[113] D. Ntourakis, R. Memeo, L. Soler, J. Marescaux, D. Mutter, and P. Pessaux, "Augmented reality guidance for the resection of missing colorectal liver metastases: An initial experience," *World J. Surg.*, vol. 40, no. 2, pp. 419–426, 2015.

[114] D. Bouget, M. Allan, D. Stoyanov, and P. Jannin, "Vision-based and marker-less surgical tool detection and tracking: A review of the literature," *Med. Image Anal.*, vol. 35, pp. 633–654, Jan. 2017.

[115] D. Sarikaya, J. J. Corso, and K. A. Guru, "Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1542–1549, 2017.

[116] T. Kurmann *et al.*, "Simultaneous recognition and pose estimation of instruments in minimally invasive surgery," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2017, pp. 505–513.

[117] A. Jin *et al.*, "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Tahoe, NV, USA, Mar. 2018, pp. 691–699.

[118] L. C. Garcia-Peraza-Herrera *et al.*, "ToolNet: Holistically-nested real-time segmentation of robotic surgical tools," in *Proc. IROS*, 2017, pp. 5717–5722.

[119] I. Laina *et al.*, "Concurrent segmentation and localization for tracking of surgical instruments," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2017, pp. 664–672.

[120] S. Bodenstedt *et al.*, "Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery," 2018. *arXiv:1805.02475*. [Online]. Available: https://arxiv.org/abs/1805.02475

[121] M. Allan *et al.*, "2017 robotic instrument segmentation challenge," 2019, *arXiv:1902.06426*. [Online]. Available: https://arxiv.org/abs/1902.06426

[122] L. Lejeune, J. Grossrieder, and R. Sznitman, "Iterative multi-path tracking for video and volume segmentation with sparse point supervision," *Med. Image Anal.*, vol. 50, pp. 65–81, Dec. 2018.

[123] F. Fuentes-Hurtado, A. Kadkhodamohammadi, E. Flouty, S. Barbarisi, I. Luengo, and D. Stoyanov, "EasyLabels: Weak labels for scene segmentation in laparoscopic videos," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 7, pp. 1247–1257, 2019.

[124] E. Flouty *et al.*, "CaDIS: Cataract dataset for image segmentation," (2019), *arXiv:1906.11586*. [Online]. Available: https://arxiv.org/abs/1906.11586

[125] E. Gibson *et al.*, "Deep residual networks for automatic segmentation of laparoscopic videos of the liver," *Proc. SPIE*, vol. 10135, p. 101351M, Mar. 2017.

[126] M. A. Everson *et al.*, "Artificial intelligence for the real-time classification of intrapapillary capillary loop patterns in the endoscopic diagnosis of early oesophageal squamous cell carcinoma: A proof-of-concept study," *United Eur. Gastroenterol. J.*, vol. 7, no. 2, pp. 297–306, 2019.

[127] M. Misawa *et al.*, "Artificial intelligence-assisted polyp detection for colonoscopy: Initial experience," *Gastroenterology*, vol. 154, no. 8, pp. 2027–2029, 2018.

[128] J. Bernal *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Feb. 2017.

[129] F. Mahmood, Z. Yang, R. Chen, D. Borders, W. Xu, and N. J. Durr, "Polyp segmentation and classification using predicted depth from monocular endoscopy," *Proc. SPIE*, vol. 10950, p. 1095011, Mar. 2019.

[130] S. Leonard *et al.*, "Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on *in vivo* clinical data," *IEEE Trans. Med. Imag.*, vol. 37, no. 10, pp. 2185–2195, May 2018.
[131] M. Visentini-Scarzanella, T. Sugiura, T. Kaneko, and S. Koto, "Deep monocular 3D reconstruction for assisted navigation in bronchoscopy," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 7, pp. 1089–1099, 2017.
[132] J. Wang, S. Song, H. Ren, C. M. Lim, and M. Q.-H. Meng, "Surgical instrument tracking by multiple monocular modules and a sensor fusion approach," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 2, pp. 629–639, Apr. 2019.
[133] X. Liu *et al.*, "Self-supervised learning for dense depth estimation in monocular endoscopy," 2019, *arXiv:1902.07766*. [Online]. Available: https://arxiv.org/abs/1806.09521
[134] N. L. Rodas, F. Barrera, and N. Padoy, "See it with your own eyes: Markerless mobile augmented reality for radiation awareness in the hybrid room," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 429–440, Feb. 2017.
[135] J. T. James, "A new, evidence-based estimate of patient harms associated with hospital care," *J. Patient Saf.*, vol. 9, no. 3, pp. 122–128, 2013.
[136] J. W. Suliburk, Q. M. Buck, and C. J. Pirko, "Analysis of human performance deficiencies associated with surgical adverse events," *JAMA Netw. Open*, vol. 2, no. 7, p. e198067, 2019.
[137] N. Padoy, "Vers une tour de contrôle des blocs opératoires?" in *Santé et Intelligence artificielle*, B. N. C. Villani, Ed. Paris, France: CNRS Editions, 2018.
[138] N. Padoy, "Machine and deep learning for workflow recognition during surgery," *Minimally Invasive Therapy Allied Technol.*, vol. 28, no. 2, pp. 82–90, 2019.
[139] P. Jannin and X. Morandi, "Surgical models for computer-assisted neurosurgery," *NeuroImage*, vol. 37, no. 3, pp. 783–791, 2007.
[140] T. Blum, N. Padoy, H. Feußner, and N. Navab, "Workflow mining for visualization and analysis of surgeries," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 3, no. 5, pp. 379–386, 2008.
[141] B. Gibaud *et al.*, "Toward a standard ontology of surgical process models," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 9, pp. 1397–1408, 2018.
[142] A. Haque *et al.*, "Towards vision-based smart hospitals: A system for tracking and monitoring hand hygiene compliance," in *Proc. Mach. Learn. Res. Mach. Learn. Healthcare Conf. (MLHC)*, 2017, pp. 75–87.
[143] A. J. Ma *et al.*, "Measuring patient mobility in the ICU using a novel noninvasive sensor," *Crit. care Med.*, vol. 45, no. 4, pp. 630–636, 2017.
[144] S. Yeung *et al.*, "A computer vision system for deep learning-based detection of patient mobilization activities in the icu," *NPJ Digit. Med.*, vol. 2, no. 1, p. 11, 2019.
[145] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and F. Li, "Viewpoint invariant 3D human pose estimation with recurrent error feedback," in *Proc. ECCV*, 2016, pp. 160–177.
[146] V. Belagiannis *et al.*, "Parsing human skeletons in an operating room," *Mach. Vis. Appl.*, vol. 27, no. 7, p. 1035–1046, Oct. 2016.
[147] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin, and N. Padoy, "A multi-view RGB-D approach for human pose estimation in operating rooms," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 363–372.
[148] I. Chakraborty, A. Elgammal, and R. S. Burd, "Video based activity recognition in trauma resuscitation," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–8.
[149] C. Lea, J. C. Facker, G. D. Hager, R. H. Taylor, and S. Saria, "3D sensing algorithms towards building an intelligent intensive care unit," in *Proc. AMIA Summits Translational Sci.*, 2013, p, 136.
[150] A. P. Twinanda, E. O. Alkan, A. Gangi, and M. de Mathelin, "Data-driven spatio-temporal RGBD feature encoding for action recognition in operating rooms," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 10, no. 6, pp. 737–747, Jun. 2015.
[151] A. P. Twinanda, P. Winata, A. Gangi, M. D. Mathelin, and N. Padoy, "Multi-stream deep architecture for surgical phase recognition on multi-view RGBD videos," in *Proc. M2CAI Workshop MICCAI*, 2016, pp. 1–8.
[152] V. Srivastav, T. Issenhuth, K. Abdolrahim, M. de Mathelin, A. Gangi, and N. Padoy, "MVOR: A multi-view RGB-D operating room dataset for 2D and 3D human pose estimation," in *Proc. MICCAI-LABELS*, 2018, pp. 1–10.
[153] T. Issenhuth, V. Srivastav, A. Gangi, and N. Padoy, "Face detection in the operating room: Comparison of state-of-the-art methods and a self-supervised approach," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 6, pp. 1049–1058, 2019.
[154] V. Srivastav, A. Gangi, and N. Padoy, "Privacy-preserving human pose estimation on low-resolution depth images," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2019.
[155] A. Malpani, C. Lea, C. C. G. Chen, and G. D. Hager, "System events: Readily accessible features for surgical phase detection," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, no. 6, pp. 1201–1209, 2016.
[156] Y. Gu *et al.*, "Language-based process phase detection in the trauma resuscitation," in *Proc. IEEE ICHI*, Aug. 2017, pp. 239–247.
[157] N. Navab, C. Hennersperger, B. Frisch, and B. Fuerst, "Personalized, relevance-based multimodal robotic imaging and augmented reality for computer assisted interventions," *Med. Image Anal.*, vol. 33, pp. 64–71, Oct. 2016.
[158] F. Lalys and P. Jannin, "Surgical process modelling: A review," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, no. 3, pp. 495–511, 2014.
[159] C. E. Reiley and G. D. Hager, "Using robots to train the surgeons of tomorrow," *IEEE Spectr.*, to be published.
[160] H. C. Lin, I. Shafran, T. E. Murphy, A. M. Okamura, D. D. Yuh, and G. D. Hager, "Automatic detection and segmentation of robot-assisted surgical motions," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2005, pp. 802–810.
[161] H. C. Lin, I. Shafran, D. Yuh, and G. D. Hager, "Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions," *Comput. Aided Surg.*, vol. 11, no. 5, pp. 220–230, 2006.
[162] F. Despinoy *et al.*, "Unsupervised trajectory segmentation for surgical gesture recognition in robotic training," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 6, pp. 1280–1291, Oct. 2015.
[163] B. Varadarajan, C. E. Reiley, H. Lin, S. Khudanpur, and G. D. Hager, "Data-derived models for segmentation with application to surgical assessment and training," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2009, pp. 426–434.
[164] N. Padoy and G. D. Hager, "Human-Machine Collaborative surgery using learned models," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, Shanghai, China, May 2011, pp. 5285–5292.
[165] J. D. Birkmeyer *et al.*, "Surgical skill and complication rates after bariatric surgery," *New England J. Med.*, vol. 369, no. 15, pp. 1434–1442, 2013.
[166] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. A. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 3, pp. 443–455, 2018.
[167] T. S. Kim *et al.*, "Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 6, pp. 1097–1105, 2019.
[168] J. Fotouhi *et al.*, "Interactive flying frustums (IFFs): Spatially aware surgical data visualization," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 6, pp. 913–922, 2019.
[169] S. Matinfar *et al.*, "Surgical soundtracks: Automatic acoustic augmentation of surgical procedures," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 9, pp. 1345–1355, Sep. 2018.
[170] O. Pauly, B. Diotte, P. Fallavollita, S. Weidert, E. Euler, and N. Navab, "Machine learning-based augmented reality for improved surgical scene understanding," *Comput. Med. Imag. Graph.*, vol. 41, pp. 55–60, Apr. 2015.
[171] A. Malpani, S. S. Vedula, H. C. Lin, G. D. Hager, and R. H. Taylor, "Real-time teaching cues for automated surgical coaching," 2017, *arXiv:1704.07436*. [Online]. Available: https://arxiv.org/abs/1704.07436
[172] A. Reiter, P. K. Allen, and T. Zhao, "Feature classification for tracking articulated surgical tools," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2012, pp. 592–600.
[173] X. Du *et al.*, "Articulated multi-instrument 2-D pose estimation using fully convolutional networks," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1276–1287, May 2018.
[174] D. Pakhomov, V. Premachandran, M. Allan, M. Azizian, and N. Navab, "Deep residual learning for instrument segmentation in robotic surgery," 2017, *arXiv:1703.08580*. [Online]. Available: https://arxiv.org/abs/1703.08580
[175] N. Rieke *et al.*, "Real-time localization of articulated surgical instruments in retinal microsurgery," *Med. Image Anal.*, vol. 34, pp. 82–100, Dec. 2016.
[176] T. Erickson, "Some problems with the notion of context-aware computing," *Commun. ACM*, vol. 45, no. 2, pp. 102–104, Feb. 2002.

## ABOUT THE AUTHORS

**Tom Vercauteren** graduated from Columbia University, New York, NY, USA, and the École Polytechnique, Palaiseau, France. He received the Ph.D. degree from INRIA, Sophia-Antipolis, France, in 2008.

From 2004 to 2014, he was with Mauna Kea Technologies, Paris, France, where he led the research and development team designing image computing solutions for the company's CE-marked and FDA-cleared optical biopsy device. From 2014 to 2018, he was an Associate Professor with University College London (UCL), London, U.K., where he was the Deputy Director of the Wellcome/EPSRC Centre for Interventional and Surgical Sciences from 2017 to 2018. He has been a Professor of interventional image computing with King's College London, London, since 2018, where he holds the Medtronic/Royal Academy of Engineering Research Chair in Machine Learning for Computer-Assisted Neurosurgery. His current research interests include translational medical image computing, machine learning, and interventional imaging devices with a specific interest in their development for surgery and interventional sciences.

**Mathias Unberath** received the B.Sc. degree *(summa cum laude)* in physics, the M.Sc. degree in optical technologies, and the Ph.D. degree in computer science from the Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany.

He was an ERASMUS Scholar with the University of Eastern Finland, Joensuu, Kuopio, Finland, and a DAAD Fellow with Stanford University, Stanford, CA, USA. He joined Johns Hopkins University, Baltimore, MD, USA, as a Postdoctoral Fellow, where he is currently an Assistant Research Professor with the Department of Computer Science, the Laboratory for Computational Sensing and Robotics, and the Malone Center for Engineering in Healthcare, Johns Hopkins University. His research interests include the intersection of computer vision, including augmented reality, machine learning, and medical physics, which have been recognized with multiple national and international awards and aim at pushing the boundaries of computer assistance in medical imaging and image-guided interventions.

**Nicolas Padoy** received the Maîtrise degree in computer science from the École Normale Supérieure de Lyon, Lyon, France, in 2003, the Diploma degree in computer science from Technische Universität München (TUM), Munich, Germany, in 2005, and the joint Ph.D. degree from the Chair for Computer Aided Medical Procedures, TUM, and the INRIA Group MAGRIT, Nancy, France.

From 2009 to 2011, he was a Postdoctoral Researcher and an Assistant Research Professor with the Laboratory for Computational Interactions and Robotics, Johns Hopkins University, Baltimore, MD, USA. He joined the Chair of Excellence, University of Strasbourg, Strasbourg, France, as an Assistant Professor, in 2012, where he is currently a Full Professor of computer science. He created and is currently leading the research group CAMMA on Computational Analysis and Modeling of Medical Activities, Strasbourg, that focuses on computer vision, activity recognition, artificial intelligence, and the applications thereof to surgical workflow analysis and human–machine cooperation during surgery.

**Nassir Navab** received the Ph.D. degree from INRIA, Sophia-Antipolis, France, and the University of Paris XI, Orsay, France.

He held a postdoctoral fellowship at the MIT Media Laboratory, Cambridge, MA, USA, for two years. He was with Siemens Corporate Research, Princeton, NJ, USA, from 1994 to 2003. He is currently a Full Professor and the Director of the Laboratory for Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany, and Johns Hopkins University, Baltimore, MD, USA. He is the inventor of 47 granted U.S. patents and more than 50 international ones. His current research interests include multimodal imaging, medical augmented reality, computer-assisted surgery, medical robotics, and machine learning.

Dr. Navab has acted as a member of the Board of Directors of the MICCAI Society from 2007 to 2012 and 2014 to 2017. He serves on the Steering Committee of the IEEE Symposium on Mixed and Augmented Reality (ISMAR) and the Information Processing in Computer-Assisted Interventions (IPCAI). In 2012, he was elected as a Fellow of the MICCAI Society. He received the Siemens Inventor of the Year Award in 2001, the SMIT Society Technology Award in 2010, and the 10 Years Lasting Impact Award of IEEE ISMAR in 2015.