The Impact of Machine Learning on 2D/3D Registration for Image-Guided Interventions: A Systematic Review and Perspective

Mathias Unberath*, Cong Gao, Yicheng Hu, Max Judish, Russell H Taylor, Mehran Armand and Robert Grupp

Advanced Robotics and Computationally Augmented Environments (ARCADE) Lab, Department of Computer Science, Johns Hopkins University, Baltimore, MD, United States

Image-based navigation is widely considered the next frontier of minimally invasive surgery. It is believed that image-based navigation will increase the access to reproducible, safe, and high-precision surgery as it may then be performed at acceptable costs and effort. This is because image-based techniques avoid the need of specialized equipment and seamlessly integrate with contemporary workflows. Furthermore, it is expected that image-based navigation techniques will play a major role in enabling mixed reality environments, as well as autonomous and robot-assisted workflows. A critical component of image guidance is 2D/3D registration, a technique to estimate the spatial relationships between 3D structures, e.g., preoperative volumetric imagery or models of surgical instruments, and 2D images thereof, such as intraoperative X-ray fluoroscopy or endoscopy. While image-based 2D/3D registration is a mature technique, its transition from the bench to the bedside has been restrained by well-known challenges, including brittleness with respect to optimization objective, hyperparameter selection, and initialization, difficulties in dealing with inconsistencies or multiple objects, and limited single-view performance. One reason these challenges persist today is that analytical solutions are likely inadequate considering the complexity, variability, and high-dimensionality of generic 2D/3D registration problems. The recent advent of machine learning-based approaches to imaging problems that, rather than specifying the desired functional mapping, approximate it using highly expressive parametric models holds promise for solving some of the notorious challenges in 2D/3D registration. In this manuscript, we review the impact of machine learning on 2D/3D registration to systematically summarize the recent advances made by introduction of this novel technology. Grounded in these insights, we then offer our perspective on the most pressing needs, significant open problems, and possible next steps.

Keywords: artificial intelligence, deep learning, surgical data science, image registration, computer-assisted interventions, robotic surgery, augmented reality

1 INTRODUCTION

1.1 Background

Advances in interventional imaging, including the miniaturization of high-resolution endoscopes and the increased availability of C-arm X-ray systems, have driven the development and adoption of minimally invasive alternatives to conventional, invasive and open surgical techniques across a wide variety of clinical specialities. While minimally invasive approaches are generally considered safe and effective, the indirect visualization of surgical instruments relative to anatomical structures complicates spatial cognition and the more confined room for maneuvers requires precise command of the surgical instruments. It is well known that due to the aforementioned challenges, among others, outcomes after minimally invasive surgery are positively correlated with technical proficiency, experience, and procedural volume of the operator (Birkmeyer et al., 2013; Pfandler et al., 2019; Hafezi-Nejad et al., 2020; Foley and Hsu, 2021). To mitigate the impact of experience on complication risk and outcomes, surgical navigation solutions that register specialized tools with 3D models of the anatomy using additional tracking hardware are now commercially available (Mezger et al., 2013; Ewurum et al., 2018). Surgical navigation promotes reproducibly good patient outcomes, and when combined with robotic assistance systems, may enable novel treatment options and improved techniques (van der List et al., 2016). Unfortunately, navigation systems are not widely adopted due to, among other things, high purchase price despite limited versatility, increased procedural time and cost, and potential for disruptions to surgical workflows due to line-of-sight occlusions or other system complications (Picard et al., 2014; Joskowicz and Hazan, 2016). While frustrations induced by workflow disruption affect every operator equally, the aforementioned limitations regarding cost particularly inhibit the adoption of surgical navigation systems in geographical areas with less specialized healthcare providers with lower volumes for any procedure; areas where routine use of navigation would perhaps be most impactful.

To mitigate the challenges of conventional surgical navigation systems that introduce dedicated tracking hardware and instrumentation as well as workflow alterations, the computer-assisted interventions community has contributed purely image-based alternatives to surgical navigation, e.g., (Nolte et al., 2000; Mirota et al., 2011; Leonard et al., 2016; Tucker et al., 2018). Image-based navigation techniques do not require specialized equipment but rely on traditional intra-operative imaging that enabled the minimally invasive technique in the first place. Therefore, these techniques do not introduce economic trade-offs. Further, because image-based navigation techniques are designed to seamlessly integrate into conventional surgical workflows, their use should—in theory—not cause frustration or prolonged procedure times (Vercauteren et al., 2019). A central component to many if not most image-guided navigation solutions is image-based 2D/3D registration, which estimates the spatial relationship between a 3D model of the scene (potentially including anatomy and instrumentation) and 2D interventional images thereof (Markelj et al., 2012; Liao et al., 2013). Two examples of using 2D/3D registration for intra-operative guidance are shown in **Figure 1**: Image-guidance of periacetabular osteotomy (left, and discussed in greater detail in **Section 3.3**) and robot-assisted femoroplasty (right). One may be tempted to assume that after several decades of research on this topic, image-based 2D/3D registration is a largely solved problem. While, indeed, analytical solutions now exist to precisely recover 2D/3D spatial relations under certain conditions (Markelj et al., 2012; Uneri et al., 2013; Gao et al., 2020b; Grupp et al., 2020b), several hard challenges prevail.

1.2 Problem Formulation

Generally speaking, in 2D/3D registration we are interested in finding the optimal geometric transformation that aligns a (typically pre-operative) 3D representation of objects or anatomy with (typically intra-operative) 2D observations thereof. For the purposes of this review, we will assume that the reduction in dimensionality originates from a projective, not an affine, transformation.
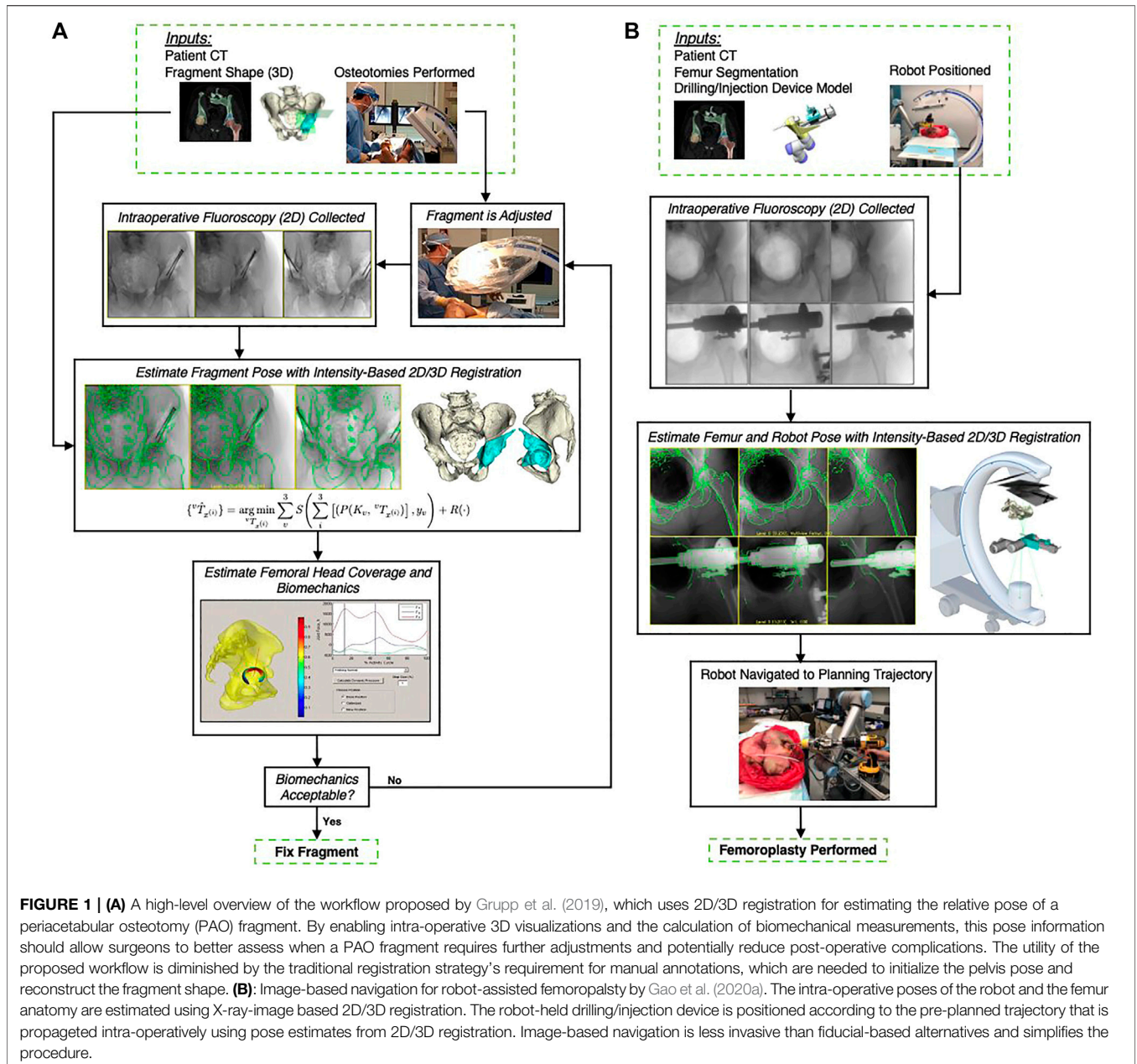
Given a set of 3D data x_i and 2D observations y_v, where subscripts i, v suggest that there may be multiple objects and multiple 2D observations, respectively, a generic way of writing the optimization problem for the common case of a single object but multiple views is:

$$\left\{\hat{K}_v, \ {}^v\hat{T}_x, \ \hat{\omega}_D\right\} = \arg\min_{K_v, \ {}^vT_x, \omega_D} \sum_v S\big(\big(P\big(K_v, \ {}^vT_x\big) \circ D_{\omega_D}\big)(x), y_v\big) + R(\cdot).$$

$$(1)$$

In **Eq. 1**, D_{ω_D} is a 3D non-rigid deformation model with parameters $\omega_D \in \mathbb{R}^{N_D}$, $P(K_v, {}^vT_x)$ is a rigid projection operation using a camera with intrinsic parameters $K_v \in \mathbb{R}^{3\times3}$ and pose ${}^vT_x \in SE(3)$ with respect to the 3D data x, and $S(\cdot, \cdot)$ is a cost function (in case of images often a similarity metric). We use the composite function $P \circ D(\cdot)$ to capture the variability, including their order, with which these operations may be applied. Finally, $R(\cdot)$ is a regularizing term that can act and/or depend on any combination of variables and parameters; its choice most often depends on the specific application since regularization can represent "common sense" or prior knowledge, which tends to vary with the problem domain.

2D/3D registration then amounts to estimating the $5 + 6 + N_D$ degrees of freedom (DoFs) for $\{\hat{K}_v, {}^v\hat{T}_x, \hat{\omega}_D\}$, respectively, that minimize the optimization objective. Clearly, there are special cases to **Eq. 1**, e.g., for rigid registration where only $\hat{K}_v$, ${}^v\hat{T}_x$ must be estimated and D_{ω_D} is known, or vice versa.

In traditional image-based 2D/3D registration, this optimization is usually performed iteratively where parameters are initialized with some values $\{K_v, {}^vT_x, \omega_D\}_0$ and then adjusted incrementally. The updates $\delta\{K_v, {}^vT_x, \omega_D\}$ are derived from **Eq. 1** using gradient-based or gradient-free methods, such as BFGS (Liu and Nocedal, 1989; Berger et al., 2016) or CMA-ES (Hansen et al., 2003) and BOBYQA (Powell, 2009), respectively. In certain cases when 2D and 3D data representations are not pixel or voxel grid-

FIGURE 1 | (A) A high-level overview of the workflow proposed by Grupp et al. (2019), which uses 2D/3D registration for estimating the relative pose of a periacetabular osteotomy (PAO) fragment. By enabling intra-operative 3D visualizations and the calculation of biomechanical measurements, this pose information should allow surgeons to better assess when a PAO fragment requires further adjustments and potentially reduce post-operative complications. The utility of the proposed workflow is diminished by the traditional registration strategy's requirement for manual annotations, which are needed to initialize the pelvis pose and reconstruct the fragment shape. **(B)**: Image-based navigation for robot-assisted femoropalsty by Gao et al. (2020a). The intra-operative poses of the robot and the femur anatomy are estimated using X-ray-image based 2D/3D registration. The robot-held drilling/injection device is positioned according to the pre-planned trajectory that is propageted intra-operatively using pose estimates from 2D/3D registration. Image-based navigation is less invasive than fiducial-based alternatives and simplifies the procedure.

based but sparse, e.g., keypoints, analytic solutions to **Eq. 1**, such as the perspective n point (PnP) algorithm (Lepetit et al., 2009), exist.

The above traditional approach to solving 2D/3D image registration has spawned solutions that precisely recover the desired geometric transformations under certain conditions (Markelj et al., 2012; Uneri et al., 2013; Gao et al., 2020b; Grupp et al., 2020b). Unfortunately, despite substantial efforts over the past decades, 2D/3D registration is not yet enjoying wide popularity as a workhorse component in image-based navigation platforms at the bedside. Rather, it is shackled to the bench top because several hard open challenges inhibit its widespread adoption. They include:

- **Narrow capture range of similarity metrics**: Conventional intensity-based methods mostly use hand-crafted similarity metrics between y_v and its current best estimate $\hat{y}_v = (P(K_v, {}^vT_x) \circ D_{\omega_D})(x)$fx1 as loss function. Common choices for the similarity metric are Normalized Cross Correlation (NCC), gradient information (Berger et al., 2016), or Mutual Information (MI) (Maes et al., 1997). While these metrics are positively correlated with pose differences when the perturbations in $\{\hat{K}_v, {}^vT_x, \hat{\omega}_D\}$ are small, they are generally non-convex and fail to accurately represent pose offsets when perturbations are large. Thus, without proper initialization, the optimization is prone to get stuck in local minima, returning wrong registration

results. The initial estimate of the target parameters $\{\hat{K}_v, {}^v\hat{T}_x, \hat{\omega}_D\}$ must hence be close enough to the true solution in order for the optimization to converge to the global minima. Estimating good initial parameters is commonly achieved using some manual interaction, which is cumbersome and time consuming, or—in research papers—neglected all-together. The magnitude by which the initial parameter guesses may be incorrect for the downstream algorithm to still produce a successful registration is referred to as the capture range (Markelj et al., 2012; Esteban et al., 2019). Its magnitude depends, among other things, on the similarity function as well as the optimizer, and it can be stated quantitatively as the mean target registration error (mTRE) between 3D keypoints at initialization[1].

The resulting challenges are two-fold: On the one hand, it is important to develop robust and automated initialization strategies for the existing image-based 2D/3D registration algorithms to succeed. On the other hand, there is interest in and opportunity for the development of better similarity metrics that better capture the evolution of image (dis)similarity. Doing so is challenging, however, because of the complexity of the task including various contrast mechanisms, imaging geometries, and inconsistencies.

- **Ambiguity**: The aforementioned complexity also leads to registration ambiguity, which is most pronounced in single-view registration (Otake et al., 2013; Uneri et al., 2013). Because the spatial information along the projection line is collapsed onto the imaging plane, it is hard to precisely recover the information in the projective direction. A well-known example is the difficulty of accurately estimating the depth of 3D scene from the camera center using a single 2D image. These challenges already exist for rigid 2D/3D registration and are further exacerbated in rigid plus deformable registration settings.
- **High dimensional optimization problems**: Even in the simplest case, 2D/3D registration describes a non-convex optimization problem with at least six DoFs to describe a rigid body transform. In the context of deformable 2D/3D registration, the high dimensional parameter ω_D that describes the 3D deformation drastically increases the optimization search space. However, since the information within the 2D and 3D images remains constant, the optimization problem may easily become ill-posed. Although statistical modeling techniques exist to limit the parameter search space, the registration accuracy and sensitivity to key features remain an area of concern (Zhu et al., 2021).
- **Verification and uncertainty**: As a central component of image-based surgical navigation platforms, 2D/3D registration supplies critical information to enable precise manipulation of anatomy. To enable users to

assess risk and make better decisions, there is a strong desire for registration algorithms to verify the resulting geometric parameters or supply uncertainty estimates. Perhaps the most straightforward way of verifying a registration result is to visually inspect the 2D overlay of the projected 3D data—this approach, however, is neither quantitative nor does it scale since it is based on human intervention.

These open problems can be largely attributed to the variability in the problem settings (e.g., regarding image appearance and contrast mechanisms, pose variability, ...) that cannot easily be handled algorithmically because the desired properties cannot be formalized explicitly. Machine learning methods, including deep neural networks (NNs), have enjoyed a growing popularity across a variety of image analysis problems (Vercauteren et al., 2019), precisely because they do not require explicit definitions of complex functional mappings. Rather, they optimize parametric functions, such as convolutional NNs (CNNs), on training data such that the model learns to approximate the desired mapping between input and output variables. As such, they provide opportunities to supersede heuristic components of traditional registration pipelines with learning-based alternatives that were optimized for the same task on much larger amounts of data. This allows us to expand **Eq. 1** into:

$$
\begin{aligned}
\{\hat{K}_v,\ {}^v\hat{T}_x, \hat{\omega}_D\} \\
= \arg\min_{K_v, {}^vT_x, \omega_D} \sum_v S^{\theta_S}\Big(\big(P(K_v,\ {}^vT_x) \circ D_{\omega_D}^{\theta_D}\big)\big(G_x^{\theta_x}(x)\big), G_y^{\theta_y}(y_v)\Big) \\
+ R^{\theta_R}(\cdot),
\end{aligned}
\tag{2}
$$

where we have introduced parameters θ to several components of the objective function to indicate that they may now be machine learning models, such as CNNs. Similarly, the registration may not rely on the original 3D and 2D data itself but some higher-level representation thereof, e.g., anatomical landmarks, that are generated using some learned function $G^\theta(\cdot)$.

In this manuscript, first we summarize a systematic review of the recent literature on machine learning-based techniques for image-based 2D/3D registration, and explain how they relate to **Eq. 2**. Based on those observations, we identify the impact that the introduction of contemporary machine learning methodology has had on 2D/3D registration for image-guided interventions. Concurrently, we identify open challenges and contribute our perspective on possible solutions.

2 SYSTEMATIC REVIEW

2.1 Search Methodology
The aim of the systematic review is to survey those machine learning-enhanced 2D/3D registration methods in which the 3D data and 2D observations thereof are related through one or multiple perspective projections (and potentially some non-rigid deformation). This scenario arises, for example, in the registration between 3D CT and 2D X-ray, 3D magnetic resonance angiography (MRA) and 2D digital subtraction

---

[1]It should be noted that the definition of capture range is by no means standardized or even similar across papers, which we will also comment on in **Section 3**.

angiography (DSA), or 3D anatomical models and 2D endoscopy images. 3D/3D registration methods (such as between 3D CT and intra-operative CBCT) or 2D/3D slice-to-volume registration (as it arises, among others, in ultrasound to CT/MR registration) are beyond the scope of this review. Because we are primarily interested in surveying the impact of contemporary machine learning techniques, such as deep CNNs, on 2D/3D registration, we limit our analysis to records that appeared after January 2012, which pre-dates the onset of the ongoing surge of interest in learning-based image processing (Krizhevsky et al., 2012).

To this end, we conducted a systematic literature review in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) method (Moher et al., 2009) (cf. **Figure 2**). We searched PubMed, embase, Business Source Ultimate, and Compendex to find articles pertinent to machine learning for 2D/3D image registration. The following search terms were used to screen titles, abstracts, and keywords of all available records from January 2012 through February 2021:

("2D3D registration" OR "2D 3D registration" OR "3D2D registration" OR "3D 2D registration" OR "2D/3D registration" OR "3D/2D registration" OR "2D-3D registration" OR "3D-2D registration" OR "two-dimensional/three-dimensional registration" OR "three-dimensional/two-dimensional registration") AND ("learning" OR "training" OR "testing" OR "trained" OR "tested")
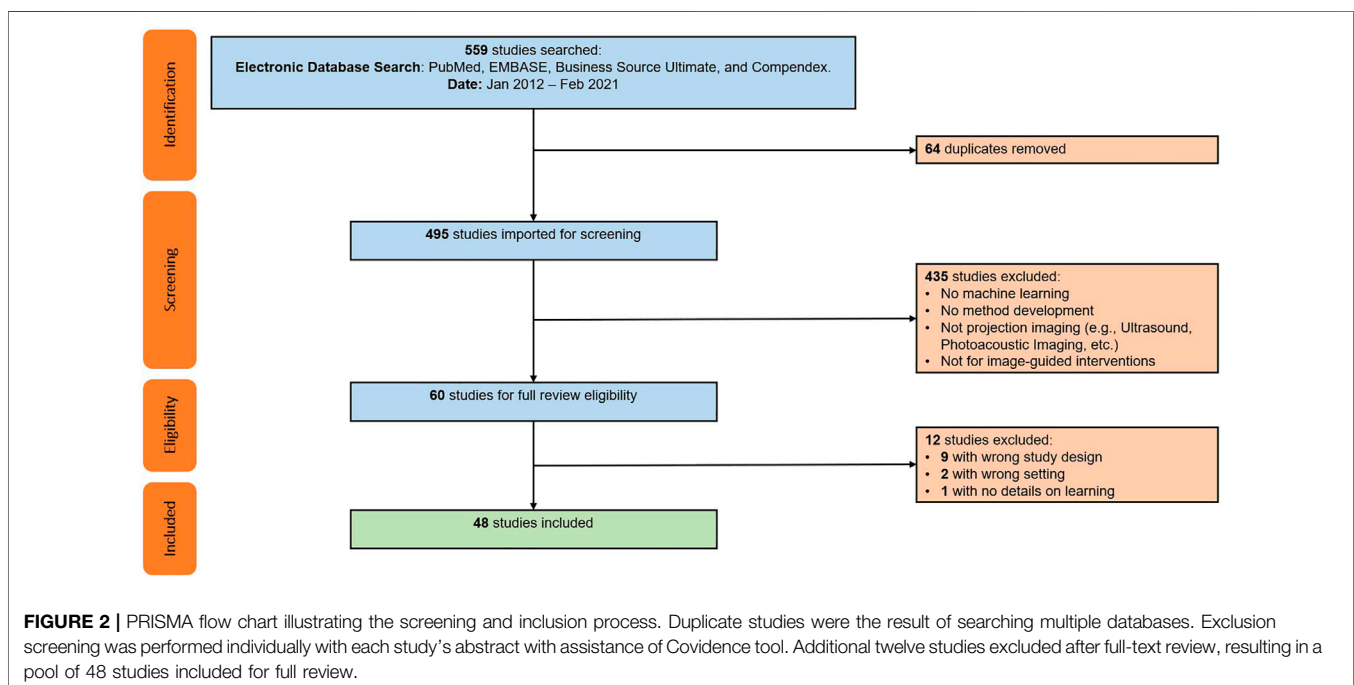
The initial search resulted in 559 records, and after removal of duplicates, 495 unique studies were included for screening. From those, 447 were excluded because they either did not describe a machine learning-based method for 2D/3D registration, or considered a slice-to-volume registration problem (e.g., as in ultrasound or magnetic resonance imaging). The remaining 48 articles were included for an in-depth full text review, analysis, and data extraction, which was performed by five of the authors (MU, CG, MJ, YH, and RG). Initially, every reviewer analyzed five articles to develop and refine the data extraction template and coding approach. The final template involved the extraction of the following information: 1) A brief summary of the method including the key contribution; 2) modalities and registration phase (including the 3D modality, 2D modality, the registration goal, whether the method requires manual interactions, whether the method is anatomy or patient-specific, and the clinical speciality), 3) the spatial transformation to be recovered (including the number of objects to be registered, the number of views used for registration, and the transformation model used), 4) information on the machine learning model and training setup (including the explicit machine learning technique, the approach to training data curation as well as to data labeling and supervision, and the application of domain generalization or adaptation techniques), 5) the evaluation strategy (including the data source used for evaluation as well as its annotation, the metrics and techniques used for quantitative and qualitative assessment, and most importantly the deterioration of performance in presence of domain shift), and finally, 6) a more subjective summary of concerns with respect to the experimental or methodological approach or the assumptions made in the design or evaluation of the method.

Every one of the 48 articles was analyzed and coded by at least two of the five authors and one author (MU) merged the individual reports into a final consensus document.

2.2 Limitations
Despite our efforts to broaden the search terms regarding 2D/3D registration, we acknowledge that the list may not be exhaustive. Newer or less popular terminology, such as

FIGURE 2 | PRISMA flow chart illustrating the screening and inclusion process. Duplicate studies were the result of searching multiple databases. Exclusion screening was performed individually with each study's abstract with assistance of Covidence tool. Additional twelve studies excluded after full-text review, resulting in a pool of 48 studies included for full review.

"pose regression", were not used. We also did not use modality specific terms, such as "video to CT registration", which may have excluded some manuscripts that focus on endoscopy or other RGB camera-based modalities. The search included terms like "learning", "training", or "testing" as per our interest in machine learning methods for 2D/3D image registration. This search may have excluded some studies that do not explicitly characterize their work as machine or deep "learning" and do not describe their training and testing approach, neither in their title nor abstract. The terminology used in the search may also have resulted in the exclusion of relevant work from the general computer vision literature. Finally, the review is limited to published manuscripts. Publication bias may have resulted in the exclusion of works relevant to this review.

2.3 Concise Summary of the Overall Trends

We first summarize the general application domain and problem setting of the 48 included papers and then review the role that machine learning plays in those applications. The specific characteristics of all included papers are summarized in **Tables 1** and **2**, respectively. We state the number of papers either in the running text or in parentheses.

The vast majority of papers (34) considers the 2D/3D registration between X-ray images and CT or cone-beam CT (CBCT) volumes, with the registration of X-ray images and 3D object models being a distant second (10). Other modality combinations included 2D RGB to 3D CT (2) and 3D MR (1), or did not specify the 3D modality (1). The clinical applications that motivate the development of those methods include orthopedics (19), with a focus on pelvis and spine, angiography (9), and radiation therapy (7), e.g., for tracking of lung or liver tumors, and cephalometry (4). We observe that eleven methods are explicitly concerned with finding a good initial parameter set to begin optimization, while 37 papers (also) describe approaches to achieve high fidelity estimates of the true geometric parameters. Further, four methods consider verification of the registration result. Resulting from the clinical task and registration phase distributions, most methods only perform rigid alignment (36), while nine methods consider non-rigid registration only and two approaches address both, rigid and non-rigid registration. To solve the alignment problem, 38 papers relied on a single 2D view, seven approaches used multiple views onto the same constant 3D scene, and two methods assumed the same view but used multiple images of a temporally dynamic 3D scene. Perhaps the most striking observation is that all but three (45) included studies only consider the registration of a single object. Two other studies that deal with multiple objects, however, are limited to object detection (Doerr et al., 2020) and inpainting (Esfandiari et al., 2021), respectively, and do not report registration results. The remaining study (Grupp et al., 2020c) performs a 2D segmentation of multiple bones, but does not apply any additional learning to perform the registration. While there are three methods that do, in fact, describe the 2D/3D registration of multiple objects, i.e., vertebral bodies (Varnavas et al., 2013; 2015b) and knee anatomy (Wang et al., 2020), the individual

registrations are solved independently which inhibits information sharing to ease the optimization problem.

The focus of this review is the impact that machine learning has had on the contemporary state of 2D/3D registration and we will briefly introduce the five main themes that we identified here and then discuss them in greater detail in subsequent sections. From a high-level perspective of abstracting 2D/3D registration problems, they follow the flow of acquiring *Data*, fitting *Model* and solving the *Objective*. The five themes which we categorize logically aim at improving certain aspects of this flow. The themes are:

- Contextualization (**Section 2.4**): Instead of relying solely on the images themselves, the 14 methods in this theme use machine learning algorithms to extract semantic information from the 2D or 3D data, including landmark or object detection, semantic segmentation, or data quality classification. Doing so enables automatic initialization techniques, sophisticated regularizers, as well as techniques that handle inconsistencies between 2D and 3D data.
- Representation learning (**Section 2.5**): Principal component analysis (PCA), among other techniques, are a common way to reduce the dimensionality of highly complex data—in this case, rigid and non-rigid geometric transformations. Twelve papers used representation learning techniques as part of the registration pipeline.
- Similarity modeling (**Section 2.5**): Optimization-based image registration techniques conventionally rely on image similarity metrics that ideally should capture appearance differences due to both large and very fine scale geometric misalignment. Ten studies describe learning-based approaches to improve on similarity quantification.
- Direct parameter regression (**Section 2.7**): In contrast to iterative methods, direct parameter regression techniques seek to infer the correct geometric parameters for 2D/3D alignment (either absolute with respect to a canonical 3D coordinate frame, or relative between a source and a target coordinate frame) directly from the 2D observation. A total of 22 manuscripts reported such approaches for either rigid or non-rigid registration.
- Verification (**Section 2.8**): Four studies used machine learning-based techniques to assess whether the estimated geometric parameters should be considered reliable.

High-level depictions of these themes are shown in **Figure 3** and the respective sections below provide details for each, along with references to the individual studies. Refer to **Table 3** for a summary of themes attributed to each included study.

2.4 Contextualization

Studies summarized in this theme use machine learning techniques to increase the information available to the 2D/3D registration problem by extracting semantic information from the 2D or 3D data (Lin and Winey, 2012; Varnavas et al., 2013, 2015b; Bier et al., 2018; Chen et al., 2018; Bier et al., 2019; Luo et al., 2019; Yang and Chen, 2019; Grupp et al., 2020c; Doerr et al., 2020;

TABLE 1 | Parameters defining the registration problems described in the studies included for review. Registration purpose refers to the registration stage being addressed, such as initialization (init.), precise retrieval of geometric parameters (fine regis.), or others.

Study ID	3D Modality	2D Modality	Regis. Action	Regis. Purpose	Speciality	Rigid/Non-rigid	No. Views
Brost et al. (2012)	Catheter Model	X-ray	Pre-proc	fine regis	Catheter	R	1
Lin and Winey (2012)	CT	X-ray	Pre-proc	fine regis	Radiotherapy	R	2
Chou and Pizer (2013)	CT (3D + t)	X-ray	Deformable regr	fine regis	Lung tumors	NR	1
Chou et al. (2013)	CBCT, s-NST	X-ray	Pose updates	fine regis	Head-and-neck, lungs	Both	R: 4, NR: 2
Varnavas et al. (2013)	CT	X-ray	Pre-proc	init., verif	Spine/Vertebrae	R	1
Chou and Pizer (2014)	CT	X-ray	Deformable regr	fine regis	Lung	NR	1
Mitrovi et al. (2014)	CT	X-ray	N/A	verif	Spine/Vertebra	R	1
Zhao et al. (2014)	CT	X-ray	Deformable regr	fine regis	Abdomen	NR	1
Baka et al. (2015)	CTA	XA	Pre-proc	fine regis	Coronary artery	Both	1 (temporal)
Mitrovi et al. (2015)	DSA	DSA	Pose regr., updates	init., fine regis	Angiography	R	1
Varnavas et al. (2015a)	CT	X-ray	Pre-proc	init., verif	Spine	R	1
Wu et al. (2015)	mdl	X-ray	Pose regr	init	Knee	R	1
Miao et al. (2016b)	Tool Model	X-ray	Pose regr	fine regis	Implants	R	1
Miao et al. (2016a)	Implant and TEE mdl	X-ray	Pose regr	fine regis	Implants/TEE transducer	R	1
Tang and Scalzo (2016)	MRA	DSA	cost func	fine regis	Angiography	R	1
Wu et al. (2016)	CT	X-ray	N/A	verif	Skull/Head	R	2
Hou et al. (2017)	CT	X-ray	Pose regr	init	Thorax	R	1
Pei et al. (2017)	CT	X-ray	Deformable regr	fine regis	Skull	NR	1
Xie et al. (2017)	CTA	X-ray	Pose regr	fine regis	Angiography	R	1
Bier et al. (2018)	CT	X-ray	Pre-proc	init	Pelvis	R	1
Chen et al. (2018)	CT	X-ray	Pre-proc	feat. extract	Spine	R	1
Hou et al. (2018)	CT	X-ray	Pose regr	init	Thorax	R	1
Miao et al. (2018)	CBCT	X-ray	Pose regr., updates	fine regis	Spine	R	2
Toth et al. (2018)	Left ventr mdl	X-ray	Pose regr., updates	fine regis	Heart	R	1
Zhang et al. (2018)	PCA Deformation Field	X-ray	Deformable regr	init., fine regis	Skull	NR	1 (temporal)
Zheng et al. (2018)	CT, implant mdl	X-ray	Domain adaptation	Pose regr., updates	Spine and TEE transducer	R	1
Bier et al. (2019)	CT	X–ray	Pre-proc	init	Pelvis	R	1
Foote et al. (2019)	CT	X-ray	Deformable regr	fine regis	Lung	NR	1
Guan et al. (2019)	Vasc mdl. (CT)	DSA	Pose regr	fine regis	Cardiovascular	R	1
Liao et al. (2019)	CT or CBCT	X-ray	cost func	fine regis	Thorax	R	1
Luo et al. (2019)	CT	Broncho - scopy	Pre-proc	fine regis	Bronchoscopy	R	1
Schaffert et al. (2019)	CBCT	X-ray	cost func	fine regis	Spine	R	1
Yang and Chen (2019)	CT/MRI	Stereo RGB	Pre-proc	feat. extract	Head/face	R	2
Doerr et al. (2020)	N/A	X-ray	Pre-proc	init	Spine, pedicle screws	R	1
Francois et al. (2020)	MRI	Laparo - scopy	Pre-proc, cost func	fine regis	Uterus	N/A	N/A
Gao et al. (2020c)	CT	X-ray	Pose regr	Pose updates	Pelvis	R	1
Grupp et al. (2020c)	CT	X-ray	Pose regr	Pose updates	Pelvis	R	1
Gu et al. (2020)	CT	X-ray	cost func	Pose updates	Pelvis	R	1
Guan et al. (2020)	Aorta mdl	DSA (synth)	Deformable regr	fine regis	Cardiovascular	NR	1
Karner et al. (2020)	CT/MR	RGB face img	Pre-proc	fine regis	Face	R	1
Li et al. (2020)	CBCT	X-ray	Deformable regr	fine regis	Skull	NR	1
Neumann et al. (2020)	MRA	DSA	cost func	fine regis	Angiography	R	1
Schaffert et al. (2020a)	CBCT	X-ray	cost func	fine regis	Spine	R	1
Schaffert et al. (2020b)	CBCT	X-ray	cost func	fine regis	Spine, head	R	1–2
Wang et al. (2020)	CT	Bi - plane Fluoroscopy	Pre-proc	init	Knee	R	2

(Continued on following page)

TABLE 1 | (*Continued*) Parameters defining the registration problems described in the studies included for review. Registration purpose refers to the registration stage being addressed, such as initialization (init.), precise retrieval of geometric parameters (fine regis.), or others.

Study ID	3D Modality	2D Modality	Regis. Action	Regis. Purpose	Speciality	Rigid/Non-rigid	No. Views
Xiangqian et al. (2020)	CT	X-ray	Pose regr	fine regis	Pelvis	R	1
Zhang et al. (2020)	CT	X-ray	Post-proc	fine regis	Liver tumors	NR	20
Esfandiari et al. (2021)	CT	X-ray	Pre-proc	fine regis	Spine, pedicle screws	R	2

Francois et al., 2020; Karner et al., 2020; Wang et al., 2020; Esfandiari et al., 2021).

Using the notation of **Eq. 2**, these methods specify $G_y^{\theta_y}(y)$, $G_x^{\theta_x}(x)$, or $R^{\theta_R}(\cdot)$ although not all methods are necessarily integrated in the iterative optimization procedure. Perhaps the most prevalent approach here is the detection of anatomical landmarks on 2D images (Bier et al., 2018, 2019; Yang and Chen, 2019; Grupp et al., 2020c; Wang et al., 2020) to define correspondences with the respective 3D locations, which allows for either, explicit determination of $\hat{y}_v = \{K_v, {}^v T_x\}$ using PnP (Bier et al., 2019, 2018) or stereo-reconstruction following 3D-3D matching (Yang and Chen, 2019; Wang et al., 2020), or for the introduction of soft re-projection constraints as a regularizing term (Grupp et al., 2020c). Another way of benefiting initialization through contextualization is object detection (Lin and Winey, 2012; Varnavas et al., 2013, 2015b; Chen et al., 2018; Doerr et al., 2020). While all methods for landmark detection rely on deep CNNs, object detection achieved satisfactory results already using less complex learning models, i.e., templates in (Lin and Winey, 2012), PCA over object contours in (Chen et al., 2018), and the Generalized Hough Transform in (Varnavas et al., 2013; 2015b). The drawback, however, is that most of these less complex approaches require patient-specific training since the models are unable to generalize beyond a single shape. Doerr et al. (2020) describe a deep learning-based alternative, where the Fast R-CNN object detector is re-trained to return bounding boxes of 30 different screw types in varied poses.

A complementary trend is the identification of image regions (Francois et al., 2020; Esfandiari et al., 2021) or whole images (Luo et al., 2019) that should not contribute to the optimization problem because of inconsistency. Francois et al. (2020) use a U-net-like fully convolutional NN (FCN) to segment occluding contours of the uterus to reject regions while Luo et al. (2019) identify and reject poor quality frames in bronchoscopy. Esfandiari et al. (2021) consider mismatch introduced by intra-operative instrumentation. They contribute a non-blind image inpainting method using a FCN that seeks to restore the background anatomy after image regions corresponding to instruments were identified.

It is undeniable that the introduction of machine learning to contextualize 2D and 3D data enables novel techniques that quite substantially expand the tools one may rely on when designing a 2D/3D registration algorithm, and as such, are likely to become impactful. However, a general trend that we observed in most of these studies was that the impact of the contextualization component on the downstream registration task was not, in fact, evaluated. For example, while (Bier et al., 2019) report

quantitative results on real data of cadaveric specimens, it remains unclear whether the performance would be sufficient to actually initialize an image similarity-based 2D/3D registration algorithm. There are, of course, positive examples including (Varnavas et al., 2015b; Luo et al., 2019; Grupp et al., 2020c) that demonstrate the benefit of contextualization on overall pipeline performance—Empirical demonstrations should strongly be preferred over arguments from authority.

2.5 Representation Learning

As highlighted in **Section 1.2**, 2D/3D registration, especially in deformable scenarios, suffers from high dimensional parameter spaces and changes in any one of those parameters are not easily resolved due to limited information which creates ambiguity. In our review we found that unsupervised representation learning techniques are a widely adopted technique to reduce the dimensionality of the parameter space while introducing implicit regularization by confining possible solutions to the principal modes of variation across population- or patient-level observations. We identified 12 studies that propose such techniques or use them as part of the registration pipeline (Brost et al., 2012; Lin and Winey, 2012; Chou and Pizer, 2013, 2014; Chou et al., 2013; Zhao et al., 2014; Baka et al., 2015; Pei et al., 2017; Chen et al., 2018; Zhang et al., 2018; Foote et al., 2019; Li et al., 2020; Zhang et al., 2020).

PCA is by far the most prevalent method for representation learning and is used in all but one study. This specific study, however, used by far the most views $v = 20$ for initial estimation of a low resolution vector field, which was then regularized by projection onto a deep learning-based population model (Zhang et al., 2020). We found that methods designed for cephalometry were distinct from all other approaches as their primary goal is not generally 2D/3D registration, but 3D reconstruction of the skull given a 2D X-ray. Among the papers included in this review, this problem is often formulated as the deformable 2D/3D registration between a lateral X-ray image of the skull and a 3D atlas using a PCA deformation model, the principal components ω_D of which are estimated via a prior set of 3D/3D registrations (Pei et al., 2017; Zhang et al., 2018; Li et al., 2020). Consequently, these methods rely on population-level models and are thus different from methods used for radiation therapy (Chou et al., 2013; Chou and Pizer, 2013, 2014; Zhao et al., 2014; Foote et al., 2019) and angiography (Brost et al., 2012; Baka et al., 2015), which rely on patient-specific models that are built pre- and intra-operatively, respectively. It is worth mentioning that, while most methods rely

TABLE 2 | A summary of the training and testing details for each study reviewed.

| Study ID | Network | Training | | | | | Testing |
	Architecture	Data	Domain Transfer	Object Number	Anatomy Specificity	Technique	Data
Brost et al. (2012)	PCA	Real	Segmentation	Single	N/A	Unsupervised	Real
Lin and Winey (2012)	N/A	N/A	N/A	Single	Patient	Unsupervised	N/A
Chou et al. (2013)	Linear Regression	Synthetic	N/A	Single	Anatomy	Supervised	Synthetic
Chou and Pizer (2013)	PCA	Synthetic and Real	N/A	Single	Anatomy	Supervised	Synthetic and Real
Varnavas et al. (2013)	GHT	Real	N/A	Single	Anatomy and Patient	Unsupervised	Real
Chou and Pizer (2014)	Random Forest	Synthetic	Gaussian Normalization	Single	Patient	Supervised	Synthetic and Real
Mitrovi et al. (2014)	N/A	Synthetic	N/A	Single	Patient	Supervised	Real
Zhao et al. (2014)	Linear Regression	Synthetic	N/A	Single	Patient	Supervised	Real and Synthetic
Baka et al. (2015)	N/A	Real	N/A	Single	Patient	N/A	Real
Mitrovi et al. (2015)	N/A	N/A	N/A	Single	Patient	N/A	Synthetic and Real
Varnavas et al. (2015a)	GHT	Synthetic	N/A	Single	Anatomy and Patient	Unsupervised	Real
Wu et al. (2015)	PCA	Synthetic	N/A	Single	Anatomy	Unsupervised	Synthetic and Real
Miao et al. (2016a)	Siamese CNNs	Synthetic	Realism Tuning	Single	Anatomy	Supervised	Real
Miao et al. (2016a)	Siamese CNNs	Synthetic	N/A	Single	Anatomy	Supervised	Real
Tang and Scalzo (2016)	Spectral Regression	Real	Abstraction	Single	Anatomy	Supervised	Real
Wu et al. (2016)	MLP	Real	N/A	Single	N/A	Supervised	Real
Hou et al. (2017)	CaffeNet	Synthetic	N/A	Single	Anatomy	Supervised	Synthetic
Pei et al. (2017)	CNNs	Real	N/A	Single	Anatomy	Supervised	Synthetic
Xie et al. (2017)	CNNs	Synthetic	N/A	Single	Anatomy	Supervised	Synthetic
Bier et al. (2018)	Sequential CNNs	Synthetic	Domain Generalization	Single	Anatomy	Supervised	Synthetic
Chen et al. (2018)	PCA	Synthetic	N/A	Single	N/A	Supervised	Real
Hou et al. (2018)	N/A	Synthetic	N/A	Single	Anatomy	Supervised	Synthetic
Miao et al. (2018)	Dilated CNNs	Synthetic	N/A	Single	Anatomy	Supervised	Real
Toth et al. (2018)	CNNs	Synthetic	N/A	Single	Anatomy	Supervised	Synthetic
Zhang et al. (2018)	VGG	Synthetic	N/A	Single	Anatomy	Supervised	Synthetic
Zheng et al. (2018)	DA Module	Pairwise Synthetic and Real	Domain Adaptation	Single	Anatomy	Unsupervised	Real
Bier et al. (2019)	Sequential CNNs	Synthetic	Domain Generalization	Single	Anatomy	Supervised	Synthetic and Cadaver
Foote et al. (2019)	DenseNet	Synthetic	Equalization methods	Single	Patient	Supervised	Synthetic
Guan et al. (2019)	CNNs	Synthetic	Retrain on patient data	Single	Anatomy	Supervised	Synthetic
Liao et al. (2019)	U-Net	Real	N/A	Single	Anatomy	Supervised	Real
Luo et al. (2019)	Instance Learning	Real	N/A	Single	Anatomy	Supervised	Real
Schaffert et al. (2019)	PointNet	Real	N/A	Single	Anatomy	Supervised	Real
Yang and Chen (2019)	Stacked Hourglass	Real	N/A	Single	Anatomy	Supervised	Synthetic
Doerr et al. (2020)	Fast R-CNN	Synthetic	N/A	Multiple	Patient	Supervised	Synthetic
Francois et al. (2020)	U-Net	Real	N/A	Single	Anatomy	Supervised	Real
Gao et al. (2020c)	Spatial Transformer	Synthetic	Realistic Simulation	Single	Anatomy	Supervised	Synthetic and Real
Grupp et al. (2020c)	U-Net	Real	N/A	Single	Anatomy	Supervised	Real
Gu et al. (2020)	DenseNet	Synthetic	Realistic Simulation	Single	Anatomy	Supervised	Synthetic and Real
Guan et al. (2020)	CNNs	Synthetic	Refine on patient data	Single	Anatomy	Supervised	Synthetic
Karner et al. (2020)	Face-to-3D	N/A	N/A	Single	Anatomy	N/A	Real
Li et al. (2020)	ResNet	Synthetic	N/A	Single	Anatomy	Self-supervised	Synthetic and Real
Neumann et al. (2020)	Siamese ResNet	Synthetic	N/A	Single	Anatomy	Supervised	Synthetic
Schaffert et al. (2020a)	PointNet	Real	N/A	Single	Anatomy	Supervised	Real

TABLE 2 | (*Continued*) A summary of the training and testing details for each study reviewed.

Study ID	Network	Training					Testing
	Architecture	Data	Domain Transfer	Object Number	Anatomy Specificity	Technique	Data
Schaffert et al. (2020b)	FlowNet-S	Real	N/A	Single	Anatomy	Supervised	Real
Wang et al. (2020)	VGGs	Real	Domain Adaptation	Multiple	Anatomy	Supervised	Real
Xiangqian et al. (2020)	GoogleNet	Synthetic	Histogram Matching	Single	Anatomy	Supervised	Real
Zhang et al. (2020)	U-Net	Real	N/A	Single	Anatomy	Supervised	Real
Esfandiari et al. (2021)	PConvS	Synthetic	Heavy Augmentation	Single	Anatomy	Supervised	Synthetic
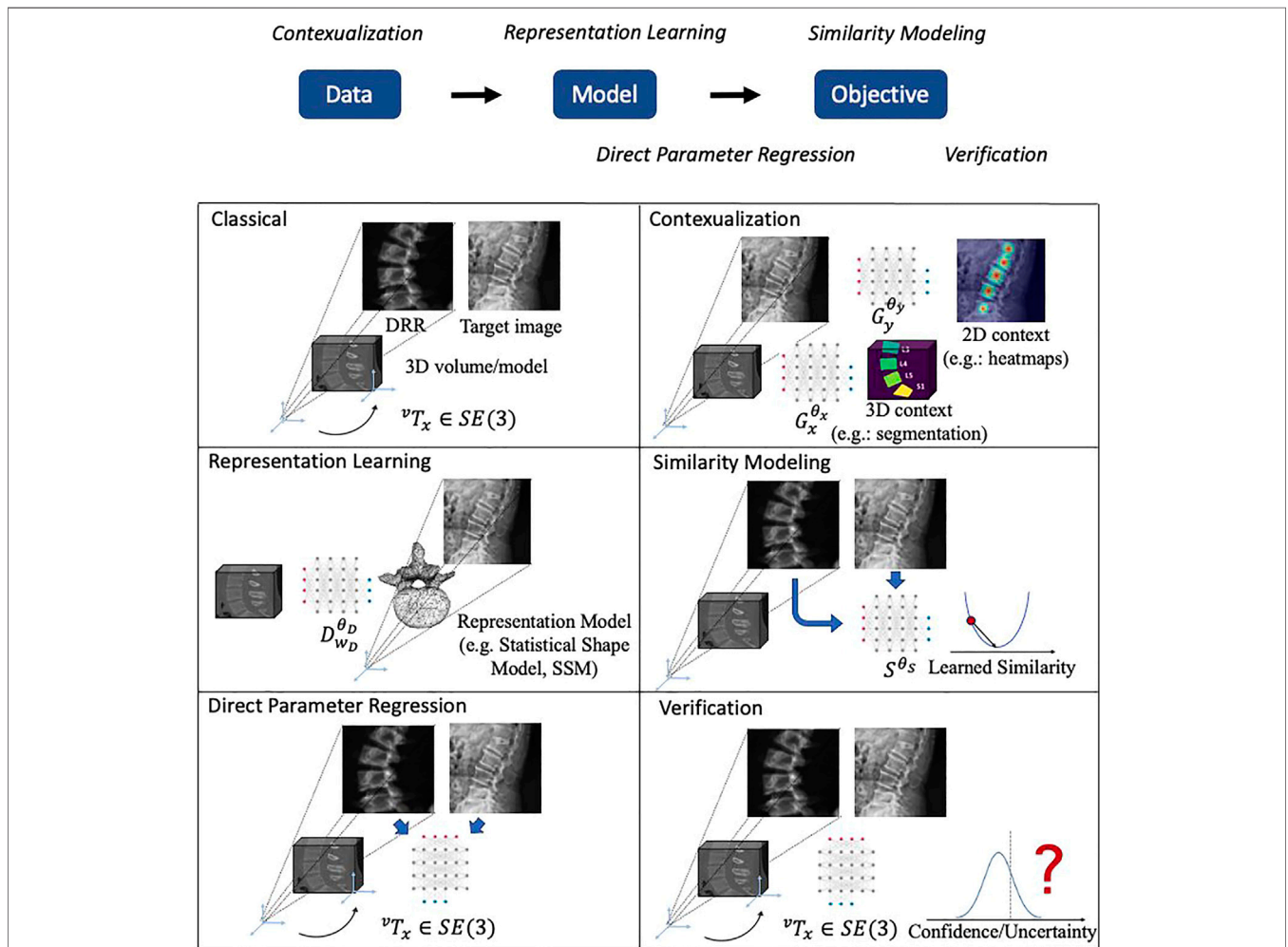
FIGURE 3 | Illustrations of the main themes of machine learning in 2D/3D registration. The logic relationships of these themes are shown on top. We use a spine CT volume and a spine X-ray image as an example to show the generic 2D/3D projection geometry. Machine learning models are represented with a neural network icon. Key labels and parameters are presented and map to **Eqs. 1**, **2**.

on PCA to condense deformable motion parametrizations, it has also been found useful to identify and focus on the primary modes of variation in rigid registration (Brost et al., 2012).

We note that most studies in this theme do not consider rigid alignment prior to deformable parameter estimation, e.g., by assuming perfectly lateral radiographs of the skull for cephalometry or perfectly known imaging geometry in radiation therapy. Except for few exceptional cases with highly specialized instrumentation, for example (Chou and Pizer, 2013) that relied on on-board CBCT imaging,

TABLE 3 | A summary of each study's relation with the five themes of Contextualization, Representation learning, Direct parameter regression, Similarity modeling, and Verification.

Study ID	Contextualization	Representation learning	Direct parameter regression	Similarity modeling	Verification
Brost et al. (2012)	■	■			
Lin and Winey (2012)	■				
Chou and Pizer (2013)		■	■		
Chou et al. (2013)		■	■		
Varnavas et al. (2013)	■				■
Chou and Pizer (2014)		■	■		
Mitrovi et al. (2014)					■
Zhao et al. (2014)		■	■		
Baka et al. (2015)	■	■			
Mitrovi et al. (2015)			■		
Varnavas et al. (2015a)	■				■
Wu et al. (2015)			■		
Miao et al. (2016b)			■		
Miao et al. (2016a)			■		
Tang and Scalzo (2016)				■	
Wu et al. (2016)					■
Hou et al. (2017)			■		
Pei et al. (2017)		■	■		
Xie et al. (2017)			■		
Bier et al. (2018)	■				
Chen et al. (2018)	■	■			
Hou et al. (2018)			■		
Miao et al. (2018)			■		
Toth et al. (2018)			■		
Zhang et al. (2018)		■	■		
Zheng et al. (2018)			■		
Bier et al. (2019)	■				
Foote et al. (2019)		■	■		
Guan et al. (2019)			■		
Liao et al. (2019)				■	
Luo et al. (2019)	■				
Schaffert et al. (2019)				■	
Yang and Chen (2019)	■				
Doerr et al. (2020)	■				
Francois et al. (2020)	■			■	
Gao et al. (2020c)			■	■	
Grupp et al. (2020c)	■			■	
Gu et al. (2020)				■	
Guan et al. (2020)			■		
Karner et al. (2020)	■				
Li et al. (2020)		■	■		
Neumann et al. (2020)				■	
Schaffert et al. (2020a)				■	
Schaffert et al. (2020b)				■	
Wang et al. (2020)	■				
Xiangqian et al. (2020)			■		
Zhang et al. (2020)		■			
Esfandiari et al. (2021)	■				

assumptions around rigid alignment seem to be unjustified, which may suggest that the performance estimates need to be interpreted with care. This is further emphasized by the fact that many studies are only evaluated on synthetic data (which we fear may have sometimes been generated by sampling the PCA model also used for registration, introducing inverse crime) and may not have paired 3D data for extensive quantitative evaluation.

2.6 Similarity Modeling

As we established in **Section 1.2** for optimization-based 2D/3D registration algorithms, the cost function—or similarity metric—$S(\cdot, \cdot)$ is among the most important components since it will determine parameter updates. It is well known that most commonly used metrics fail to accurately represent the distances in geometric parameter space that generate the mismatch between the current observations. It is thus not surprising that ten studies describe methods to better model and quantify the similarity between the source and target images to increase the capture range, and thus, the likelihood of registration success (Tang and Scalzo, 2016; Liao et al., 2019; Schaffert et al., 2019, 2020b; Schaffert et al., 2020a; Gao et al., 2020c; Grupp et al., 2020c; Francois et al., 2020; Gu et al., 2020; Neumann et al., 2020).

Some studies propose novel image similarity functions $S^{\theta_S}(\cdot, \cdot)$ that, analogous to traditional similarity metrics, accept as input the source and target image and return a scalar or vector that is related to the mismatch in parameter space (Francois et al., 2020; Gu et al., 2020; Tang and Scalzo, 2016; Neumann et al., 2020; Grupp et al., 2020c; Gao et al., 2020b). Among those, two methods rely on regularization: Grupp et al. (2020c) detect anatomical landmarks to expand an analytic similarity function with landmark-reprojection constraints to enhance the capture range of an intensity-based strategy, while Francois et al. (2020) segment occluding contours to constrain similarity evaluation to salient regions. The other four methods use machine learning models to approximate a geometric parameter distance function based on the input images. To this end, Gu et al. (2020) and Gao et al. (2020b) estimate the geodesic in Riemannian tangent space between the source and target camera poses, which in an ideal case results in a convex similarity function. Tang and Scalzo (2016) learn a more expressive feature descriptor to better quantify the mismatch in vasculature registration, and for a similar application, Neumann et al. (2020) regress the disparity between corresponding points in source and target images to quantify image dissimilarity. Both methods rely on concepts that are limited to sparse objects, such as vessels. Different to image-based similarity metrics, four studies describe methods for keypoint matching to compute image similarity (Schaffert et al., 2019, 2020a,b; Liao et al., 2019). To this end, Liao et al. (2019) train a network to establish keypoint correspondences between the source and target images. Because the geometric parameters of the source image are known, the unknown target parameters can be recovered relative to the source image using PnP-like methods. Finally, a series of three papers (Schaffert et al., 2019, 2020a,b) characterizes a learning-based method to

adaptively weight point correspondences that are established to quantify the degree of misalignment.

As methods in this theme have primarily focused on expanding the capture range of contemporary similarity metrics, the potential shortcomings of other components of the registration pipeline, such as the optimizers, remain unaffected. While we introduced (Gao et al., 2020b) in the context of similarity learning, the method also describes a fully differentiable 2D/3D registration pipeline that addresses optimization aspects. This enables both, end-to-end learning of $G_y^{\theta_y}(y), G_x^{\theta_x}(x)$ and/or $S^{\theta_S}(\cdot, \cdot)$ during training as well as analytic gradient-based optimization using backpropagation during application. Further, we found that most emphasis was given on increasing the capture range of the registration pipeline and very little, if any, attention is paid to increasing the resolution and precision of these metrics. Especially in single-view registration scenarios, which we have identified to be most prevalent, it is well known that certain DoFs cannot be resolved with high accuracy. Therefore, developing methods that not only increase the capture range but also the precision of 2D/3D registration pipelines should be of high priority.

2.7 Direct Parameter Regression

So far and especially in the context of **Section 2.6**, 2D/3D registration was motivated as an optimization-based process that compares the source image, generated using the current geometric parameter estimate, with the desired target y_v using some cost function. However, this problem can also be formulated in the context of regression learning, where a machine learning algorithm directly predicts the desired geometric parameters K_v, ^vT_x, and/or ω_D from y_v, or from both y_v and $\hat{y}_v$. Such methods partially or completely skip the step of precise modeling of image formation or similarity, and instead build up the knowledge in a data-driven manner. We have identified 22 studies that describe methods for direct parameter regression (Chou and Pizer, 2013, 2014; Chou et al., 2013; Zhao et al., 2014; Mitrovi et al., 2015; Wu et al., 2015; Miao et al., 2016a; Miao et al., 2016b; Hou et al., 2017; Pei et al., 2017; Xie et al., 2017; Hou et al., 2018; Miao et al., 2018; Toth et al., 2018; Zhang et al., 2018; Zheng et al., 2018; Guan et al., 2019, 2020; Foote et al., 2019; Gao et al., 2020c; Li et al., 2020; Xiangqian et al., 2020).

Relying on parameter regression solely based on the target image y_v is particularly prevalent for radiation therapy, where the main application is the regression of the principal components of a patient-specific PCA motion model (Zhao et al., 2014; Chou et al., 2013; Chou and Pizer, 2014, 2013; Foote et al., 2019). The importance of regression learning is primarily attributed to the substantially decreased run-time that enables close to real-time tumor tracking in 3D. Methods directed at cephalometry (Li et al., 2020; Zhang et al., 2018; Pei et al., 2017) are identical in methodology to the radiation therapy methods. As noted in **Section 2.5**, most methods here limit themselves to shape estimation and assume that a global rigid alignment is either performed prior or unnecessary. The remaining 14 methods consider rigid parameter regression, and we differentiate methods that infer pose directly from the target y_v (Xiangqian

et al., 2020; Wu et al., 2015; Hou et al., 2017, 2018; Xie et al., 2017; Guan et al., 2019, 2020), and methods that process both y_v and $\hat{y}_v$ (Toth et al., 2018; Miao et al., 2016a,b, 2018; Zheng et al., 2018; Gao et al., 2020c; Mitrovi et al., 2015), and therefore, can run iteratively. Methods that rely on the target image only are relatively straight-forward and generally train a standard feed-forward CNN architecture to regress pose on large datasets comprising of multiple independent objects or anatomies. While the simplicity of these approaches is appealing, a general concern is that poses are absolute which requires the definition of a canonical 3D coordinate system. This challenge is mitigated for applications that consider an instrument or tool, because such a canonical system can be readily defined; however, establishing this reference frame is considerably more effortful for patient anatomy and may require (group-wise) 3D/3D registrations. Unfortunately, none of the studies reviewed here describes a dedicated effort to establish such a canonical reference frame, suggesting that those methods will eventually hit a performance ceiling that is determined by the misalignment within the reference coordinate systems in the training data. Methods that regress pose between y_v and $\hat{y}_v$ avoid the aforementioned concern because poses are relative rather than absolute, however similarly to conventional techniques, they require an initialization. There is some flexibility in how the information from source and target images is combined. (Miao et al., 2016a,b, 2018; Zheng et al., 2018). use the image residual between y_v and $\hat{y}_v$ in a region of interest around the projected tool model. Because only a small part of the target image is used, the initialization must be sufficiently close. Toth et al. (2018) extract features using a CNN from both source and target image independently and then concatenate them for pose regression using fully connected layers. Rather than regressing pose directly, Gao et al. (2020c) introduce a fully differentiable 2D/3D pipeline and compare $G_y^{\theta_y}(y)$ and $G_{y'}^{\theta_y}(y')$ using a simple L_2 distance as the similarity function $S(\cdot, \cdot)$. The parameters θ_y and θ_x are then optimized using a double backward pass on the computational graph, such that the gradient $\partial S/\partial \, {}^v T_x$ aligns with the geodesic between the current pose estimate and the desired target pose.

Pose regression directly from images is appealing because it may result in substantially faster convergence, potentially with a single forward pass of a CNN. We found that all methods included in this review are limited to registration of a single object and it remains unclear how these methods would apply to multiple objects, in part, because of the combinatorial explosion of relative poses. Further, methods that solely rely on the target image never involve neither the 3D data nor the source images created from it. This may be problematic, because it is unclear how these methods would verify that this specific 2D/3D registration data satisfy, among other things, the canonical coordinate frame assumption.

2.8 Verification

We identified four studies that leverage machine learning techniques for verifying whether a registration process produced satisfactory geometric parameter estimates (Wu et al., 2016; Mitrovi et al., 2014; Varnavas et al., 2013, 2015b).

An interesting observation is that none of the included methods make use of the resulting images or overlays, but rather rely on low-dimensional data. Varnavas et al. (2013, 2015b), for example, rely on the cost function value and the relative poses of multiple objects that are registered independently as input to a support vector machine classifier. Similarly, Wu et al. (2016) train a shallow NN to classify registration success based on hand-crafted features of the objective function surface around the registration estimate. Finally, Mitrovi et al. (2014) compare a registration estimate to known local minima and thresholds to determine success/failure, which worked well but may be limited in practice as the approach seems to assume knowledge of the correct solution.

Some studies included in this theme stand out, in that they are certainly mature and were demonstrated to work well on comparably large amounts of real clinical data, such as (Varnavas et al., 2015b) and (Wu et al., 2016); unfortunately however, these methods are not general purpose as they rely on the registration of multiple objects and the availability of two orthogonal views, respectively. Compared to the other four themes and maybe even in general, there has been very little emphasis on and innovation in the development of more robust and general purpose methods for the verification of 2D/3D registration results, which we perceive to be a regrettable omission.

3 PERSPECTIVE

The introduction of machine learning methodology to the 2D/3D registration workflow was partly motivated by persistent challenges, which were not yet satisfactorily addressed by heuristic and purely algorithmic approaches. Upon review of the recent literature in **Section 2**, perhaps the most pressing question is: Has machine learning resolved any of those open problems? We begin our discussion using the categorization of **Section 1.2**:

- **Narrow capture range of similarity metrics**: We identified many methods that quantitatively demonstrate increased capture range of 2D/3D registration pipelines. The means to accomplishing this, however, are diverse. Several methods describe innovative learning-based similarity metrics that better reflect distances in geometric parameter space, while other studies present (semi-)global initialization or regularization techniques which may rely on contextual data. The demonstrated improvements are generally significant, suggesting huge potential for machine learning in regards to this particular challenge. Contextualization, such as landmark detection and segmentation, can mimic user input to enable novel paradigms for initialization while also providing a clear interface for human-computer interaction (Amrehn et al., 2017; Du et al., 2019; Zapaishchykova et al., 2021). Similarity modeling using learning-based techniques—potentially combined with contextual information—is a similarly powerful concept that finds

broad application also in slice-to-volume registration, e.g., for ultrasound to MRI (Hu et al., 2018), which was beyond this review.

- **Ambiguity**: While several methods report an overall improved registration performance when using their novel, machine learning-enhanced algorithms, we did not register any method with a particular focus on reducing ambiguity. Because performance is usually reported as a summary statistic over multiple DoFs and instances, it is unclear to what extent performance increases should be attributed to 1) registering individual instances more precisely (which would suggest reduced ambiguity) or 2) succeeding more often (which would rather emphasize the importance of the capture range).

- **High dimensional optimization problems**: We found that representation learning techniques, currently dominated by PCA, are a clearly established tool to reduce the dimensionality of deformable 2D/3D registration problems, and may even be useful for rigid alignment. In those lower dimensional spaces, e.g., the six parameters of a rigid transformations or the principal components of a PCA model, direct pose regression from the target and/or source images is a clearly established line of research. These approaches supersede optimization with a few forward passes of a machine learning model, which makes them comparably fast. This is particularly appealing for traditionally time critical applications, such as tumor tracking in radiation therapy. Complementary approaches that seek to enable fully differentiable 2D/3D registration pipelines for end-to-end training and analytical optimization of geometric parameters, such as (Gao et al., 2020c; Shetty et al., 2021), combine elements of optimization and inference. While these methods are in a development stage, their flexibility may prove a great strength in developing solutions that meet clinical needs.

Certainly, some of the above ideas will become increasingly important in the quest to accelerate and improve 2D/3D registration pipelines. This is because image-based navigation techniques (Sugano, 2003; Hummel et al., 2008; Tucker et al., 2018) as well as visual servoing of surgical robots (Yi et al., 2018; Gao et al., 2019; Unberath et al., 2019) will also require high-precision 2D/3D registration at video frame-rates.

- **Verification and uncertainty**: Compared to the other challenges and themes, very few studies used machine learning to benefit verification of registration results. The studies that did, however, reported promising performance even with rather simple machine learning techniques on low dimensional data, i.e., cost function properties rather than images themselves. Quantifying uncertainty in registration slowly emerges as a research thrust in 2D/2D and 3D/3D registration (Pluim et al., 2016; Sinha et al., 2019). It is our firm belief that the first generally applicable methods for confidence assignment and uncertainty estimation in 2D/3D registration will become trend-setting due to the nature of the clinical applications that 2D/3D registration enables.

Despite this positive prospect on the utility of machine learning for 2D/3D registration, we noted certain trends and recurring shortcomings in our review that we will discuss next. As was done before, we either specify the number of studies satisfying a specific condition in the text or state it in parentheses.

3.1 Preserving Improvements Under Domain Shift From Training to Deployment

An omnipresent concern in the development of machine learning-based components for 2D/3D registration, but everywhere really, is the availability of or access to large amounts of relevant data. In some cases, the data problem amounts to a simple opportunity cost, e.g., for automation of manually performed tasks such as landmark detection. It should be noted, however, that even for this "simple" case to succeed, many conditions must be met including ethical review board approval, digital medicine infrastructure, and methods to reliably annotate the data. In many other—from a research perspective perhaps more exciting—cases, this retrospective data collection paradigm is infeasible because the task to be performed with a machine learning algorithm is not currently performed in clinical practice. The more obvious examples are visual servoing of novel robotic surgery platforms (Gao et al., 2019) or robotic imaging paradigms that alter how data is acquired (Zaech et al., 2019; Thies et al., 2020). Despite the fact that most studies included in this review address use-cases that fall under the "opportunity cost" category, we found that only 16 out of the 48 studies used real clinical or cadaveric data to train the machine learning algorithms. All remaining papers relied on synthetic data, namely digitally reconstructed radiographs (DRRs), that were simulated from 3D CT scans to either replace or supplement (small) real datasets.

Training on synthetic data has clear advantages because large datasets and corresponding annotations can be generated with relatively little effort. In addition, rigid pose and deformation parameters are perfectly known by design thus creating an unbiased learning target. Contemporary deep learning-based techniques enable the mapping of very complex functions directly from high-dimensional input data at the cost of heavily over-parameterized models that require as much data as possible to learn sensible associations. These unrelenting requirements, especially with respect to annotation, are not easily met with clinical data collection. Indeed, of the 32 studies that describe deep learning-based methods, 24 trained on synthetic data (seven trained on real data, and one did not train at all but used a pre-trained network). It is evident that data synthesis is an important idea that enables research on creative approaches that contribute to the advancement of 2D/3D registration.

Unfortunately, there are also substantial drawbacks of synthetic data training. Trained machine learning algorithms approximate the target function only on a compact domain (Zhou, 2020), and their behaviour outside this domain is unspecified. Because synthesized data is unlikely to capture all characteristics of data acquired using real systems and from real patients, the domains defined by the synthetic data used for

training and the real data used during application will not, or only partially, overlap. This phenomenon is known as domain shift. Therefore, applying a synthetic data-trained machine learning model to real data is likely to result in substantially deteriorated performance (Unberath et al., 2018, 2019). While this problem exists for all machine learning algorithms, it is particularly prevalent for modern deep learning algorithms as they operate on high-resolution images directly, where mismatch in characteristics (such as noise, contrast, …) is most pronounced.[2] Indeed, among the seven studies that trained deep CNNs on synthetic data and evaluated in a way that allowed for comparisons between synthetic and real data performance, we found quite substantial performance drops (Miao et al., 2018; Bier et al., 2019; Gao et al., 2020c; Doerr et al., 2020; Gu et al., 2020; Guan et al., 2020; Li et al., 2020). Worse, three studies used different evaluation metrics in synthetic and real experiments so that comparison was not possible (Miao et al., 2016a; Toth et al., 2018; Esfandiari et al., 2021), and perhaps worst, ten studies that trained on synthetic data never even tested (meaningfully) on real data (Hou et al., 2017, 2018; Pei et al., 2017; Xie et al., 2017; Bier et al., 2018; Foote et al., 2019; Guan et al., 2019; Yang and Chen, 2019; Neumann et al., 2020; Zhang et al., 2020). To mitigate the negative impact of domain shift, there is growing interest in domain adaptation and generalization techniques. Several methods do, in fact, already incorporate some of those techniques (Zheng et al. (2018); Wang et al. (2020) use domain adaptation to align feature representations of real and synthetic data: Gu et al. (2020); Esfandiari et al. (2021); Foote et al. (2019) use heavy pixel-level transformations that approximate domain randomization, and Xiangqian et al. (2020); Bier et al. (2018, 2019); Gu et al. (2020); Gao et al. (2020c); Miao et al. (2016a) rely on realistic synthesis to reduce domain shift, e.g., using open-source physics-based DRR engines (Unberath et al., 2018, 2019)). However, as we have outlined above their impact is not yet strongly felt. For the new and exciting 2D/3D registration pipelines reviewed here to impact image-based navigation, we must develop novel techniques to increase the robustness under domain shift to preserve the method's level of performance when transferring from training to deployment domain.

3.2 Experimental Design, Reporting, and Reproducibility

Quantitatively evaluating registration performance is clearly important. The error metrics that were used included the standard registration pose error (commonly separated into translational and rotational DoFs), keypoint distances (Chou and Pizer, 2014; Zhang et al., 2020), and the mean target registration error (mTRE) in 3D, and varied other metrics in 2D, such as reprojection distances (Bier et al., 2019), segmentation or overlap DICE score (Zhang et al., 2020), or

contour differences (Chen et al., 2018). Other metrics that are not uniquely attributable to a domain include the registration capture range (Schaffert et al., 2019; Esfandiari et al., 2021) and the registration success rate (Varnavas et al., 2013; Mitrovi et al., 2014; Miao et al., 2016b; Yang and Chen, 2019; Schaffert et al., 2020a). There are no standard routines to define the successful registrations.

The wealth of evaluation strategies and metrics can, in part, be attributed to the fact that different clinical applications necessitate different conditions to be met. For example, in 2D/3D deformable registration for tumor tracking during radiation therapy, accurately recovering the 3D tumor shape and position (quantified well using, e.g., the DICE score of true and estimated 3D position over time) is much more relevant than a Euclidean distance between the deformation field parameters, which would describe irrelevant errors far from the region of interest. We very clearly advocate *for* the use of task-specific evaluation metrics, since ultimately those metrics are the ones that will distinguish success from failure in the specific clinical application. However, we also believe that the lack of universally accepted reporting guidelines, error metrics, and datasets is a severe shortcoming that has unfortunate consequences, such as a high risk of duplicated efforts and non-interpretable performance reporting. We understand the most pressing needs to be:

- **Standardizing evaluation metrics**: An issue that appears to have become more prevalent with the introduction of machine learning methods that are developed and trained on synthetic data is the lack of substantial results on clinically relevant, real data. While experimental conditions, including ground truth targets, are perfectly known for simulation, they are much harder to obtain for clinical or cadaveric data. A common approach to dealing with this situation is to provide detailed quantification of mTRE, registration accuracy, etc. on synthesized data where the algorithm will perform well (cf. **Section 3.1**) while only providing much simpler, less informative, and sometimes purely qualitative metrics for real data experiments. Clearly, this practice is undesirable because 1) synthetic data experiments now cannot serve as a baseline (since they use different metrics and are thus incomparable), and 2) the true quantities of interest remain unknown (for example, a 3D mTRE is more informative than a 2D reprojection TRE since it can adequately resolve depth).

While it is evident that not all evaluation paradigms that are easily available on synthetic data can be readily transferred to clinical data, the reverse is not true. If real data experiments require simplified evaluation protocols because some gold standard quantities cannot be assessed, then these simplified approaches should at a minimum also be implemented on synthetic data to further complement the evaluation. While this approach may still leave some questions regarding real data performance unanswered, it will at least provide reliable information to assess the deterioration from sandbox to real life.

---

[2]Non-deep learning techniques usually operate on lower dimensional data that is abstracted from the images (such as cost function values (Wu et al., 2016) or centerlines (Tang and Scalzo, 2016)) such that domain shift is handled elsewhere in the pipeline, e.g., in a segmentation algorithm.
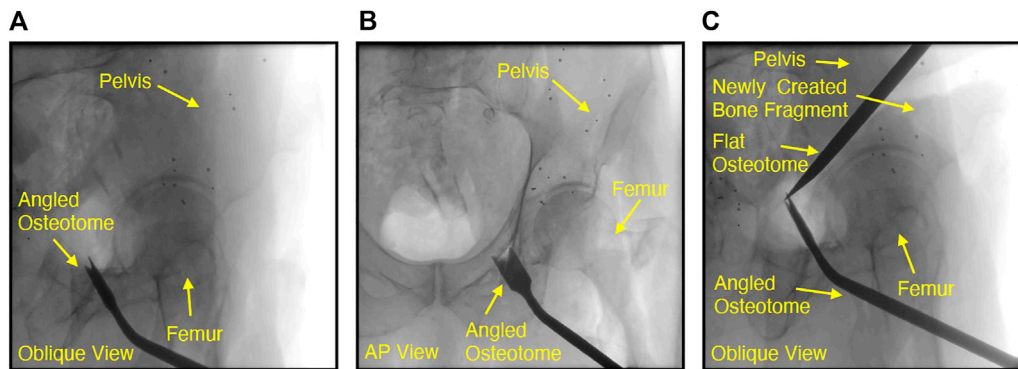
FIGURE 4 | Fluoroscopic images of chiseling performed during a cadaveric periacetabular osteotomy procedure. A cut along the ischium of the pelvis is shown by the oblqiue view in **(A)**. Due to the difficulty with manually interpreting lateral orientations of the tool in an oblique view, the very next fluoroscopy frame was collected at an approximate anterior-posterior (AP) view, shown in **(B)**. Although the osteotome location was confirmed by changing view points, the process of adjusting the C-arm by such a large offset can potentially increase operative time or cause the clinician to lose some context in the previous oblique view. Another oblique view is shown in (c), where the flat osteotome is used to complete the posterior cut, resulting in the creation of two bone fragments from the pelvis. In **(C)**, the angled osteotome was left in the field of view and used as a visual aid for navigating the flat chisel and completing the osteotomy.

- **Reporting problem difficulty**: A confounding factor that needs to be considered even when consistent metrics are being used is the fact that different datasets are likely to posit 2D/3D registration problems of varied difficulty. For example, evaluation on synthetic data may include data sampled from a broad range of viewpoints or deformations that are approximately uniformly distributed. Real data, on the other hand, is not uniformly sampled from such a distribution, but rather, will be clustered around certain viewpoints (see (Grupp et al., 2020c) for a visualization of viewpoints used during a cadaveric surgery vs the synthetic data for the same machine learning task in (Bier et al., 2019). Because those viewpoints are optimized for human interpretation, they are likely to contain more easily interpretable information which suggests that algorithmic assessment is also more likely to succeed. Then, in those cases, even if the quantitative metrics reported across those two datasets would suggest similar performance, in reality there is degradation due to evaluation on a simpler problem.

One way to address this challenge would be to attempt the harmonization of problem complexity by recreating the real dataset synthetically. Another, perhaps more feasible approach would be to develop reporting guidelines that allow for a more precise quantification of problem complexity, e.g., by more carefully describing the variation in the respective datasets.

- **Enabling reproduction**: The recent interest in deep learning has brought about a transformational move towards open science, where large parts of the community share their source code publicly and demonstrate the performance of their solutions on public benchmarks, which allows for fair comparisons. While a too strong focus on "winning benchmarks" is certainly detrimental to the creativity of novel approaches, we feel that the lack of public benchmarks for 2D/3D registration

and related problems is perhaps an even greater worry. In addition to the use of different metrics for validation purposes discussed above, studies may even use different definitions of the same quantity (such as the capture range, that is defined, e.g., using the decision boundary of an SVM classifier in (Esfandiari et al., 2021) and using mTRE in (Schaffert et al., 2020a). Further, most code-bases and more importantly datasets are kept private which inhibits reproduction, since re-implementation is particularly prone to biased conclusions. Consequently, creating public datasets with well-defined and standardized testing conditions should be a continued and reinforced effort, and wherever possible, the release of source code should be considered.

To this end, our group has previously released a relatively large dataset of CTs and >350 X-rays across multiple viewpoints of six cadaveric specimens prior to undergoing periacetabular osteotomy (cf. **Figure 4**) (Grupp et al., 2020). Further, we have made available the core registration components of our intensity-based framework, xReg, as well as our open-source framework for fast and physics-based synthesis of DRRs from CT, DeepDRR, which may allow for the creation of semi-realistic but very large and precisely controlled data. Increasing the rate with which we share code and data will likely result in research that truly advances the contemporary capabilities since the baselines are transparent and much more clearly defined.

These challenges clearly restrict research, but even more dramatically inhibit translational efforts, simply because we cannot reliably understand whether a specific 2D/3D registration problem should be considered "solved" or what the open problems are. Finding answers to the posited questions will become especially important as 2D/3D registration technology matures and is integrated in image-based navigation solutions that are subject to regulatory approval. Then, compelling evidence will need to be provided

on potential sources and extent of algorithmic bias as well as reliable estimates of real world performance (US Food and Drug Administration, 2021). Adopting good habits around standardized reporting will certainly be a good first step to translate research successes into patient outcomes through productization.

3.3 Registration of Multiple Objects, Compound or Non-rigid Motion, and Presence of Foreign Objects

Image-guidance systems often need to process and report information regarding the relative poses between several objects of interest, such as bones, organs and surgical instruments. We found that only three studies in our review register multiple objects through learning-based approaches (Varnavas et al., 2013, 2015b; Wang et al., 2020), and moreover, these studies independently registered single objects in order to obtain relative poses. Although combining the results of two distinct single object registrations is perhaps the most straight-forward approach for obtaining the relative pose between two objects, it fails to leverage the combined information of their relative pose during any optimization process, which could have potentially yielded a less-challenging search landscape. As an example, consider the case of two adjacent objects whose independent, single view, depth estimates are erroneous in opposite directions. The relative pose computed from these two independent poses will have an exacerbated translation error resulting from the compounding effect of the independent depth errors. A multiple object registration strategy could alternatively parameterize the problem to *simultaneously* solve for the pose of one object with respect to the imaging device and also for the relative pose between the objects. This approach partially compounds motion of the objects during registration and completely eliminates the possibility of conflicting depth estimates.

Despite the small number of learning-based multiple object registration strategies, traditional intensity-based registration methods are now routinely employed to solve compound, multiple object, registration problems across broad applications, such as: kinematic measurements of bone (Chen et al., 2012; Otake et al., 2016; Abe et al., 2019), rib motion and respiratory analysis (Hiasa et al., 2019), intra-operative assessment of adjusted bone fragments (Grupp et al., 2019, 2020b; Han et al., 2021), confirmation of screw implant placement during spine surgery (Uneri et al., 2017) and the positioning of a surgical robot with respect to target anatomy (Gao et al., 2020a). We believe that the lack of new multiple object registration learning-based strategies is indicative of the substantial challenges involved with their development, rather than any perceived lack of the problem's importance by the community. In order to better understand this "gap" between learning-based and traditional intensity-based methods in the multiple object domain, we first revisit (1) and update it to account for N 3D objects. For $i = 1, \ldots, N$, let $\omega_D^{(i)}$ denote the deformation parameters of the ith object and let $^v T_{x^{(i)}}$ denote it's pose with respect to the vth view. Since the vast majority of

studies examined concern 2D X-ray images, we also assume that 2D view modality is X-ray. This allows us to take advantage of the line integral nature of X-ray projection physics and represent the synthetic X-ray images formed from multiple objects as the sums of independent synthetic images created from each individual object, yielding the updated registration objective function:

$$\left\{\hat{K}_v, {}^v\hat{T}_{x^{(i)}}, \hat{\omega}_D^{(i)}\right\} = \operatorname*{arg\,min}_{K_v, {}^v T_{x^{(i)}}, \omega_D^{(i)}} \sum_v S\left(\sum_i \left[\left(P\left(K_v, {}^v T_{x^{(i)}}\right) \circ D_{\omega_D^{(i)}}\right)\left(x^{(i)}\right)\right], y_v\right) + R(\cdot).$$

(3)

Solving this optimization becomes more challenging as objects are added and the dimensionality of the search space grows. These challenges are somewhat mitigated by the compositional nature of the individual components of (**Eq. 3**). Indeed, updating an intensity-based registration framework to compute (3) instead of (**Eq. 1**) is relatively straight forward from an implementation perspective: compute N synthetic radiographs instead of one and perform a pixel-wise sum of the synthetic images together before calculating the image similarity metric, $S(\cdot, \cdot)$. The compositional structure of (**Eq. 3**) also enables the high-dimensional problem to be solved by successively solving lower-dimensional sub-problems. For example, the pose of a single object may be optimized while keeping the poses of all other objects fixed at their most recent estimates. After this optimization of a single object is complete, another object's pose is optimized and all other's are kept constant. This process is cycled until all objects have been registered once or some other termination criteria is met.

Extending (3) to the ML case, requires new per-object model parameters to be introduced: $\theta_D^{(i)}$ and $\theta_x^{(i)}$ for $i = 1, \ldots, N$. The updated ML-based objective function for multiple objects is written as:

$$\left\{\hat{K}_v, {}^v\hat{T}_{x^{(i)}}, \hat{\omega}_D^{(i)}\right\} = \operatorname*{arg\,min}_{K_v, {}^v T_{x^{(i)}}, \omega_D^{(i)}}$$
$$\times \sum_v S^{\theta_s}\left(\sum_i \left[\left(P\left(K_v, {}^v T_{x^{(i)}}\right) \circ D_{\omega_D^{(i)}}^{\theta_D^{(i)}}\right)\left(G_x^{\theta_x^{(i)}}\left(x^{(i)}\right)\right)\right], G_y^{\theta_y}(y_v)\right) + R^{\theta_R}(\cdot).$$

(4)

When considering the $G_x^{\theta_x^{(i)}}(x^{(i)})$ terms in (**Eq. 4**), it appears that contextualization methods are able to isolate some object-dependent parameters from other components of the registration problem. However, we found that each of the contextualization studies which considered multiple objects (Grupp et al., 2020c; Doerr et al., 2020; Wang et al., 2020; Esfandiari et al., 2021) did not explicitly isolate the parameters corresponding to each object, but instead used a single NN with data in the output layer indexed according to the appropriate object. This is likely effective for very specific and simple applications as the NN is able to learn the relative spatial relationships of the objects. For more complex applications, it may be more appropriate to learn separate,

well-trained, models which may be composed together. As an example, consider two separate models, one trained to identify contextual information of the hip joint and the other trained to produce contextual information of a robotic device. Although these two models may be used in conjunction, or combined through "fine-tuning," for the application of robotic hip surgery, the models may also be composed with models developed for other applications, such as measuring hip biomechanics or robotic shoulder surgery. However, a singular model trained to jointly contextualize hip and robot features would most likely fail to generalize to any additional applications. We anticipate that the development of independent and robust contextualization models, capable of composition, shall accelerate the number learning-based methods applied to the multiple object problem.

Although representation learning methods may function in the presence of multiple objects without additional modification, better performance is usually obtained by adding structure to account for the known spatial relationships between objects (Yokota et al., 2013). We anticipate that methods relying on representation learning shall migrate away from PCA, towards auto-encoder style approaches, with embedded rigid transformer modules so that all spatial and shape relationships may be learned. As representation learning methods have the potential to reconstruct the bony anatomy of joints from a sparse number of 2D views, these could potentially become very popular by enabling navigation to a 3D anatomical model with substantially reduced radiation exposure.

Similarity modeling approaches are implicitly affected by the introduction of additional objects due to the inputs of $S^{\theta_S}(\cdot, \cdot)$ having a dependence on $G_x^{\theta_x}(x^{(i)})$ and $\theta_D^{(i)}$. None of the studies reviewed in this paper used multiple objects as part of similarity modeling. Although Grupp et al. (2020c) registered multiple objects, the regularization function employed learned landmark annotations from a only a single object. We envision several challenges with extending these methods to properly accommodate multiple objects. Obtaining convexity of the learned similarity models in the ideal case for (Gao et al., 2020c; Gu et al., 2020) is most likely not attainable when using a weighted sum of rotation and translation components of each object's pose offset, especially as the number of objects increases. Methods which model 2D/3D correspondences (Schaffert et al., 2019, 2020a,b; Liao et al., 2019) will require an additional dimension to handle the assignment of points to various obejcts, adding complexity and potentially increasing the challenges associated with training.

Probably most handicapped by the introduction of multiple objects to the registration problem, are methods relying on the direct regression of pose parameters, as they attempt to model the entire objective function. One may be tempted to solve this problem by simply adding additional model outputs for each object's estimated pose, but this does not guarantee that the limitations of independent single object pose regression are addressed. Therefore, direct

multiple object pose regression methods should attempt to model the relative poses between objects in addition to a single absolute pose (or single pose relative to some initialization). Another limitation of regression approaches lies in the combinatorial explosion which occurs as new objects are added to the registration problem, making training difficult in the presence of large fluctuations of the loss function.

Although not unique to learning-based methods, multiple object registration verification also becomes a much more complex problem as the number of objects considered increases. Questions arise, such as: should verification be reported on each relative pose or should an overall pass/fail be reported? The three verification studies examined in this paper only considered verification of a single object's pose estimate. As single object verification methods mature, their issues when expanding to multiple objects will likely become more apparent.

In light of these challenges associated with learning-based approaches, it is easy to see how contemporary intensity-based methods currently dominate the multiple object domain given their relative ease of implementation and reasonable performance. However, there are multiple object registration problems which remain unsolved since multiple object intensity-based methods continue to suffer from the limitations previously identified in Section 1.2. Some frequent properties of these unsolved problems are misleading views with several objects having substantial overlap in 2D, the potential for a varying number of surgical instruments to be present in a view and the existence of objects which are dynamically changing shape or possibly "splitting" into several new objects. Real-time osteotome navigation using single view fluoroscopy, illustrated in **Figure 4**, is exemplary of the many unsolved challenges outlined above and will be used as a motivating example throughout this discussion.

Each time the osteotome is advanced through bone, a single fluoroscopic view is collected and interpreted by the surgeon in order to accurately adjust the chiseling trajectory and safely avoid sensitive components of the anatomy which must not be damaged, such as the acetabulum shown in 4. Although the oblique views shown in **Figure 4** 1) and (c) help ensure that the osteotome tip is distinguishable from the acetabulum, the lateral orientation of the tool is challenging to interpret manually and may need to be confirmed by collecting subsequent fluoroscopic views at substantially different orientations, such as the approximate anterior-posterior (AP) view shown in **Figure 4** (b). **Figure 4** (c) also demonstrates the intraoperative creation of a new object which may move independently of all others, further compounding the difficulty associated with non-navigated interpretation of these views. An additional challenge with osteotomy cases is the ability to determine when all osteotomies are complete and the new bone fragment is completely free from its parent. **Figure 4** (c) demonstrates that retaining the angled osteotome as a static object in the field of view during performance of the posterior cut helps to

guide the flat osteotome along the desired trajectory, but does not guarantee that the acetabular fragment has in fact been freed from the pelvis. A navigation system capable of accurately tracking the osteotome with respect to the anatomy using a series of oblique views would eliminate the need for additional, "confirmation," views, potentially reduce operative time, decrease radiation exposure to the patient and clinical team and reduce the frequency of breaches into sensitive anatomy.

We envision that this problem, and others like it, will eventually be solvable as the capabilities of learning-based methods advance further. Future contextualization and similarity modeling methods could enable a large enough capture range to provide an automatic registration to the initial oblique view used for chiseling, and regression methods, coupled with contextualization, could enable real-time pose estimates in subsequent views. Reconstructions of newly created bone fragments should be possible through similarity modeling approaches, such as by extending the framework introduced by Gao et al. (2020c) to include multiple objects and non-rigid deformation. Finally, recurrent NN approaches, such as long short-term memory (LSTM) components, could provide some temporal segmentation of the intervention into phases and gestures, as is already done for laparoscopic surgery (Vercauteren et al., 2019; Garrow et al., 2021; Wu et al., 2021). This segmentation could be useful in identifying when 1) certain objects need to be tracked, 2) a radically different view has been collected, or 3) new objects have been split off from an existing object.

Although foreign objects such as screws, K-wires, retractors, osteotomes, etc. frequently confound traditional registration methods, their location or poses often have clinical relevance. Therefore, contrary to the inpainting approach of Esfandiari et al. (2021), we advocate that learning-based methods should make attempts to register these objects.

3.4 Estimating Uncertainty and Assuring Quality

The four studies which consider registration verification in this review all attempt to provide a low dimensional classification of registration success, e.g., pass/fail or correct/poor/incorrect (Mitrovi et al., 2014; Varnavas et al., 2013, 2015a; Wu et al., 2016). These strategies effectively attempt to label registration estimates as either a global (correct) or local minimum (poor/incorrect) of (Eq. 1). Unfortunately, even when correctly classifying a global optimum, these low-dimensional categorizations of a registration result may unintentionally fail to report small, but clinically relevant, errors. This is perhaps easiest to recognize by first revisiting (1) and noting that registration strategies attempt to find singular solutions, or point estimates, which best minimize the appropriate objective function. However, several factors, including sensor noise, modeling error or numerical imprecision, may influence the landscape of the objective

function and potentially even cause some variation in the location of the global minimum. This possibility of obtaining several different pose estimates under nominally equivalent conditions, reveals the inherent uncertainty of registration. Even though a failure to identify cases of large uncertainties may lead surgeons to take unintentional risks with potentially catastrophic implications for patients, to our knowledge, only one prior work has attempted to estimate the error associated with 2D/3D registration estimates (Hu et al., 2016). We therefore believe that the development of methods which quantify 2D/3D registration uncertainty is of paramount importance and essential for the eventual adoption of 2D/3D registration into routine use, and even more in the advent of autonomous robotic surgery.

Inspiration can be drawn from the applications of 3D/3D and 2D/2D deformable image registration (Chen et al., 2020; Fu et al., 2020; Haskins et al., 2020), where machine learning techniques have dominated most existing research into registration uncertainty. Some of these approaches report interval estimates, either by reformulating the registration objective function as a probability distribution and drawing samples (Risholm et al., 2013; Le Folgoc et al., 2016; Schultz et al., 2018), approximating the sampling process using test-time NN drop-out (Yang et al., 2017) or sampling using the test-time deformation covariance matrices embedded within a variational autoencoder (Dalca et al., 2019). Due to the interventional nature of most 2D/3D registration applications, care needs to be taken in order to ensure that program runtimes are compatible with intra-operative workflows.

Although we have mostly advocated for the development of new uncertainty quantification methods, there is likely still room for improvement of the local/global minima classification problem. Given the computational capacity of contemporary GPUs and sophistication of learning frameworks, it should be feasible to extend the approach of Wu et al. (2016) and pass densely sampled regions of the objective function through a NN, relying on the learning process to extract optimal features for distinguishing local and global minimal.

As registration methods will inevitably report failure or large uncertainties under certain conditions, we also believe that a promising topic of future research extends to intelligent agents which would determine subsequent actions to optimally reduce uncertainty, such as collecting another 2D view from a specific viewpoint. Finally, we note that registration uncertainties have the potential to augment existing robotic control methods which rely on intermittent imaging (Alambeigi et al., 2019).

3.5 The Need for More Generic Solutions

Traditional image-based 2D/3D registration that relies on optimization of a cost function is limited in many ways (Section 1.2); however, a major advantage of the algorithmic approach is that it is very generic, i.e., a pipeline configured for 2D/3D registration of the pelvis would be equally applicable to the spine. The introduction of machine learning to the 2D/3D pipeline, however, has in a

way taken away this generality and methods have become substantially more specialized. This is not because specific machine learning models are only suitable for one specific task or anatomy; on the contrary, e.g., pose or PCA component regression techniques are largely identical across clinical targets. Rather, it is yet another consequence of the compact domain assumption of machine learning models that confine solutions to the data domain they were developed on. While this specificity of solutions may be acceptable, and perhaps, unavoidable for methods that seek to contextualize data, it inhibits the use of 2D/3D registration pipelines as a tool to answer clinical or research questions. We feel that, in part due to its complexity, 2D/3D registration already leads a niche existence and the necessity to re-train or even re-develop some algorithm components for the specific task at hand further exacerbates this situation.

Some approaches already point towards potential solutions to this issue: Methods like (Varnavas et al., 2013; 2015b) rely on patient- and object-specific training sets, and therefore, avoid the task-specificity that comes with one-time training. Another noteworthy method learns to establish sparse point correspondences between random points rather than anatomically relevant ones (Liao et al., 2019). However, despite matching random points, this method is likely still scene-specific because matching is performed using a very deep CNN with very large receptive field which may have learned to exploit the global context of the random keypoints, which would not be preserved in different anatomy. Contributing machine learning-based solutions that address some of the large open challenges in 2D/3D registration while making the resulting tools general purpose and easy to use will be an important goal in the immediate future.

4 CONCLUSION

Machine learning-based improvements to image-based 2D/3D registration were already of interest before the deep learning era (Gouveia et al., 2012), and deep learning has only accelerated and diversified the contributions to the field. Contextualization of data, representation learning to reduce problem dimensionality, similarity modeling for increased capture range, direct pose regression to avoid iterative optimization, as well as confidence assessment are all well established research thrusts, which are geared towards developing automated registration pipelines. While convincing performance improvements are reported across a variety of clinical tasks and problem settings already today, most of those studies are performed "on the benchtop."

Coordinated research efforts are desirable towards 1) developing more robust learning paradigms that succeed under domain shift, 2) creating standardized reporting templates and devising evaluation metrics to enhance reproducibiliy and enable comparisons, 3) researching multi-object registration methods that can deal with the presence of foreign objects, 4) advancing uncertainty quantification and confidence estimation methodology to better support human decisions, and finally 5) developing generalist machine learning components of 2D/3D registration pipelines to improve accessibility.

Even though learning-based methods show great promise and have supplanted traditional methods in many aspects, their rise should not render traditional methods unusable or irrelevant. New researchers typically spend a great deal of time implementing a traditional registration pipeline so that traditional methods may be leveraged in conjunction with the development of learning-based approaches. In order to facilitate more rapid research and development towards learning-based methods, we have made the core registration components of our intensity-based framework (xReg[3]), as well as our physics-based DRR generation tools for realistic synthesis and generalizable learning (DeepDRR[4]) available as open source software projects.

We have no doubt that progress on the aforementioned fronts will firmly establish machine learning methodology as an important component for 2D/3D registration workflows that will substantially contribute to 2D/3D registration growing out of its niche existence, establishing itself as a reliable, precise, and easy-to-use component for research, and more importantly, at the bedside.

DATA AVAILABILITY STATEMENT

The original contribution presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

MU conceived of the presented idea was in charge of overall direction and planning. MU, CG, and RG refined the scope of the presented review and perspective. MU, CG, R.G, M.J, and YH screened abstracts and full texts during review and extracted information from the included studies. MU, CG, and RG merged extracted information and carried out the systematic review. All authors contributed to writing the manuscript. They also provided critical feedback, and helped shape the research and analysis.

ACKNOWLEDGMENTS

---

[3]https://github.com/rg2/xreg
[4]https://github.com/arcadelab/deepdrr

REFERENCES

Abe, S., Otake, Y., Tennma, Y., Hiasa, Y., Oka, K., Tanaka, H., et al. (2019). Analysis of Forearm Rotational Motion Using Biplane Fluoroscopic Intensity-Based 2D-3D Matching. *J. Biomech.* 89, 128–133. doi:10.1016/j.jbiomech.2019.04.017

Alambeigi, F., Pedram, S. A., Speyer, J. L., Rosen, J., Iordachita, I., Taylor, R. H., et al. (2019). Scade: Simultaneous Sensor Calibration and Deformation Estimation of Fbg-Equipped Unmodeled Continuum Manipulators. *IEEE Trans. Robot* 36, 222–239. doi:10.1109/tro.2019.2946726

Amrehn, M., Gaube, S., Unberath, M., Schebesch, F., Horz, T., Strumia, M., et al. (2017). Ui-net: Interactive Artificial Neural Networks for Iterative Image Segmentation Based on a User Model. *Proc. Eurographics Workshop Vis. Comput. Biol. Med.*, 143–147.

Baka, N., Lelieveldt, B. P. F., Schultz, C., Niessen, W., and Van Walsum, T. (2015). Respiratory Motion Estimation in X-ray Angiography for Improved Guidance during Coronary Interventions. *Phys. Med. Biol.* 60, 3617–3637. doi:10.1088/0031-9155/60/9/3617

Berger, M., Müller, K., Aichert, A., Unberath, M., Thies, J., Choi, J. H., et al. (2016). Marker-free Motion Correction in Weight-bearing Cone-beam CT of the Knee Joint. *Med. Phys.* 43, 1235–1248. doi:10.1118/1.4941012

Bier, B., Goldmann, F., Zaech, J.-N., Fotouhi, J., Hegeman, R., Grupp, R., et al. (2019). Learning to Detect Anatomical Landmarks of the Pelvis in X-Rays from Arbitrary Views. *Int. J. CARS* 14, 1463–1473. doi:10.1007/s11548-019-01975-5

Bier, B., Unberath, M., Zaech, J.-N., Fotouhi, J., Armand, M., Osgood, G., et al. (2018). *X-ray-transform Invariant Anatomical Landmark Detection for Pelvic Trauma Surgery.* Granada, Spain: Springer-Verlag, 55–63. doi:10.1007/978-3-030-00937-3_7

Birkmeyer, J. D., Finks, J. F., O'Reilly, A., Oerline, M., Carlin, A. M., Nunn, A. R., et al. (2013). Surgical Skill and Complication Rates after Bariatric Surgery. *N. Engl. J. Med.* 369, 1434–1442. doi:10.1056/nejmsa1300625

Brost, A., Wimmer, A., Liao, R., Bourier, F., Koch, M., Strobel, N., et al. (2012). Constrained Registration for Motion Compensation in Atrial Fibrillation Ablation Procedures. *IEEE Trans. Med. Imaging* 31, 870–881. doi:10.1109/tmi.2011.2181184

Chen, X., Graham, J., Hutchinson, C., and Muir, L. (2012). Automatic Inference and Measurement of 3d Carpal Bone Kinematics from Single View Fluoroscopic Sequences. *IEEE Trans. Med. Imaging* 32, 317–328. doi:10.1109/TMI.2012.2226740

Chen, X., Diaz-Pinto, A., Ravikumar, N., and Frangi, A. (2020). Deep Learning in Medical Image Registration. *Prog. Biomed. Eng.* doi:10.1088/2516-1091/abd37c

Chen, Z.-Q., Wang, Z.-W., Fang, L.-W., Jian, F.-Z., Wu, Y.-H., Li, S., et al. (2018). Real-time 2d/3d Registration of Vertebra via Machine Learning and Geometric Transformation. *Zidonghua Xuebao/Acta Automatica Sinica* 44, 1183–1194.

Chou, C.-R., Frederick, B., Mageras, G., Chang, S., and Pizer, S. (2013). 2d/3d Image Registration Using Regression Learning. *Computer Vis. Image Understanding* 117, 1095–1106. doi:10.1016/j.cviu.2013.02.009

Chou, C.-R., and Pizer, S. (2014). *Local Regression Learning via forest Classification for 2d/3d Deformable Registration.* Nagoya, Japan: Springer-Verlag, 24–33. doi:10.1007/978-3-319-14104-6_3

Chou, C.-R., and Pizer, S. (2013). *Real-time 2d/3d Deformable Registration Using Metric Learning.* Nice, France: Springer-Verlag, 1–10. doi:10.1007/978-3-642-36620-8_1

Dalca, A. V., Balakrishnan, G., Guttag, J., and Sabuncu, M. R. (2019). Unsupervised Learning of Probabilistic Diffeomorphic Registration for Images and Surfaces. *Med. image Anal.* 57, 226–236. doi:10.1016/j.media.2019.07.006

Doerr, S. A., Uneri, A., Huang, Y., Jones, C. K., Zhang, X., Ketcha, M. D., et al. (2020). *Data-driven Detection and Registration of Spine Surgery Instrumentation in Intraoperative Images.* Houston, TX, United states: The Society of Photo–Optical Instrumentation Engineers.

Du, M., Liu, N., and Hu, X. (2019). Techniques for Interpretable Machine Learning. *Commun. ACM* 63, 68–77. doi:10.1145/3359786

Esfandiari, H., Weidert, S., Kövesházi, I., Anglin, C., Street, J., and Hodgson, A. J. (2021). Deep Learning-Based X-ray Inpainting for Improving Spinal 2d-3d Registration. *The Int. J. Med. robotics + Comput. Assist. Surg. : MRCAS.*

Esteban, J., Grimm, M., Unberath, M., Zahnd, G., and Navab, N. (2019). Towards Fully Automatic X-ray to Ct Registration. In International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 631–639. doi:10.1007/978-3-030-32226-7_70

Ewurum, C. H., Guo, Y., Pagnha, S., Feng, Z., and Luo, X. (2018). Surgical Navigation in Orthopedics: Workflow and System Review. *Intell. Orthopaedics*, 47–63. doi:10.1007/978-981-13-1396-7_4

Foley, J. P., and Hsu, W. K. (2021). Effectiveness of Bioskills Training in Spinal Surgery. *Contemp. Spine Surg.* 22, 1–7. doi:10.1097/01.css.0000734864.37046.9b

Foote, M. D., Zimmerman, B. E., Sawant, A., and Joshi, S. C. (2019). *Real-time 2d-3d Deformable Registration with Deep Learning and Application to Lung Radiotherapy Targeting.* Hong Kong, China: Springer-Verlag, 265–276. doi:10.1007/978-3-030-20351-1_20

François, T., Calvet, L., Madad Zadeh, S., Saboul, D., Gasparini, S., Samarakoon, P., et al. (2020). Detecting the Occluding Contours of the Uterus to Automatise Augmented Laparoscopy: Score, Loss, Dataset, Evaluation and User Study. *Int. J. CARS* 15, 1177–1186. doi:10.1007/s11548-020-02151-w

Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., and Yang, X. (2020). Deep Learning in Medical Image Registration: a Review. *Phys. Med. Biol.* 65, 20TR01. doi:10.1088/1361-6560/ab843e

Gao, C., Farvardin, A., Grupp, R. B., Bakhtiarinejad, M., Ma, L., Thies, M., et al. (2020a). Fiducial-free 2d/3d Registration for Robot-Assisted Femoroplasty. *IEEE Trans. Med. Robot. Bionics* 2, 437–446. doi:10.1109/tmrb.2020.3012460

Gao, C., Grupp, R. B., Unberath, M., Taylor, R. H., and Armand, M. (2020b). Fiducial-free 2d/3d Registration of the Proximal Femur for Robot-Assisted Femoroplasty. In Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling, 11315. Bellingham, WA: International Society for Optics and Photonics, 113151C. doi:10.1117/12.2550992

Gao, C., Liu, X., Gu, W., Killeen, B., Armand, M., Taylor, R., et al. (2020c). *Generalizing Spatial Transformers to Projective Geometry with Applications to 2d/3d Registration*, 12263. Lima, Peru: Springer Science and Business Media Deutschland GmbH, 329–339. doi:10.1007/978-3-030-59716-0_32Generalizing Spatial Transformers to Projective Geometry with Applications to 2D/3D Registration

Gao, C., Unberath, M., Taylor, R., and Armand, M. (2019). Localizing Dexterous Surgical Tools in X-ray for Image-Based Navigation. In International Conference on Information Processing for Computer-Assisted Interventions, Proceedings of.

Garrow, C. R., Kowalewski, K.-F., Li, L., Wagner, M., Schmidt, M. W., Engelhardt, S., et al. (2021). Machine Learning for Surgical Phase Recognition. *Ann. Surg.* 273, 684–693. doi:10.1097/sla.0000000000004425

Gouveia, A. I. R., Metz, C., Freire, L., and Klein, S. (2012). Comparative Evaluation of Regression Methods for 3d-2d Image Registration. In International Conference on Artificial Neural Networks. Springer, 238–245. doi:10.1007/978-3-642-33266-1_30

Grupp, R. B., Hegeman, R. A., Murphy, R. J., Alexander, C. P., Otake, Y., McArthur, B. A., et al. (2020). Pose Estimation of Periacetabular Osteotomy Fragments with Intraoperative X-ray Navigation. *IEEE Trans. Biomed. Eng.* 67, 441–452. doi:10.1109/TBME.2019.2915165

Grupp, R. B., Murphy, R. J., Hegeman, R. A., Alexander, C. P., Unberath, M., Otake, Y., et al. (2020b). Fast and Automatic Periacetabular Osteotomy Fragment Pose Estimation Using Intraoperatively Implanted Fiducials and Single-View Fluoroscopy. *Phys. Med. Biol.* 65, 245019. doi:10.1088/1361-6560/aba089

Grupp, R. B., Unberath, M., Gao, C., Hegeman, R. A., Murphy, R. J., Alexander, C. P., et al. (2020c). Automatic Annotation of Hip Anatomy in Fluoroscopy for Robust and Efficient 2d/3d Registration. *Int. J. CARS* 15, 759–769. doi:10.1007/s11548-020-02162-7

Grupp, R., Unberath, M., Gao, C., Hegeman, R., Murphy, R., Alexander, C., et al. (2020a). Data and Code Associated with the Publication: Automatic Annotation of Hip Anatomy in Fluoroscopy for Robust and Efficient 2D/3D Registration doi:10.7281/T1/IFSXNV

Gu, W., Gao, C., Grupp, R., Fotouhi, J., and Unberath, M. (2020). *Extended Capture Range of Rigid 2d/3d Registration by Estimating Riemannian Pose Gradients.* Lima, Peru: Springer Science and Business Media Deutschland GmbH, 281–291. doi:10.1007/978-3-030-59861-7_29

Guan, S., Meng, C., Sun, K., and Wang, T. (2019). *Transfer Learning for Rigid 2d/3d Cardiovascular Images Registration.* Xi'an, China: Springer Science and Business Media Deutschland GmbH, 380–390. doi:10.1007/978-3-030-31723-2_32Transfer Learning for Rigid 2D/3D Cardiovascular Images Registration

Guan, S., Wang, T., Sun, K., and Meng, C. (2020). *Transfer Learning for Nonrigid 2d/3d Cardiovascular Images Registration*.

Hafezi-Nejad, N., Bailey, C. R., Solomon, A. J., Abou Areda, M., Carrino, J. A., Khan, M., et al. (2020). Vertebroplasty and Kyphoplasty in the usa from 2004 to 2017: National Inpatient Trends, Regional Variations, Associated Diagnoses, and Outcomes. *J. NeuroInterventional Surg.*

Han, R., Uneri, A., Vijayan, R., Wu, P., Vagdargi, P., Sheth, N., et al. (2021). Fracture Reduction Planning and Guidance in Orthopaedic Trauma Surgery via Multi-Body Image Registration. *Med. Image Anal.* 68, 101917. doi:10.1016/j.media.2020.101917

Hansen, N., Müller, S. D., and Koumoutsakos, P. (2003). Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (Cma-es). *Evol. Comput.* 11, 1–18. doi:10.1162/106365603321828970

Haskins, G., Kruger, U., and Yan, P. (2020). Deep Learning in Medical Image Registration: a Survey. *Machine Vis. Appl.* 31, 1–18. doi:10.1007/s00138-020-01060-x

Hiasa, Y., Otake, Y., Tanaka, R., Sanada, S., and Sato, Y. (2019). Recovery of 3d Rib Motion from Dynamic Chest Radiography and Ct Data Using Local Contrast Normalization and Articular Motion Model. *Med. image Anal.* 51, 144–156. doi:10.1016/j.media.2018.10.002

Hou, B., Alansary, A., McDonagh, S., Davidson, A., Rutherford, M., Hajnal, J. V., et al. (2017). *Predicting Slice-To-Volume Transformation in Presence of Arbitrary Subject Motion*. Quebec City, QC, Canada: Springer-Verlag, 296–304. doi:10.1007/978-3-319-66185-8_34

Hou, B., Miolane, N., Khanal, B., Lee, M. C. H., Alansary, A., McDonagh, S., et al. (2018). *Computing Cnn Loss and Gradients for Pose Estimation with Riemannian Geometry*. Granada, Spain: Springer-Verlag, 756–764. doi:10.1007/978-3-030-00928-1_85

Hu, Y., Bonmati, E., Gibson, E., Hipwell, J. H., Hawkes, D. J., Bandula, S., et al. (2016).2d-3d Registration Accuracy Estimation for Optimised Planning of Image-Guided Pancreatobiliary Interventions. In International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 516–524. doi:10.1007/978-3-319-46720-7_60

Hu, Y., Modat, M., Gibson, E., Ghavami, N., Bonmati, E., Moore, C. M., et al. (2018).Label-driven Weakly-Supervised Learning for Multimodal Deformable Image Registration. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, 1070–1074.

Hummel, J., Figl, M., Bax, M., Bergmann, H., and Birkfellner, W. (2008). 2d/3d Registration of Endoscopic Ultrasound to Ct Volume Data. *Phys. Med. Biol.* 53, 4303–4316. doi:10.1088/0031-9155/53/16/006

Joskowicz, L., and Hazan, E. J. (2016). Computer Aided Orthopaedic Surgery: Incremental Shift or Paradigm Change?

Karner, F., Gsaxner, C., Pepe, A., Li, J., Fleck, P., Arth, C., et al. (2020). *Single-shot Deep Volumetric Regression for mobile Medical Augmented Reality*. Lima, Peru: Springer Science and Business Media Deutschland GmbH, 64–74. doi:10.1007/978-3-030-60946-7_7

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–1105.

Le Folgoc, L., Delingette, H., Criminisi, A., and Ayache, N. (2016). Quantifying Registration Uncertainty with Sparse Bayesian Modelling. *IEEE Trans. Med. Imaging* 36, 607–617. doi:10.1109/TMI.2016.2623608

Leonard, S., Reiter, A., Sinha, A., Ishii, M., Taylor, R. H., and Hager, G. D. (2016).Image-based Navigation for Functional Endoscopic Sinus Surgery Using Structure from Motion. In Medical Imaging 2016: Image Processing. Bellingham, WA: International Society for Optics and Photonics, 97840V. doi:10.1117/12.2217279

Lepetit, V., Moreno-Noguer, F., and Fua, P. (2009). EPnP: An Accurate O(n) Solution to the PnP Problem. *Int. J. Comput. Vis.* 81, 155–166. doi:10.1007/s11263-008-0152-6

Li, P., Pei, Y., Guo, Y., Ma, G., Xu, T., and Zha, H. (2020). *Non-rigid 2d-3d Registration Using Convolutional Autoencoders*, 2020. Iowa City, IA, United states: IEEE Computer Society, 700–704.

Liao, H., Lin, W.-A., Zhang, J., Zhang, J., Luo, J., and Zhou, S. K. (2019). *Multiview 2d/3d Rigid Registration via a point-of-interest Network for Tracking and Triangulation*, 2019. Long Beach, CA, United states: IEEE Computer Society, 12630–12639.

Liao, R., Zhang, L., Sun, Y., Miao, S., and Chefd'Hotel, C. (2013). A Review of Recent Advances in Registration Techniques Applied to Minimally Invasive Therapy. *IEEE Trans. Multimedia* 15, 983–1000. doi:10.1109/tmm.2013.2244869

Lin, C., and Winey, B. (2012). Shape Distribution-Based 2d/3d Registration for Fast and Accurate 6 Degrees-Of-freedom Stereotactic Patient Positioning. *Int. J. Radiat. Oncology*Biology*Physics* 84, S724. doi:10.1016/j.ijrobp.2012.07.1939

Liu, D. C., and Nocedal, J. (1989). On the Limited Memory Bfgs Method for Large Scale Optimization. *Math. programming* 45, 503–528. doi:10.1007/bf01589116

Luo, X., Zeng, H.-Q., Du, Y.-P., and Cheng, X. (2019). *Towards Multiple Instance Learning and hermann Weyls Discrepancy for Robust Image-Guided Bronchoscopic Intervention*. Shenzhen, China: Springer Science and Business Media Deutschland GmbHLNCS, 403–411. doi:10.1007/978-3-030-32254-0_45

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., and Suetens, P. (1997). Multimodality Image Registration by Maximization of Mutual Information. *IEEE Trans. Med. Imaging* 16, 187–198. doi:10.1109/42.563664

Markelj, P., Tomaževič, D., Likar, B., and Pernuš, F. (2012). A Review of 3d/2d Registration Methods for Image-Guided Interventions. *Med. image Anal.* 16, 642–661. doi:10.1016/j.media.2010.03.005

Mezger, U., Jendrewski, C., and Bartels, M. (2013). Navigation in Surgery. *Langenbecks Arch. Surg.* 398, 501–514. doi:10.1007/s00423-013-1059-4

Miao, S., Piat, S., Fischer, P., Tuysuzoglu, A., Mewes, P., Mansi, T., et al. (2018). *Dilated Fcn for Multi-Agent 2d/3d Medical Image Registration*. New Orleans, LA, United states: AAAI press, 4694–4701.

Miao, S., Wang, Z. J., and Liao, R. (2016a). A Cnn Regression Approach for Real-Time 2d/3d Registration. *IEEE Trans. Med. Imaging* 35, 1352–1363. doi:10.1109/tmi.2016.2521800

Miao, S., Wang, Z. J., Zheng, Y., and Liao, R. (2016b). *Real-time 2d/3d Registration via Cnn Regression*, 2016. Prague, Czech republic: IEEE Computer Society, 1430–1434.

Mirota, D. J., Ishii, M., and Hager, G. D. (2011). Vision-based Navigation in Image-Guided Interventions. *Annu. Rev. Biomed. Eng.* 13, 297–319. doi:10.1146/annurev-bioeng-071910-124757

Mitrović, U., Pernuš, F., Likar, B., and Špiclin, Ž. (2015). Simultaneous 3D-2D Image Registration and C-arm Calibration: Application to Endovascular Image-guided Interventions. *Med. Phys.* 42, 6433–6447. doi:10.1118/1.4932626

Mitrović, U., Špiclin, Ž., Likar, B., and Pernuš, F. (2014). *Automatic Detection of Misalignment in Rigid 3d-2d Registration*, 8361. Nagoya, Japan: Springer-Verlag, 117–124. doi:10.1007/978-3-319-14127-5_15

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the Prisma Statement. *BMJ* 339, b2535. doi:10.1136/bmj.b2535

Neumann, C., Tonnies, K. D., and Pohle-Frohlich, R. (2020). *Deep Similarity Learning Using a Siamese Resnet Trained on Similarity Labels from Disparity Maps of Cerebral Mra Mip Pairs*, 11313. Houston, TX, United states: The Society of Photo–Optical Instrumentation Engineers.

Nolte, L.-P., Slomczykowski, M. A., Berlemann, U., Strauss, M. J., Hofstetter, R., Schlenzka, D., et al. (2000). A New Approach to Computer-Aided Spine Surgery: Fluoroscopy-Based Surgical Navigation. *E Spine J.* 9, S078–S088. doi:10.1007/pl00010026

Otake, Y., Esnault, M., Grupp, R., Kosugi, S., and Sato, Y. (2016).Robust Patella Motion Tracking Using Intensity-Based 2d-3d Registration on Dynamic Bi-plane Fluoroscopy: towards Quantitative Assessment in Mpfl Reconstruction Surgery. In Medical Imaging 2016: Image-Guided Procedures, Robotic Interventions, and Modeling. International Society for Optics and Photonics, 97860B. doi:10.1117/12.2214699

Otake, Y., Wang, A. S., Webster Stayman, J., Uneri, A., Kleinszig, G., Vogt, S., et al. (2013). Robust 3D-2D Image Registration: Application to Spine Interventions and Vertebral Labeling in the Presence of Anatomical Deformation. *Phys. Med. Biol.* 58, 8535–8553. doi:10.1088/0031-9155/58/23/8535

Pei, Y., Zhang, Y., Qin, H., Ma, G., Guo, Y., Xu, T., et al. (2017). *Non-rigid Craniofacial 2d-3d Registration Using Cnn-Based Regression*. Quebec City, QC, Canada: Springer-Verlag, 117–125. doi:10.1007/978-3-319-67558-9_14

Pfandler, M., Stefan, P., Mehren, C., Lazarovici, M., and Weigl, M. (2019). Technical and Nontechnical Skills in Surgery. *Spine* 44, E1396–E1400. doi:10.1097/brs.0000000000003154

Picard, F., Clarke, J., Deep, K., and Gregori, A. (2014). Computer Assisted Knee Replacement Surgery: Is the Movement Mainstream? *Orthop. Muscular Syst.* 3.

Pluim, J. P., Muenzing, S. E., Eppenhof, K. A., and Murphy, K. (2016). The Truth Is Hard to Make: Validation of Medical Image Registration. In 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2294–2300. doi:10.1109/icpr.2016.7899978

Powell, M. J. (2009). *The Bobyqa Algorithm for Bound Constrained Optimization without derivativesCambridge NA Report NA2009/06*. Cambridge: University of Cambridge, 26–46.

Risholm, P., Janoos, F., Norton, I., Golby, A. J., and Wells, W. M., III (2013). Bayesian Characterization of Uncertainty in Intra-subject Non-rigid Registration. *Med. image Anal.* 17, 538–555. doi:10.1016/j.media.2013.03.002

Schaffert, R., Wang, J., Fischer, P., Borsdorf, A., and Maier, A. (2020a). Learning an Attention Model for Robust 2-D/3-D Registration Using point-to-plane Correspondences. *IEEE Trans. Med. Imaging* 39, 3159–3174. doi:10.1109/tmi.2020.2988410

Schaffert, R., Wang, J., Fischer, P., Borsdorf, A., and Maier, A. (2019). *Metric-driven Learning of Correspondence Weighting for 2-D/3-D Image Registration*, 11269. Stuttgart, Germany: Springer-Verlag, 140–152. doi:10.1007/978-3-030-12939-2_11

Schaffert, R., Weiß, M., Wang, J., Borsdorf, A., and Maier, A. (2020b). *Learning-based Correspondence Estimation for 2-D/3-D Registration*. Berlin, Germany: Springer Science and Business Media Deutschland GmbH, 222–228. doi:10.1007/978-3-658-29267-6_50

Schultz, S., Handels, H., and Ehrhardt, J. (2018). A Multilevel Markov Chain Monte Carlo Approach for Uncertainty Quantification in Deformable Registration. In Medical Imaging 2018: Image Processing, 10574. Bellingham, WA: International Society for Optics and Photonics), 105740O. doi:10.1117/12.2293588

Shetty, K., Birkhold, A., Strobel, N., Egger, B., Jaganathan, S., Kowarschik, M., et al. (2021). Deep Learning Compatible Differentiable X-ray Projections for Inverse Rendering. arXiv preprint arXiv:2102.02912doi:10.1007/978-3-658-33198-6_70

Sinha, A., Ishii, M., Hager, G. D., and Taylor, R. H. (2019). Endoscopic Navigation in the Clinic: Registration in the Absence of Preoperative Imaging. *Int. J. CARS* 14, 1495–1506. doi:10.1007/s11548-019-02005-0

Sugano, N. (2003). Computer-assisted Orthopedic Surgery. *J. Orthopaedic Sci.* 8, 442–448. doi:10.1007/s10776-002-0623-6

Tang, A., and Scalzo, F. (2016). *Similarity Metric Learning for 2d to 3d Registration of Brain Vasculature*, 10072. Las Vegas, NV, United states: Springer-Verlag, 3–12. doi:10.1007/978-3-319-50835-1_1

Thies, M., Zäch, J.-N., Gao, C., Taylor, R., Navab, N., Maier, A., et al. (2020). A Learning-Based Method for Online Adjustment of C-Arm Cone-Beam Ct Source Trajectories for Artifact Avoidance. *Int. J. CARS* 15, 1787–1796. doi:10.1007/s11548-020-02249-1

Toth, D., Miao, S., Kurzendorfer, T., Rinaldi, C. A., Liao, R., Mansi, T., et al. (2018). 3d/2d Model-To-Image Registration by Imitation Learning for Cardiac Procedures. *Int. J. CARS* 13, 1141–1149. doi:10.1007/s11548-018-1774-y

Tucker, E., Fotouhi, J., Unberath, M., Lee, S. C., Fuerst, B., Johnson, A., et al. (2018).Towards Clinical Translation of Augmented Orthopedic Surgery: from Pre-op Ct to Intra-op X-ray via Rgbd Sensing. In Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications. Bellingham, WA: International Society for Optics and Photonics, 105790J. doi:10.1117/12.2293675

Unberath, M., Zaech, J.-N., Gao, C., Bier, B., Goldmann, F., Lee, S. C., et al. (2019). Enabling Machine Learning in x-ray-based Procedures via Realistic Simulation of Image Formation. *Int. J. CARS* 14, 1517–1528. doi:10.1007/s11548-019-02011-2

Unberath, M., Zaech, J.-N., Lee, S. C., Bier, B., Fotouhi, J., Armand, M., et al. (2018).DeepDRR - A Catalyst for Machine Learning in Fluoroscopy-Guided Procedures. In International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 98–106. doi:10.1007/978-3-030-00937-3_12

Uneri, A., De Silva, T., Goerres, J., Jacobson, M. W., Ketcha, M. D., Reaungamornrat, S., et al. (2017). Intraoperative Evaluation of Device Placement in Spine Surgery Using Known-Component 3D-2D Image Registration. *Phys. Med. Biol.* 62, 3330–3351. doi:10.1088/1361-6560/aa62c5

Uneri, A., Otake, Y., Wang, A. S., Kleinszig, G., Vogt, S., Khanna, A. J., et al. (2013). 3D-2D Registration for Surgical Guidance: Effect of Projection View Angles on Registration Accuracy. *Phys. Med. Biol.* 59, 271–287. doi:10.1088/0031-9155/59/2/271

US Food and Drug Administration (2021). *Artificial Intelligence/machine Learning (Ai/ml)-based Software as a Medical Device (Samd) Action Plan*. White Oak, MD, USA: US Food Drug Admin. Tech. Rep 145022.

van der List, J. P., Chawla, H., Joskowicz, L., and Pearle, A. D. (2016). Current State of Computer Navigation and Robotics in Unicompartmental and Total Knee Arthroplasty: a Systematic Review with Meta-Analysis. *Knee Surg. Sports Traumatol. Arthrosc.* 24, 3482–3495. doi:10.1007/s00167-016-4305-9

Varnavas, A., Carrell, T., and Penney, G. (2015a). Fully Automated 2D-3D Registration and Verification. *Med. image Anal.* 26, 108–119. doi:10.1016/j.media.2015.08.005

Varnavas, A., Carrell, T., and Penney, G. (2015b). Fully Automated 2d-3d Registration and Verification. *Med. image Anal.* 26, 108–119. doi:10.1016/j.media.2015.08.005

Varnavas, A., Carrell, T., and Penney, G. (2013). *Fully Automated Initialisation of 2d-3d Image Registration*. San Francisco, CA, United states: IEEE Computer Society, 568–571.

Vercauteren, T., Unberath, M., Padoy, N., and Navab, N. (2019). CAI4CAI: The Rise of Contextual Artificial Intelligence in Computer Assisted Interventions. *Proc. IEEE Inst. Electr. Electron. Eng.* 108, 198–214. doi:10.1109/JPROC.2019.2946993

Wang, C., Xie, S., Li, K., Wang, C., Liu, X., Zhao, L., et al. (2020). Multi-view point-based Registration for Native Knee Kinematics Measurement with Feature Transfer Learning.

Wu, J., Fatah, E. E., and Mahfouz, M. R. (2015). Fully Automatic Initialization of Two-Dimensional-Three-Dimensional Medical Image Registration Using Hybrid Classifier. *J. Med. Imaging (Bellingham)* 2, 024007. doi:10.1117/1.JMI.2.2.024007

Wu, J., Su, Z., and Li, Z. (2016). A Neural Network-Based 2d/3d Image Registration Quality Evaluator for Pediatric Patient Setup in External Beam Radiotherapy. *J. Appl. Clin. Med. Phys.* 17, 22–33. doi:10.1120/jacmp.v17i1.5235

Wu, J. Y., Tamhane, A., Kazanzides, P., and Unberath, M. (2021). Cross-modal Self-Supervised Representation Learning for Gesture and Skill Recognition in Robotic Surgery. *Int. J. Comp. Assist. Radiol. Surg.*, 1–9. doi:10.1007/s11548-021-02343-y

Xiangqian, C., Xiaoqing, G., Gang, Z., Yubo, F., and Yu, W. (2020). 2d/3d Medical Image Registration Using Convolutional Neural Network. *Chin. J. Biomed. Eng.* 39, 394–403.

Xie, Y., Meng, C., Guan, S., and Wang, Q. (2017). *Single Shot 2d3d Image Regisraton*, 2018. Shanghai, China: Institute of Electrical and Electronics Engineers Inc, 1–5.

Yang, X., Kwitt, R., Styner, M., and Niethammer, M. (2017). Quicksilver: Fast Predictive Image Registration - A Deep Learning Approach. *NeuroImage* 158, 378–396. doi:10.1016/j.neuroimage.2017.07.008

Yang, Z., and Chen, L. (2019). *A Novel Neurosurgery Registration Pipeline Based on Heat Maps and Anatomic Facial Feature Points*. Huaqiao, China: Institute of Electrical and Electronics Engineers Inc).

Yi, T., Ramchandran, V., Siewerdsen, J. H., and Uneri, A. (2018). Robotic Drill Guide Positioning Using Known-Component 3D-2D Image Registration. *J. Med. Imaging (Bellingham)* 5, 021212. doi:10.1117/1.JMI.5.2.021212

Yokota, F., Okada, T., Takao, M., Sugano, N., Tada, Y., Tomiyama, N., et al. (2013).Automated Ct Segmentation of Diseased Hip Using Hierarchical and Conditional Statistical Shape Models. In International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 190–197. doi:10.1007/978-3-642-40763-5_24

Zaech, J.-N., Gao, C., Bier, B., Taylor, R., Maier, A., Navab, N., et al. (2019).Learning to Avoid Poor Images: Towards Task-Aware C-Arm Cone-Beam Ct Trajectories. In International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 11–19. doi:10.1007/978-3-030-32254-0_2

Zapaishchykova, A., Dreizin, D., Li, Z., Wu, J. Y., Roohi, S. F., and Unberath, M. (2021). An Interpretable Approach to Automated Severity Scoring in Pelvic Trauma.

Zhang, Y., Huang, X., and Wang, J. (2020). Automatic Cone Beam Projection-Based Liver Tumor Localization by Deep Learning and Biomechanical Modeling. *Int. J. Radiat. Oncology*Biology*Physics* 108, S171. doi:10.1016/j.ijrobp.2020.07.946

Zhang, Y., Pei, Y., Qin, H., Guo, Y., Ma, G., Xu, T., et al. (2018). *Temporal Consistent 2d-3d Registration of Lateral Cephalograms and Cone-Beam Computed Tomography Images*, 11046. Granada, Spain: Springer-Verlag, 371–379. doi:10.1007/978-3-030-00919-9_43

Zhao, Q., Chou, C.-R., Mageras, G., and Pizer, S. (2014). Local Metric Learning in 2d/3d Deformable Registration with Application in the Abdomen. *IEEE Trans. Med. Imaging* 33, 1592–1600. doi:10.1109/tmi.2014.2319193

Zheng, J., Miao, S., Jane Wang, Z., and Liao, R. (2018). Pairwise Domain Adaptation Module for Cnn-Based 2-D/3-D Registration. *J. Med. Imaging (Bellingham)* 5, 021204. doi:10.1117/1.JMI.5.2.021204

Zhou, D.-X. (2020). Universality of Deep Convolutional Neural Networks. *Appl. Comput. harmonic Anal.* 48, 787–794. doi:10.1016/j.acha.2019.06.004

Zhu, J., Li, H., Ai, D., Yang, Q., Fan, J., Huang, Y., et al. (2021). Iterative Closest Graph Matching for Non-rigid 3d/2d Coronary Arteries Registration. *Comp. Methods Programs Biomed.* 199, 105901. doi:10.1016/j.cmpb.2020.105901