The Johns Hopkins University
EN.601.656 Advanced Computer-Integrated Surgery Course Project
Instructor: Russell H. Taylor

Real-time Integration of Fully Automatic 2D/3D Pelvic Registration with Robotic X-ray Acquisition

Group 7 Background Readings Summary

Jiaming Zhang
jzhan282@jhu.edu

Zhangcong She
zshe1@jhu.edu

Mentors: Benjamin Killeen; Prof. Mathias Unberath

Contents

1 Introduction

In minimally invasive surgery, clinicians use intraoperative fluoroscopy to overcome the occlusion and ascertain the poses of anatomy, surgical instruments, or artificial implants[1]. Registration is used to align the pre- and intra-interventional fluoroscopy just before and also during an operation in such a way that corresponding anatomical structures in the intraoperative dataset and preoperative dateset are aligned [2]. Amongst all the registration techniques, intraoperative 2D/3D registration is a commonly used technique for finding a rigid pose of a 3D image, for instance, CT Scan so that it aligns with the target 2D image.

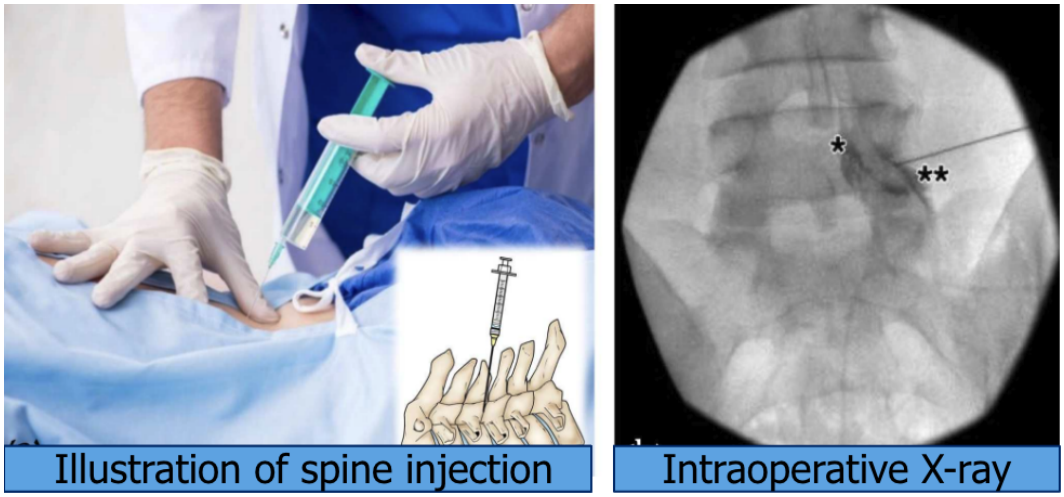

Figure 1: Demonstrates how intraoperative X-ray is applied in sugery[2]

In general, 2D/3D registration aims at finding the 3D pose of shape parameters of an object by optimizing a similarity function of the observed target image and digitally reconstructed radiograph (DRR) image obtained from the 3D volumetric data. The optimization of the similarity function can be formulated as follows:

$$\{\theta_m, m \in \{0, ..., M\}\} = \arg\min_{\theta_m} \sum_{n=0}^{N} \mathcal{S}(I_n, \sum_{m=0}^{M} \mathcal{P}(V_m; \Theta_m)) + \mathcal{R}(\theta_m) \tag{1}$$

Here, V_m means a set of 3D volumetric data and I_m represents 2D X-ray images. θ_m is the object pose. $\mathcal{P}$ means the DRR projection operator; $\mathcal{S}$ means the similarity function; $\mathcal{R}$ stands for regularizer function.

The similarity function $\mathcal{S}$ can be determined by multiple features of the image, including anatomical landmarks, intensity features and gradient features [3]. Grupp utilized landmark positions in the initialization procedure and then solve the optimization prob-

lem using Covariance Matrix Adaptation: Evolutionary Search (CMA-ES), and Bounded Optimization by Quadratic Approximation (BOBYQA) [4] in his implementation of xReg. The pipeline used in xReg depends on the acquisition of landmark positions prior to the optimization process. Traditional landmark detection is suffered from relatively low accuracy and the varying appearance of the same landmark due to the fact that the viewing direction changes substantially between views [5].

Generally, these problems can be potentially solved by a deep neural network. However, acquiring large-scale real clinical X-ray data with expert annotation is particularly challenging due to its incompatibility with the current clinical routine and the time-consuming annotation process [2]. To address this issue, Cong Gao proposed a data synthesis paradigm, which is called SyntheX, and showed that a simulated training set can be a viable replacement for training the medical AI. The pipeline we are implementing mainly relies on a Trans-UNet model to detect the landmark of the X-ray image, and the model is trained over SyntheX. Therefore, we will mainly focus on 2 papers in our report, the first one will be about SyntheX, and the second paper will introduce the techniques used in 2D/3D registration.

2 Paper 1: SyntheX

2.1 Paper Selection and Relavance

Title - SyntheX: Scaling Up Learning-based X-ray Image Analysis Through In Silico Experiments
Author - Cong Gao, Benjamin D. Killeen, Yicheng Hu, Robert B. Grupp, Russell H. Taylor, Mehran Armand, Mathias Unberath

As stated in Introduction, landmark detection is a key step in our workflow. Gao et al. proposed a model transfer paradigm for X-ray image analysis, which is referred as SyntheX. It demonstrates that training learning-based models with synthetic X-ray image dataset is a viable alternative in several surgical tasks. Furthermore, this paper provides a Trans-UNet model that can be directly integrated into our workflow to automatically annotate the raw X-ray images.

2.2 Background

Chen et al. proposed Trans-UNet [6], which is a powerful neural network for medical image processing. By applying Trans-UNet, we can easily annotate the X-ray images automatically. However, developing the AI backbones of such systems currently depends on collecting training data during routine surgeries. This remains one of the largest barriers to the widespread use of AI systems in interventional clinical settings, versus triage or diagnostic settings, as the acquisition and annotation of interventional data is time intensive and costly [1]. A promising alternative to these strategies is a simulation, i.e., the *in silico* generation of synthetic interventional training data and imagery from human models.

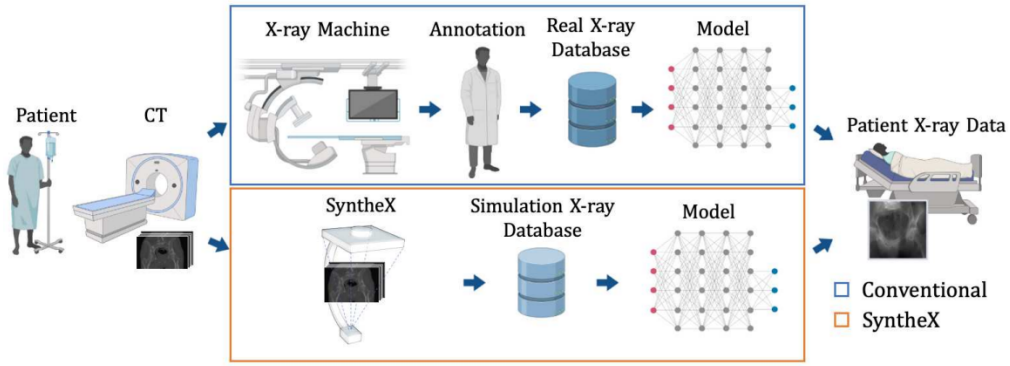

Figure 2: SyntheX Workflow [1]

The overall scheme for SyntheX has demonstrated in Fig 2. This paper wants to show that training the deep learning models on synthetic datasets is a possible choice for replacing realistic datasets, especially in three scenarios, namely, hip-imaging, surgical tool detection, and lesion segmentation.

2.3 Technical Summary

Essentially, to lessen the need of real clinical data, the simulation process has to be as real as possible. However, there exists a huge domain gap between the realistic data and the simulated data mostly because the clinical images are always not perfectly taken. To mitigate the domain gap between realistic data and simulated data, several approaches have been tested in this paper. The technique for mitigating the domain gaps is called domain generalization. Domain randomization and domain adaptation are two major methods of generalization. Other than domain randomization which does not assume knowledge or sampling of the target domain at training time, domain adaptation techniques attempt to

mitigate the domain gap's detrimental effect by aligning features across the source (training domain, here: simulated data) and the target domain (deployment domain, here: real X-ray images). As such, domain adaptation techniques require samples from the target domain at training time. Recent domain adaptation techniques have increased the suitability of the approach for Sim2Real transfer because they now allow for the use of unlabeled data in the target domain. We conducted experiments using two common domain adaptation methods: CycleGAN and adversarial discriminative domain adaptation (ADDA). The two methods are similar in that they attempt to align properties of real and synthetic domains and differ based on what properties they seek to align. The overall objective for CycleGAN training is expressed as:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_c(G, F) \qquad (2)$$

Although the state-of-art approaches are mainly focusing on domain adaptations, this paper proves that strong domain randomization is as good as adaptation in the test cases. The way of proving that hypothesis is to repeat the same task with data sets generated with different domain generalization methods and directly compare their Dice score with the ground truth. The domain randomization methods that are evaluated in this paper includes: 1) Gaussian noise injection; 2) Gamma Transform; 3) Random Crop; 4) Inverting; 5) Affine Transform; 6) Contrast; 7) Blurring. Method 1 to 3 is referred as basic randomization and 4 to 7 is called as strong randomization. For each image, the author first applies basic randomization and then randomly concatenates two strong randomization methods. The details will be described in the experiments section.

2.4 Experiments

The model is trained and evaluated on 3 types of combinations of the datasets, namely Sim2Sim, Real2Real, and Sim2Real. Sim2Sim stands for the case that the model is trained and evaluated both on the synthetic images, mainly to test the functionality. Real2Real means the model is trained on the real image and evaluated on the real image, this paradigm is used as the baseline for the model. Sim2Real represents that the training set comes from synthetic images, and the model is evaluated on the real x-ray. This paradigm is mainly used to compare the performance of different domain generalization methods.

The real images are collected from three different surgical applications, namely hip imaging task, surgical tool detection task, and lung segmentation. For the hip imaging task, the realistic dataset contains High-Res CT scans from 6 different cadaveric pelves, and 366 corresponding X-ray images. The simulated dataset contains 20,000 Digitally

Reconstructed Radiographs (DRR) generated from 20 CT scans. 3 DRR methods are applied in this process, namely naiveDRR, xreg DRR and DeepDRR. A comparison between these methods is shown in Fig 3.

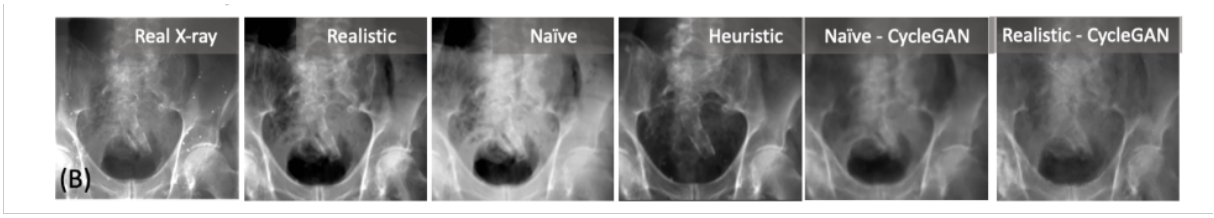

Figure 3: Comparison between different DRR and domain adaptation methods [1]

2.5 Experiment Results

For the hip imaging task, the author provides Segmentation Dice Score as a mean of 5-fold individual testing on 366 real hip X-ray images. The Dice score ranges from 0 to 1, with larger values corresponding to better segmentation performance.

Training Data	180x180		360x360		480x480	
Domain	regular DR	strong DR	regular DR	strong DR	regular DR	strong DR
RealXray (*Real2Real*)	0.775 ± 0.235	0.784 ± 0.214	0.783 ± 0.232	0.759 ± 0.248	0.739 ± 0.266	0.751 ± 0.265
Realistic	0.730 ± 0.240	0.787 ± 0.211	0.751 ± 0.241	0.760 ± 0.250	0.720 ± 0.256	0.700 ± 0.279
Heuristic	0.669 ± 0.273	0.737 ± 0.249	0.683 ± 0.265	0.682 ± 0.286	0.655 ± 0.277	0.668 ± 0.298
Naïve	0.689 ± 0.256	0.680 ± 0.278	0.687 ± 0.266	0.572 ± 0.309	0.653 ± 0.278	0.578 ± 0.305
Realistic-Cyc	0.778 ± 0.217	0.778 ± 0.220	0.760 ± 0.248	0.733 ± 0.267	0.741 ± 0.255	0.688 ± 0.291
Naïve-Cyc	0.784 ± 0.198	0.750 ± 0.230	0.739 ± 0.254	0.736 ± 0.258	0.726 ± 0.254	0.673 ± 0.292
Realistic-ADDA	0.767 ± 0.224	0.754 ± 0.231	0.726 ± 0.292	0.731 ± 0.256	0.704 ± 0.279	0.727 ± 0.256
Naïve-ADDA	0.491 ± 0.405	0.678 ± 0.266	0.693 ± 0.297	0.662 ± 0.265	0.693 ± 0.273	0.592 ± 0.306
Realistic-Scaled	$\mathbf{0.857 \pm 0.184}$	$\mathbf{0.853 \pm 0.179}$	$\mathbf{0.838 \pm 0.221}$	$\mathbf{0.818 \pm 0.221}$	0.783 ± 0.262	$\mathbf{0.823 \pm 0.221}$
Realistic-Cyc-Scaled	0.822 ± 0.213	0.794 ± 0.232	0.824 ± 0.225	0.789 ± 0.240	$\mathbf{0.789 \pm 0.241}$	0.812 ± 0.227

Figure 4: Segmentation Dice Score of different methods [1]

Note that the best performance results are bolded. Training/testing image resolutions are listed in the top row. DR is short for domain randomization. Cyc is short for CycleGAN. ADDA refers to adversarial discriminative domain adaptation. "-Scaled" means training on the scaled-up dataset.

Also, Landmark Detection Errors (mm) at 90% activation percentage as a mean of 5-fold individual testing on 366 real hip X-ray images. Lower values are better.

By inspection of the results, the author concluded that these findings suggest that scaling-up data for training is an effective strategy to improve performance both in- and

Training Data	180x180		360x360		480x480	
Domain	regular DR	strong DR	regular DR	strong DR	regular DR	strong DR
RealXray (*Real2Real*)	9.98 ± 22.58	7.78 ± 11.94	8.93 ± 19.76	8.15 ± 15.30	8.98 ± 21.38	7.59 ± 16.12
Realistic	14.33 ± 32.61	8.41 ± 14.47	12.62 ± 27.68	9.05 ± 19.37	14.96 ± 34.45	13.06 ± 26.52
Heuristic	20.84 ± 44.16	10.69 ± 22.92	14.54 ± 32.88	12.25 ± 26.82	17.59 ± 39.55	12.85 ± 29.49
Naïve	13.12 ± 30.87	13.03 ± 21.20	16.22 ± 39.71	20.55 ± 35.92	18.53 ± 40.88	19.46 ± 37.48
Realistic-Cyc	8.65 ± 19.84	8.19 ± 13.55	8.57 ± 18.69	8.90 ± 17.47	10.78 ± 26.9	8.75 ± 17.4
Naïve-Cyc	8.56 ± 18.46	9.05 ± 10.75	7.63 ± 16.10	9.29 ± 18.73	9.14 ± 21.82	11.73 ± 25.06
Realistic-ADDA	11.19 ± 24.80	11.43 ± 25.83	11.01 ± 24.69	12.92 ± 23.76	16.42 ± 37.61	9.24 ± 17.53
Naïve-ADDA	7.90 ± 17.56	11.84 ± 25.63	10.42 ± 25.05	14.17 ± 30.89	16.62 ± 40.12	22.88 ± 41.53
Realistic-Scaled	$\mathbf{5.91 \pm 8.43}$	$\mathbf{6.06 \pm 7.10}$	$\mathbf{6.80 \pm 9.53}$	$\mathbf{6.29 \pm 6.29}$	8.12 ± 19.35	5.99 ± 12.19
Realistic-Cyc-Scaled	6.79 ± 9.70	6.87 ± 9.58	6.87 ± 13.19	6.59 ± 7.25	$\mathbf{6.43 \pm 13.67}$	$\mathbf{5.52 \pm 4.85}$

Figure 5: L2 norm error of the landmark detection of different methods [1]

outside of the training domain. Scaling up training data is costly or impossible in real settings, but in comparison is easily possible using data synthesis. Having access to more varied data samples during training helps the network parameter optimization find a more stable solution, that also transfers better.

2.6 Assessment

In summary, this paper proposed a viable model transfer paradigm for training the learning-based models on synthetic datasets. It contains a thorough and detailed description of the testing methodology. Besides, it clears the path of evaluating different models for medical imaging tasks and makes it easy to follow. The author also released the python package that he developed for this paper. The package is well-implemented and convenient to integrate into other medical image processes.

On the other hand, although the proposed paradigm outperforms the state-of-the-art AI models, the author didn't mention the results of traditional annotation approaches. As a matter of fact, this AI-based landmark detection model is fundamentally imprecise compared to the existing annotation algorithms, making it insufficient for our project. We cannot simply rely on landmarks only to perform 2D/3D registration, which leads us to the second paper, where a combinatorial method is introduced.

3 Paper 2: Intensity-based 2D-3D Registration

3.1 Paper Selection and Relavance

Title - Robust Patella Motion Tracking using Intensity-based 2D-3D Registration on Dynamic Bi-plane Fluoroscopy: Towards Quantitative Assessment in MPFL reconstruction surgery.
Author - Otake, Yoshito; Esnault, Matthieu; Grupp, Robert; Kosugi, Shinichi.

This paper demonstrated a precise and completed example about how to conduct a 2D-3D intensity-based registration and how to evaluate the result registration based on gradient correlation. one of our critical component, Xref, is based on this robust, global, optimization component with various strategies. Furthermore, the technique addresses the weakness of SyntheX that focused on detecting landmarks. Instead of treating landmark features and intensity features separately, the detected landmark locations may be incorporated into a robust reprojection regularizer for intensity-based registration.

3.2 Background

In this paper, the authors mainly utilize intensity-based registration to register 2D fluoroscopic images and 3D CT scans. A precise registration result can provide valuable intraoperative feedback to the surgeon, especially surgeries such as MPFL reconstruction surgery. Medial patellofemoral ligament, known as MPFL, is a surgery that stabilize the kneecap and help protect the joint. Previous research has centered around the use of low-resolution MRI, which lacks the necessary framerate to accurately capture fluid motion in the knee. Additionally, computer-assisted manual registration has been utilized, but this method is hindered by the limitations of human ability to consistently detect sub-millimeter alignments in complex anatomy [7]. Therefore, the authors proposed to use a 2D-3D registration with robust, global, optimization components to improve the success rate of MPFL reconstruction surgery.

3.3 Technical Summary

Overall, there are three bones registration, femur, tibia-fibula and patella. The process of registration can be broken down into several steps. Initially, a rough transformation is manually estimated and used for the first image's transformation. Then, the first frame's registration result would act as an initial guess for subsequent frames. The Covariance Ma-

trix Evolution Strategy (CMA-ES) optimization method is utilized in an objective function to compute Digital Reconstructed Radiographs (DRRs) of bone volume from CT scans. The accuracy of registrations is evaluated by the sum of the gradient correlation between each DRR and two corresponding fluoroscopic images, where a high value of gradient correlation means precise registration.

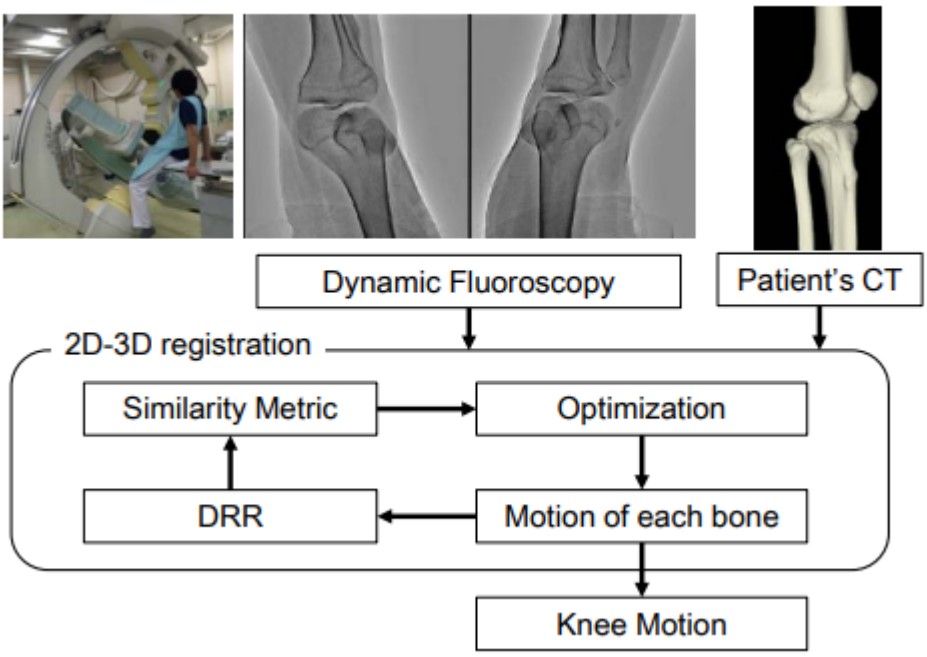

Figure 6: Overview of Technical Approach[7]

3.4 Experiments

There are four registration approaches that are tested and compared. For individual bone registration, every three bones are registered. Similarly, sequential registration is also carried out using three separate registrations, but the final registration of each bone is used for registering subsequent bones. The bones are registered in the following order: femur, tibia-fibula, and patella. Simultaneous registration is another approach that registers all three bones at the same time. The combination of sequential and simultaneous registrations involves first performing the sequential registration and using the result as an initial guess for the final simultaneous registration.

Each registration is evaluated in two scenarios, experiments with simulated images and with measured fluoroscopic images. In the first part, the authors evaluated different

registration methods by simulating fluoroscopic images using DRRs from a CT scan of a healthy knee. The initial guess for each trial was perturbed to simulate errors in manual initialization. The first frame registration error depended on the quality of the initial guess, and robustness was evaluated on only the first frame. Different frames from the simulated motion were tested, and randomized perturbations were added to the ground truth transformation to create the initial guess. In the second part, four approaches were tested using measured fluoroscopy images with a "ground truth" registration for each bone estimated using the simultaneous approach starting from an initial transformation carefully set to the visually interpreted solution. Fifty registration trials for all four registration approaches were conducted with an initialization randomly perturbed from the ground truth, and the error from the ground truth was computed.

3.5 Experiment Results

A threshold of 2mm for translation error and 2° for rotation error was used to define a successful registration [7]. In the experiment of simulated images, the individual bond approach had 87% success rate for the patella, while the success rate for all other cases is 100%.

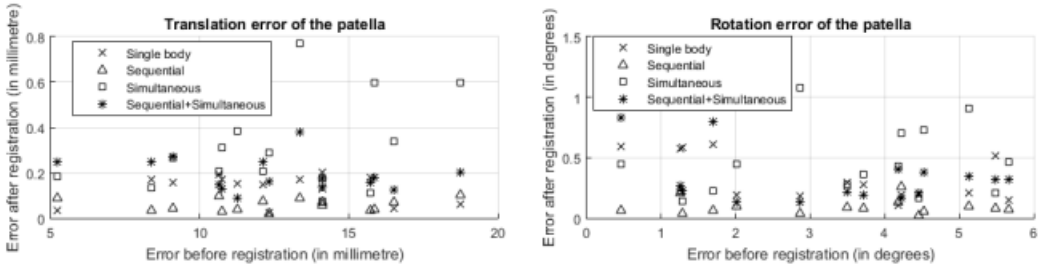

Figure 7: Individual and simultaneous registration methods in simulation method [7]

In the experiments with measured fluoroscopic images, all femur and tibia-fibula cases are successful. For patella cases, four approaches showed 38%, 74%, 70%, 68% success rates.

3.6 Assessment

The paper being discussed presents a novel method for obtaining 3D motion tracking of multiple rigid objects. This new approach utilized four different registration methods, including sequential, simultaneous, and a combination of sequential and simultaneous

approaches, all of which perform quite well. This new technique also offers significant improvements in robustness compared to traditional methods. However, despite these advantages, there are also some drawbacks to the approach presented in the paper. One of the major limitations of the study is that it was conducted using a very limited number of trials in both scenarios. This means that the results may not be generalizable to larger populations or more complex situations.

While the authors do clarify that they used the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) during their technique approach, they do not provide a detailed discussion of the input and output of CMA-ES. This omission can make it difficult for readers to fully understand how the technique works and how it produces its results. Furthermore, the authors describe that digitally reconstructed radiographs (DRR) can be computed based on the output of CMA-ES. However, the precise relationship between CMA-ES, DRRs, and the overall technique approach is not clearly defined in the paper, which can make it challenging for readers to fully grasp the methodology being employed.

In conclusion, this paper presents a concise and comprehensive 2D/3D registration method based on intensity, which was applied in the Xreg program. The method is illustrated by evaluating the gradient correlation between fluoroscopic images and DRRs of bone volume from CT. While the previous paper focuses on landmark features, this paper mainly discusses the intensity features of registration. By combining the advantages and limitations of both methods, our team gains a better understanding of how to use them together to create a more robust reprojection regularizer for registration.

References

[1] Cong Gao, Benjamin D. Killeen, Yicheng Hu, Robert B. Grupp, Russell H. Taylor, Mehran Armand, and Mathias Unberath. SyntheX: Scaling Up Learning-based X-ray Image Analysis Through In Silico Experiments. *arXiv e-prints*, page arXiv:2206.06127, June 2022.

[2] Cong Gao. *Fluoroscopic navigation for robot-assisted orthopedic surgery.* PhD thesis, Johns Hopkins University, 2022.

[3] P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš. A review of 3d/2d registration methods for image-guided interventions. *Medical Image Analysis*, 16(3):642–661, 2012. Computer Assisted Interventions.

[4] Robert B. Grupp, Mathias Unberath, Cong Gao, Rachel A. Hegeman, Ryan J. Murphy, Clayton P. Alexander, Yoshito Otake, Benjamin A. Mcarthur, Mehran Armand, and Russell H. Taylor. Automatic annotation of hip anatomy in fluoroscopy for robust and efficient 2d/3d registration. *International Journal of Computer Assisted Radiology and Surgery*, 15(5):759–769, 2020.

[5] Bastian Bier, Florian Goldmann, Jan-Nico Zaech, Javad Fotouhi, Rachel A. Hegeman, Robert Grupp, Mehran Armand, Greg M. Osgood, Nassir Navab, Andreas K. Maier, and M. Unberath. Learning to detect anatomical landmarks of the pelvis in x-rays from arbitrary views. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–11, 2019.

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv e-prints*, page arXiv:2102.04306, February 2021.

[7] Yoshito Otake, Matthieu Esnault, Robert Grupp, Shinichi Kosugi, and Yoshinobu Sato. Robust patella motion tracking using intensity-based 2D-3D registration on dynamic bi-plane fluoroscopy: towards quantitative assessment in MPFL reconstruction surgery. 9786:97860B, March 2016.