Intraoperative 2D/3D Registration via Spherical Similarity Learning and Differentiable Levenberg-Marquardt Optimization

Supplementary Material

In this supplementary material, we first provide the preprocessing details of the X-ray data used in the experiments in Sec. 6. Following that, we include the detailed descriptions of three datasets in Sec. 7. Sec. 8 describes the implementation details of the proposed method. Finally, in Sec. 9,we provide the details of domain randomization strategy we used in the this work.

6. Preprocessing of Real X-ray Data

As mentioned in numerous prior studies [18, 44, 60], X-ray imaging captures the intensity attenuation of X-rays as they traverse a medium, whereas DRRs quantify the absorption of X-ray energy by the medium. Therefore, in DRRs, higher-density bone tissue appears brighter, while lower-density soft tissue appears relatively darker. This contrast is inverted in X-ray images, where bone structures appear darker due to greater X-ray absorption, while soft tissue appears brighter due to lower attenuation. To ensure that the grayscale distribution of the DRR generated using Siddon's method [56] aligns with that of X-ray images governed by the Beer-Lambert law, we applied the following inversion strategy:

$$\tilde{I} = 1 - \frac{\log(1+I)}{\log(1+I_0)} \tag{17}$$

where I_0 represents the initial energy of the radiation beam, and all X-ray image pixel values are normalized to their maximum intensity. The transformation $\log(1+I)$ ensures numerical stability by preventing negative logarithmic values while enhancing visual contrast. Additionally, it promotes a Gaussian-like intensity distribution, improving the consistency of the processed images.

7. Datasets

We perform experiments on three publicly available datasets to evaluate the effectiveness of the proposed method:

- **DeepFluoro**: The DeepFluoro dataset contains six pelvic CT scans with a total of 366 real fluoroscopic X-ray images [22]. It provides the calibrated intrinsic matrix of the C-arm imaging system along with the ground truth extrinsic matrix for each X-ray image. Additionally, 14 manually annotated landmarks are available for each CT scan.
- **Ljubljana**: A clinical dataset [48] contains ten clinical 3D Cone-Beam Computed Tomography (CBCT) subtracted angiography scans of patients undergoing cerebral endovascular treatments. Each CBCT scan is provided with two corresponding X-ray images. Calibrated intrinsic and extrinsic matrices are provided.
- **CTSpine1k**: The CTSpine1K dataset [10] consists of 1005 CT volumes with segmentation masks collected from four different open sources, capturing diverse appearance variations. After quality control, 651 CT scans were retained. Following segmentation and denoising preprocessing, only the lumbar spine was extracted. The contour points of the spine surface were obtained by segmentation using Hounsfield unit (HU) threshold, morphological refinement, and 3D surface extraction using the Marching Cubes algorithm. The automatic landmark identification follows a strategy similar to [4, 65]. For each CT scan, 500 DRR images with random poses were generated using [60]. The intrinsic matrix used in the simulation environment is calibrated to match the settings of a mobile C-arm imaging device.

8. Implementation Details

For patient-specific 2D/3D registration, we selected ResNet-18 as the backbone for the pose regressor $\mathcal{R}(\cdot)$, using the Adam optimizer with a learning rate of 1e-3 and a cosine scheduler. The spherical similarity network, used for subsequent fine-tuning, was trained with the AdamW optimizer, also with a learning rate of 1e-3, but utilizing a cyclic scheduler. The pose regressor and similarity network were trained in a self-supervised manner using 500k and 200k synthetic images generated in real-time, respectively. For the patient-agnostic scenario, the parameter settings remained largely the same, except that the pose regressor was replaced with RTPIv3 [8], which incorporates spatial information from CT. Additionally, the number of self-supervised training iterations was increased to 750k for the pose regressor and 300k for the similarity network, to better generalize across different subjects. The termination criterion for the differentiable LM optimization is set such that the standard deviation of $\Delta\theta_i$ over the last ten iterations is less than 1e-2. This threshold was determined through grid search across all datasets to ensure an optimal balance between convergence stability and computational efficiency. Our experiments were conducted on a PC equipped with an NVIDIA RTX 3090 GPU and an Intel Core i7 CPU. Due to memory limitations, the batch sizes for training the pose regressor and the spherical similarity network were set to 4 and 1, respectively.

9. Domain Randomization Strategy

Similar to [18, 20, 64], to alleviate the domain gap between synthetic image and real X-ray image, domain randomization strategy is adopted during training:

- **Image Smoothing**: Perform random smoothing on the image with a kernel size of 3×3 or 5×5, selected with a probability of 50%.
- **Noise Injection**: Inject Gaussian noise into the image with a mean sampled uniformly from $[-0.15 \cdot \max, 0.1 \cdot \max]$.
- **Normalization**: Apply lower and upper bound normalization with intervals sampled as $[-0.04 \cdot \max, 0.02 \cdot \max]$ and $[0.9 \cdot \max, 1.05 \cdot \max]$, respectively.
- **Linear Scaling**: Scale the intensity linearly, with the scaling factor sampled uniformly from $[0.9, 1.05]$.
- **Gamma Adjustment**: Perform gamma correction, with the γ value sampled uniformly from $[0.7, 1.3]$.
- **Nonlinear Scaling**: Scale the image nonlinearly using the function $a \cdot \sin(b \cdot x + c)$, where a and b are sampled uniformly from the range $[0.8, 1.1]$, and c is sampled uniformly from $[-0.5, 0.4]$.
- **Random Erasing** [68]: Randomly erases a rectangular region of the image with an area uniformly sampled from $[0.02 \cdot \text{area}, 0.4 \cdot \text{area}]$, and an aspect ratio sampled uniformly from $[0.3, 1]$, filling the region with the mean intensity of the whole image.