Supplementary Material
Self-Supervised 2D/3D Registration for X-Ray to CT Image Fusion

Our supplementary material provides further analysis of experiments in Appendix A, where we compare our simulated baseline without domain randomization. Additionally, we illustrate the domain gap between DRR and X-ray images, followed by visualizations of the registration error distribution for different variations of our proposed self-supervised framework. In Appendix B we visualize additional samples comparing the overlays produced by the different state-of-the-art methods considered (Figure 4) and data samples from our clinical CBCT reconstruction dataset (Figure 5). We provide further implementation details in Appendix C.

A. Further Analysis of Experiments

A.1. Simulated Baseline

We compare the simulated DIRN trained without domain randomization in Table 1, evaluated on real X-ray images of our test dataset. The SR drops from 66.2% to 10% for the network trained without domain randomization (using bone projection style DRR). Domain randomization significantly improves the performance on real X-ray images, as they have seen different styles during training. Thus, enabling us to have a strong baseline for the comparison with our proposed self-supervised framework.

	mRPD [mm] ↓	SR[%] ↑
Simulated	2.97 ± 0.99	66.2
- DR	3.78 ± 0.83	10.0

Table 1: Comparison of Simulated DIRN with and without domain randomization evaluated on test dataset with real X-ray images. The simulated is our baseline which includes domain randomization and -DR indicates without domain randomization.

A.2. DRR to X-ray Domain Gap

We evaluated our simulated DIRN (includes DR) on the real X-ray and DRR images for the same start positions from our test dataset to illustrate the domain gap. As illustrated in Table 2, we achieve similar results to DIRN [3] when the source and target domain are same (DRR Eval). There is a huge drop in performance of our simulated DIRN when evaluated on real X-ray images (X-ray Eval). In Figure 1, we plot the cumulative registration error distribution for DRR and X-ray Eval of our simulated baseline network. The shift of the registration error towards higher values for X-ray Eval from the DRR Eval clearly illustrates the domain gap that exists even after the application of domain randomization.

	mRPD [mm] ↓	SR[%] ↑
DRR Eval	0.27 ± 0.60	99.3
X-ray Eval	2.97 ± 0.99	66.2

Table 2: Comparison of simulated DIRN (includes DR) evaluated on DRR (DRR Eval) and real X-ray images (X-ray Eval) from our test dataset for the same start positions.

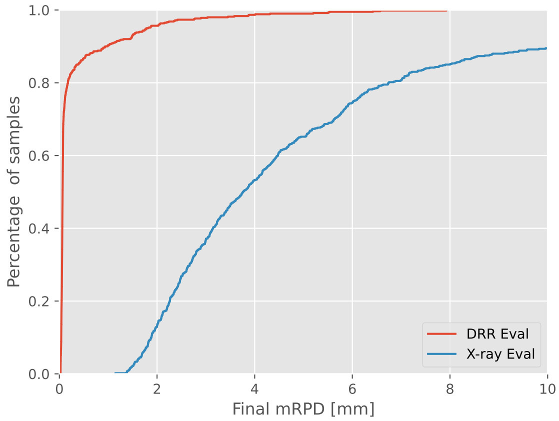

Figure 1: Comparison of final registration error using empirical cumulative distribution for DRR Eval and X-ray Eval of our simulated DIRN (includes DR), indicating the large domain gap that exists even after the application of domain randomization.

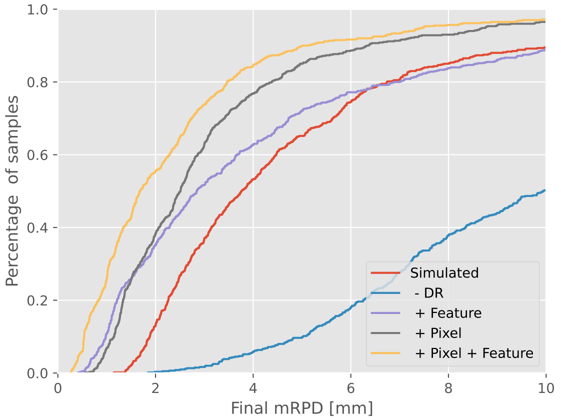

Figure 2: Comparison of final registration error using empirical cumulative distribution for different variations of our proposed framework.

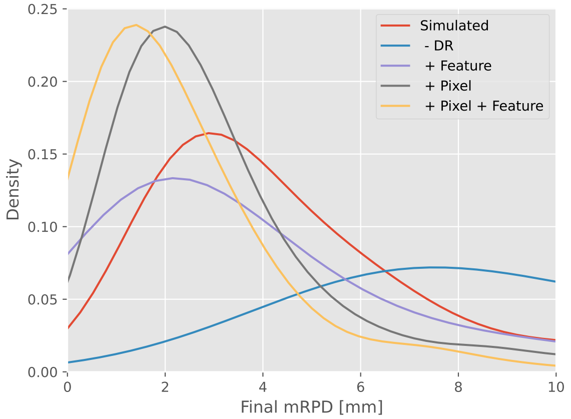

Figure 3: Comparison of final registration error using kernel density estimation for different variations of our proposed framework.

A.3. Ablation of Domain Adaptation Components

We visualize the cumulative distribution and kernel density distribution of the final registration error in Figure 2 and Figure 3 respectively for different variations of our framework. Our proposed framework shows a significant shift to lower registration error compared to the simulated baseline. The standalone feature and pixel space additions also illustrate the performance gains of each component.

A.4. Statistical Significance

We also performed a statistical significance test using a t-Test on the final mRPD achieved for all our experiments against our proposed method. We achieve statistically significant results for all the cases with a p-value < 0.001.

B. Visualization

B.1. Comparison of Registered Overlays

Figure 4 shows additional examples from our test dataset, comparing the overlays produced. Each row depicts the comparison of the overlay produced from a single test sample for the different methods considered.

B.2. Dataset Visualization

Figure 5 shows example images from our clinical CBCT reconstruction dataset which includes the CBCT reconstructed volume along with the paired X-ray images.

C. Implementation details

C.1. Image Preprocessing

The input images $\mathbf{I}$ (includes simulated $\mathbf{I}^s$ and real $\mathbf{I}^r$ X-ray images) are center cropped to a size of 480×480 from original image size of 640×480. The center cropped image is resized to 256×256 and fed as input to the networks. We normalize the pixel values using the dataset mean and standard deviation.

C.2. Network Architecture Details

Our self-supervised network consists of the registration network DIRN [3], feature adaptation components (Adversarial Feature Encoders and Barlow Twins [9]), and the unsupervised style transfer network [5]. We use the architecture proposed in the respective original works, with the specific configuration used for our framework described below. The registration network DIRN [3] consists of RAFT [7] architecture for estimating the correspondence between the fixed $\mathbf{I}_f$ and moving $\mathbf{I}_m$ images. The RAFT architecture consists of a feature encoder and a context encoder. We input $\mathbf{I}_m$ to the context encoder and perform no domain adaptation since $\mathbf{I}_m$ is the fixed style bone projection DRR for both training and evaluation. We perform all the domain adaptations on the feature encoder as we would like to replace the simulated images $\mathbf{I}_f$ with the real X-ray images $\mathbf{I}_f^r$ during evaluation. Both the feature and context encoder are based on ResNet blocks [7]. The encoded feature map from the feature encoder is of the size $[256, 32, 32]$ for both $\mathbf{I}_m$ and $\mathbf{I}_f$. The RAFT uses iterative residual flow estimation for training and evaluation. We set the number of iterations for flow estimation to 6 for both training and evaluation. We use the PointNet++ architecture [6] for correspondence weighting as proposed in DIRN [3]. The single scale grouping-based segmentation architecture of PointNet++ which can output per-point classification is used. We replace the final layer with a Sigmoid activation function for

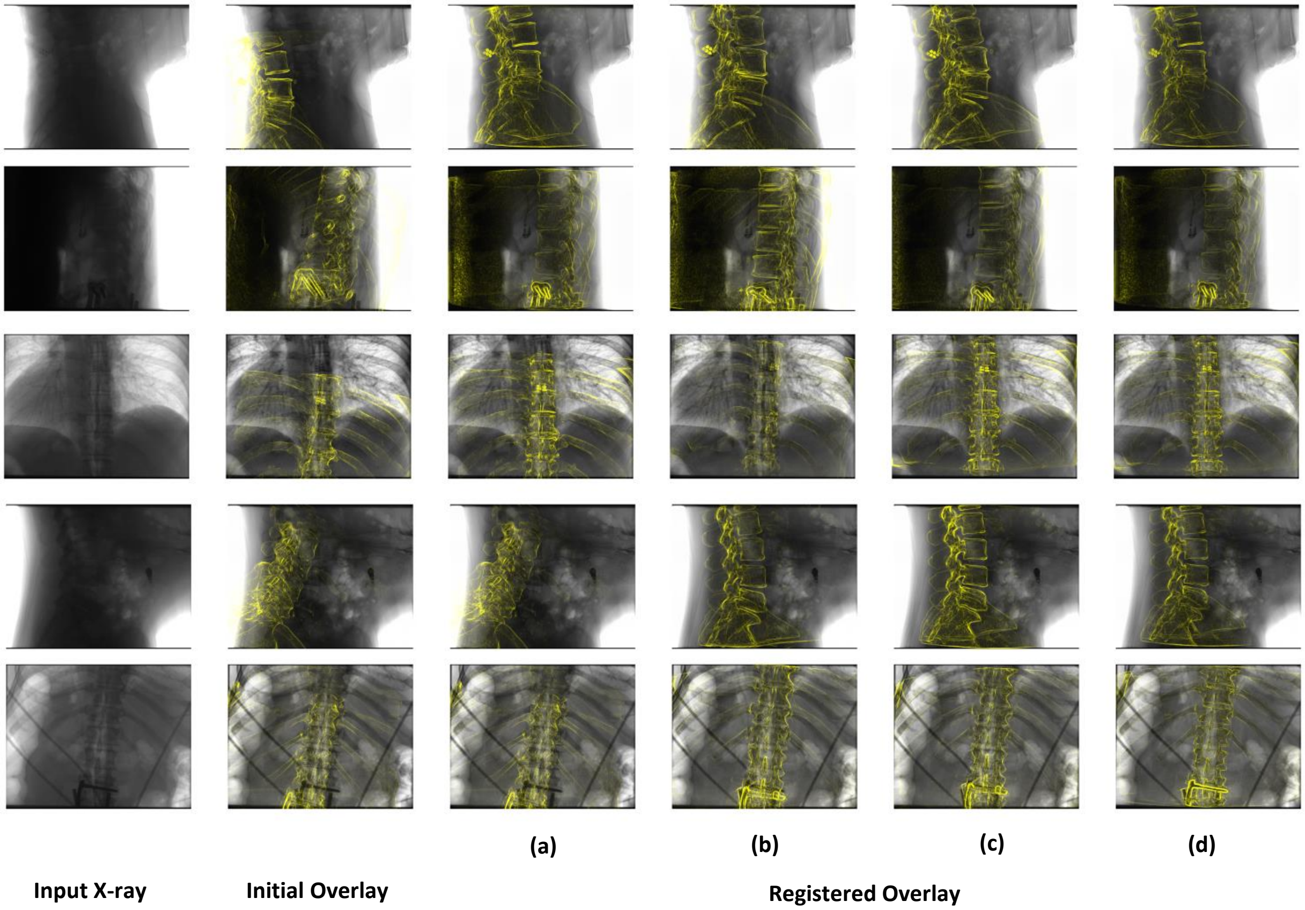

Figure 4: Additional samples from test dataset with comparison of overlays produced using (a) Optimization-based technique [8], (b) Simulated (with domain randomization [1]), (c) our proposed method, and (d) supervised [3]. Each row represents a data sample from the test dataset.

predicting per-point weights in the range of $[0, 1]$. The feature projector of the Barlow Twins consists of an MLP with three hidden layers of size $[512, 256, 128]$ that projects the encoded feature maps to 128-dimension embedding vector $\mathbf{Z}$. The feature discriminator of the adversarial feature encoder uses patch GAN [2] with a patch size of 8 and input channel dimension of 64. We use 1×1 convolution to match the encoded feature map to the input channel dimension of the patch GAN discriminator. The unsupervised style transfer network based on CUT [5] uses a ResNet based generator consisting of 9 residual blocks [4] and patch GAN discriminator [2], with a patch size of 16.

References

[1] Matthias Grimm, Javier Esteban, Mathias Unberath, and Nassir Navab. Pose-Dependent Weights and Domain Randomization for Fully Automatic X-Ray to CT Registration. *IEEE Transactions on Medical Imaging*, 40(9):2221–2232, 2021.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[3] Srikrishna Jaganathan, Jian Wang, Anja Borsdorf, Karthik Shetty, and Andreas Maier. Deep iterative 2d/3d registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 383–392. Springer, 2021.

[4] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution.

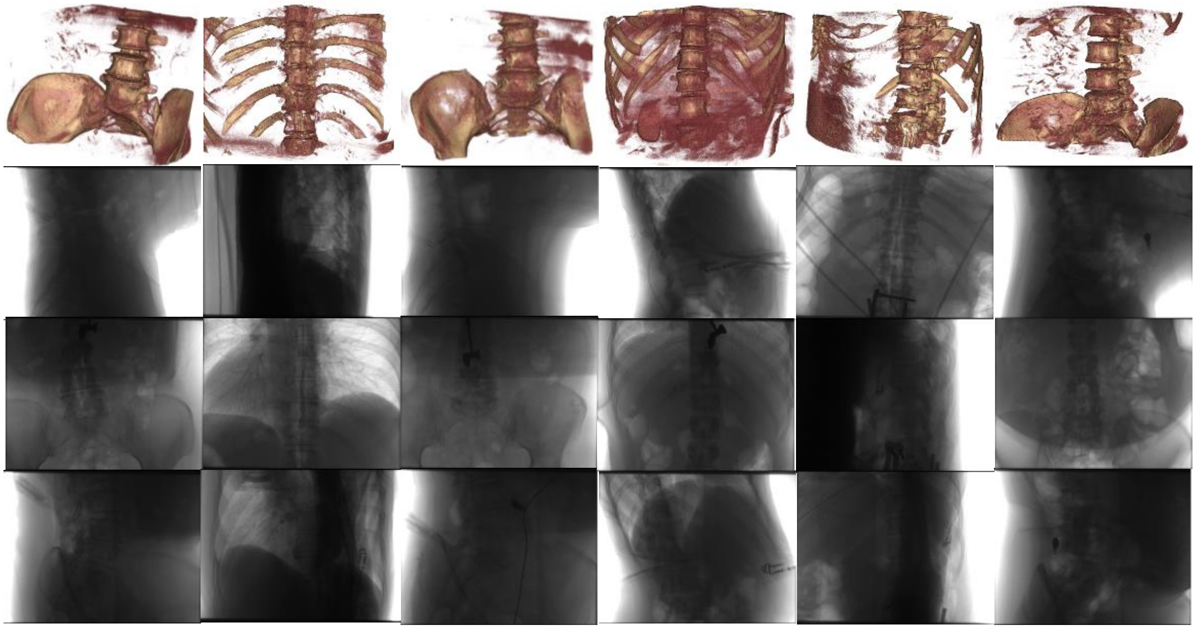

Figure 5: Exemplar data samples from our clinical CBCT dataset. The reconstructed volume is thresholded to better visualize the bone contours. Each column represents a reconstructed CBCT volume along with paired set of X-ray images used in reconstructing the CBCT volume.

In *European conference on computer vision*, pages 694–711. Springer, 2016.

[5] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020.

[6] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[7] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[8] Jian Wang. *Robust 2-D/3-D Registration for Real-time Patient Motion Compensation: Robuste 2-D/3-D Registrierung Zur Echtzeitfähigen, Dynamischen Bewegungskompensation*. Verlag Dr. Hut, 2020.

[9] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.