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

Srikrishna Jaganathan[1,2] Maximilian Kukla[2] Jian Wang[2] Karthik Shetty[1] Andreas Maier[1]
[1]FAU Erlangen-Nürnberg, Erlangen, Germany [2]Siemens Healthineers AG, Forchheim, Germany
srikrishna.jaganathan@fau.de

Abstract

Deep Learning-based 2D/3D registration enables fast, robust, and accurate X-ray to CT image fusion when large annotated paired datasets are available for training. However, the need for paired CT volume and X-ray images with ground truth registration limits the applicability in interventional scenarios. An alternative is to use simulated X-ray projections from CT volumes, thus removing the need for paired annotated datasets. Deep Neural Networks trained exclusively on simulated X-ray projections can perform significantly worse on real X-ray images due to the domain gap. We propose a self-supervised 2D/3D registration framework combining simulated training with unsupervised feature and pixel space domain adaptation to overcome the domain gap and eliminate the need for paired annotated datasets. Our framework achieves a registration accuracy of 1.83 ± 1.16 *mm with a high success ratio of 90.1% on real X-ray images showing a 23.9% increase in success ratio compared to reference annotation-free algorithms.*

1. Introduction

Image guidance for minimally invasive interventions is generally provided using live fluoroscopic X-ray imaging. The fusion of preoperative Computed Tomography (CT) volume with the live fluoroscopic image enhances the information available during the intervention. Spatial alignment of the 3D volume on the current patient position is a prerequisite for accurate fusion with the fluoroscopic image. An optimal spatial alignment between preoperative CT volume and live fluoroscopic X-ray is estimated with 2D/3D registration. Traditionally, optimization-based techniques have been used for 2D/3D registration in the interventional setting as it provides highly accurate registration [53, 29, 51]. However, optimization-based techniques are sensitive to initialization and content mismatch between X-ray and CT images. Deep Learning (DL)-based 2D/3D registration techniques have been proposed to overcome the limitations of the optimization-based techniques by improving the robustness significantly [26, 31, 32, 39], while still relying on optimization-based techniques as a subsequent refinement step to match the registration accuracy. Recently, end-to-end DL-driven solutions have been proposed that can achieve a combination of high registration accuracy and high robustness with faster computation [20].

Despite the significant improvement in learning-based registration techniques, the interventional application is still limited due to the lack of generalizability of the learned networks for different anatomy, interventions, scanner, and protocol variations [47]. The collection of large-scale annotated datasets for all variations is prohibitive since the data needed for training should be paired along with ground truth registration. Either a large-scale annotated dataset that consists of all the different variations or an annotation-free unpaired training routine based on existing DL-based technique enables us one step closer to interventional application. We focus on the latter, by removing the need for annotated paired dataset as this would immediately allow us to train the current state-of-the-art registration networks for different variations.

We propose a self-supervised 2D/3D rigid registration framework to achieve annotation-free unpaired training with minimal performance drop on real X-ray images encountered during the interventional application. The annotation-free unpaired dataset is generated from forward projections of the CT volumes. Our framework consists of simulated training combined with unsupervised feature and pixel space domain adaptation. Our novel task-specific feature space domain adaptation is trained in an end-to-end manner with the registration network. We combine the recently proposed Barlow Twins [56], adversarial feature discriminator [7, 22] and DL-based registration network [20]. This allows the features to be robust for different style variations while also being optimal for the registration task. Our feature space adaptation adds no computational cost during inference. We additionally perform unsupervised style transfer of the real X-ray to simulated X-ray image style using Contrastive Unpaired Translation [36]. We apply the style transfer network during inference, thus allowing the registration network to operate on the fixed style already encountered during training. In combination,

our proposed framework achieves a registration accuracy of 1.83 ± 1.16 mm with a high success ratio of 90.1% on real X-ray images showing a 23.9% increase compared to reference annotation-free techniques.

2. Related Work

We focus our related work discussion specific to rigid 2D/3D registration for optimization-based and learning-based 2D/3D registration algorithms. In unsupervised domain adaptation, we broadly discuss the methods applied in medical imaging tasks.

Optimization-Based 2D/3D Registration The problem of 2D/3D registration for interventional image fusion has been extensively researched with comprehensive reviews of the techniques available [27, 29]. Due to the non-convex nature of the 2D/3D registration problem, global optimization [9, 13, 35] is required to reach optimal solution. However, the high computational cost of global optimization-based techniques limits the interventional application. Faster techniques using local optimization-based methods [46, 33, 30, 12, 42] rely on image similarity measures, making it highly dependent on good initialization. Point-to-Plane Correspondence (PPC) constraint was proposed [53, 52, 51] as a more robust alternative for computing the 3D motion from the 2D misalignment visible between the 2D image and the forward projection of the 3D volume. PPC-based techniques significantly improve the registration accuracy and robustness compared to other optimization-based techniques. Extensions of the PPC-based technique proposed for multi-view scenario [40] and hybrid learning-based solutions improve the robustness significantly [39]. Recently, multi-level optimization-based technique [25] was proposed with normalized gradient field as the image similarity metric, showing further improvement in the registration accuracy.

Learning-Based 2D/3D Registration Initially, learning-based techniques were targeted to improve the computational efficiency [32] and robustness [31, 26, 39, 14, 8] of the optimization-based techniques. DL-based techniques significantly improve the robustness to initialization and content mismatch [39, 26, 31]. End-to-end DL-driven registration [20] has shown improved robustness compared to other learning-based methods [39, 19, 41], while also matching the registration accuracy of the optimization-based techniques [51] with significant improvement in computational efficiency. Recently, fully automatic DL-based registration has been proposed [6, 13, 11] that can perform both initialization and registration. A comprehensive review of the learning-based medical image registration [15] and the impact of learning-based 2D/3D registration for interventional applications [47] are available. The advances in DL-based 2D/3D registration techniques have been propelled by using supervised techniques [20, 31, 26]. The variations in imaging protocol, device manufacturer, anatomy, and intervention-specific setting alter the appearance of the acquired images significantly, preventing the adoption of the DL-based 2D/3D registration techniques in interventional scenarios. Attempts have been made to reduce the number of annotated data samples required with paired domain adaptation techniques [58, 59]. Simulated X-ray projections generated from CT volume remove the need for paired annotated data requirement. However, this leads to a domain gap due to the variations between the real and simulated X-ray projection. Unsupervised domain adaptation is required to minimize the drop in performance while not requiring annotated datasets.

Unsupervised Domain Adaptation The domain gap introduced due to the use of simulated data is bridged either by improving the realism of the simulated data [54, 48] or by performing unsupervised domain adaptation to minimize the domain gap [1, 7, 61, 36, 49, 16, 45, 17]. The use of simulated X-ray projections is increasing in training DL-based solutions [11, 3, 57, 55, 60, 43] for various medical imaging applications. DeepDRR [48], aims to bridge the domain gap by rendering realistic simulated X-ray projections which are closer to real X-ray images. Domain Randomization [45] was recently proposed to bridge the domain gap problem in learning-based 2D/3D registration networks [11]. Multiple different styles of simulated X-ray images are used during training, allowing the network to be robust to style variations encountered during inference. However, a patient-specific retraining step is required on top of domain randomization [11]. Unsupervised domain adaptation for CNN-based 6D pose regression of X-ray images was proposed in [60]. However, the evaluation ignores the use of surgical tools and content mismatch which is crucial for the interventional application. Generative Adversarial Network (GAN) [1, 57, 55, 60] and adversarial feature adaptation techniques [22, 5, 60, 50] have shown promising results for bridging the domain gap in medical imaging tasks like segmentation [57], reconstruction [55], pose regression [60], depth estimation [22] and multi-modality learning [50, 5].

3. Methods

Our proposed method is targeted towards rigid 2D/3D registration for interventional image fusion between X-ray (2D) and CT (3D) images as illustrated in Figure 1. The live X-ray image is acquired from the C-arm system during the intervention. The initial overlay depicts the fusion

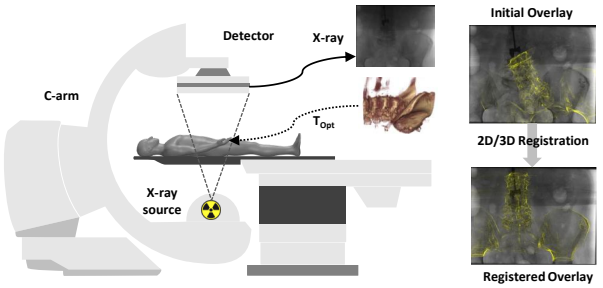

Figure 1: Interventional image fusion for C-arm system, showing the overlay before and after performing 2D/3D registration.

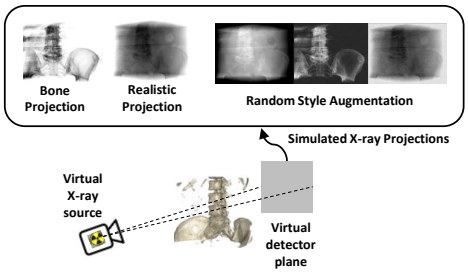

Figure 2: Rendering simulated X-ray projection from 3D CT volume with different styles.

of the 3D volume with the 2D image after performing initialization (either manually or automatically). The registered overlay depicts the overlay produced after performing 2D/3D registration which spatially transforms the preoperative volume with the patient's position and orientation. We give a brief introduction to the 2D/3D registration problem and the registration framework on top of which we build our self-supervised method in Section 3.1. Following, we describe our proposed self-supervised registration technique, with the different components that are used during training and inference in Section 3.2. We finally describe the training and inference procedure used for our registration framework in Section 3.3.

3.1. Background

Problem Formulation In 2D/3D registration, the goal is to find an optimal spatial transformation $\mathbf{T}_{\text{opt}}$ of the volume $\mathbf{V}$ from the observed X-ray projections $\mathbf{I}_f^r$ such that when the images are overlaid, there is minimal misalignment. The problem can be formulated as an optimization problem with an objective function $\mathcal{F}$ that minimizes the misalignment as described in Eq. 1. The X-ray projection $\mathbf{I}_f^r$, the preoperative volume $\mathbf{V}$ and an initial registration estimate $\mathbf{T}_{\text{init}}$ are given as input to the registration algorithm. Our focus is on recovering the registration matrix $\mathbf{T}_{\text{reg}}$ which enables us to find the optimal transformation $\mathbf{T}_{\text{opt}} = \mathbf{T}_{\text{reg}}\mathbf{T}_{\text{init}}$ that aligns the forward projected 3D volume $\mathcal{R}(\mathbf{V}, \mathbf{T})$ with the X-ray projection $\mathbf{I}_f^r$.

$$\underset{\mathbf{T}}{\operatorname{argmin}} \ \mathcal{F}(\mathbf{I}_f^r, \mathcal{R}(\mathbf{V}, \mathbf{T})) \tag{1}$$

Forward Projection A common basis for comparison between the 2D and 3D images is established using the forward projector (rendering) $\mathcal{R}(\mathbf{V}, \mathbf{T})$, which is used to generate simulated X-ray projection $\mathbf{I}_m$ also referred to as Digitally Reconstructed Radiograph (DRR). The forward projection from a CT volume to render a DRR is depicted in Figure 2. The rendering is performed by computing each detector pixel's attenuation response from a virtual X-ray source analytically using ray tracing [48]. The rendering can be done using GPUs allowing for real-time computation [4]. Arbitrary DRRs can be rendered from CT volume for different viewing angles by altering the position and orientation $\mathbf{T}$ of the virtual X-ray source and detector. The style of the rendering can be controlled by selecting the desired materials to be rendered from CT volume either using segmentation or a simple threshold. We show examples of bone projection and realistic projection styles in Figure 2. The bone projection uses thresholding to render only the bones from the CT volume. The realistic projection renders all the materials present in the CT volume. Additionally, random style augmentations (Figure 2) of the projected DRR are obtained by adjusting contrast, brightness, inverting, and adding noise.

PPC-Based Registration Framework Point-to-Plane Correspondence (PPC)-based registration framework [53, 52, 51] constraints the global 3D motion $\mathbf{dv}$ from the visible 2D misalignment using the PPC constraint described in Eq. 2. The framework requires as input, the 3D volume $\mathbf{V}$, X-ray image $\mathbf{I}_f^r$ and initial registration estimate $\mathbf{T}_{init}$. The contour points $\mathbf{w}$ and their gradients $\mathbf{g}$ are computed from $\mathbf{V}$ using a 3D canny edge detector [53]. The 3D motion $\mathbf{dv}$ is estimated by solving the PPC constraint (Eq. 2). The motion estimation $\mathbf{dv}$ is applied iteratively until convergence [53]. During each motion estimation step, the previous registration estimate $\mathbf{T}_{i-1}$ is used for rendering the DRR $\mathbf{I}_m = \mathcal{R}(\mathbf{V}, \mathbf{T}_{i-1})$ with $\mathbf{T}_0 = \mathbf{T}_{init}$. Correspondence is estimated for a set of projected contour points $\mathbf{p}$ between $\mathbf{I}_m$ and $\mathbf{I}_f^r$. The corresponding projected contour point $\mathbf{p}'$ in $\mathbf{I}_f^r$ is used to compute the 2D misalignment $\mathbf{dp} = \mathbf{p}' - \mathbf{p}$. The plane normal $\mathbf{n}$ in Eq. 2 can be computed from $\mathbf{w}$, $\mathbf{g}$ and $\mathbf{dp}$.

$$\mathbf{W} \underbrace{[\mathbf{n} \times \mathbf{w}, -\mathbf{n}]}_{\text{A}} \mathbf{dv} = \operatorname{diag}(\mathbf{W}) \underbrace{\mathbf{n}^T\mathbf{w}}_{\text{b}} \tag{2}$$

The 3D motion $\mathbf{dv}$ is in axis-angle representation and can be directly converted to a 3D transformation matrix $\mathbf{T}_i$

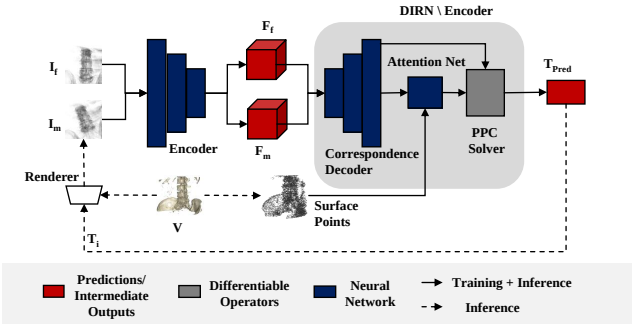

Figure 3: Schematic of DIRN with simulated training using DRRs generated from arbitrary views of the CT volume. The real X-ray image $\mathbf{I}_f^r$ is replaced with bone projection style DRR $\mathbf{I}_f$ for simulated training.

serving as the current registration estimate. To account for noisy correspondence, per correspondence weight (diagonal) matrix $\mathbf{W}$ is used in Eq. 2.

Deep Iterative 2D/3D Registration The closed form solution of Eq. 2 is differentiable [39], allowing the PPC constraint to be embedded as a known operator [28] in learning-based 2D/3D registration methods [39, 20, 41, 19]. We use the recently proposed Deep Iterative 2D/3D Registration Network (DIRN) [20] as the base architecture for our DL-based registration. Figure 3 depicts the schematic of the DIRN-based registration framework. In DIRN, the correspondence search and correspondence weighting of the classical PPC-based registration framework [53] is replaced by learned components. RAFT [44] architecture (encoder and correspondence decoder in Figure 3) is used for estimating the correspondence between the $\mathbf{I}_f$ and $\mathbf{I}_m$ images. The per correspondence weights are learned using a PointNet++ [38] architecture (attention net in Figure 3) which takes $\mathbf{w}, \mathbf{g}, \mathbf{p}', \mathbf{n}$. The predicted correspondences along with the predicted weights are used as inputs to the PPC solver. The 3D motion $\mathbf{dv}$ is computed using the closed form solution of Eq. 2. The network is trained for a single registration update using Eq. 3. During inference, iterative application of the learned network for a fixed number of iteration is used for registration.

$$\mathcal{L}_{dirn} = \mathcal{L}_{\text{reg}} + w_{\text{flow}}\mathcal{L}_{\text{flow}} + w_{\text{m}}\mathcal{L}_m \tag{3}$$

$\mathcal{L}_{\text{reg}} = ||\mathbf{T}(\mathbf{w}) - \hat{\mathbf{T}}(\mathbf{w})||$ is the registration loss with predicted transformation $\mathbf{T}$ computed from the predicted motion $\mathbf{dv}$, ground truth transformation $\hat{\mathbf{T}}$ and 3D contour points $\mathbf{w}$. $\mathcal{L}_{\text{flow}}$ is the flow loss [19, 20] and $\mathcal{L}_m = ||\mathbf{dv}||^2$ is the regularization loss on the predicted 3D motion $\mathbf{dv}$. w_{flow}, w_{m} are the weighting parameters to control the influence of the flow and motion regularization loss respectively, which are set to 0.5 and $1e-3$ respectively [20].

3.2. Self-Supervised Registration Framework

In our self-supervised 2D/3D registration framework depicted in Figure 4, we replace the use of paired training data ($\mathbf{I}_f^r, \mathbf{I}_m = \mathcal{R}(V, \mathbf{T}_{init}), \hat{\mathbf{T}}$) of DIRN framework with simulated training data. The simulated data is obtained using the bone projection style DRR images (Figure 2) $\mathbf{I}_f = \mathcal{R}(V, \mathbf{T}_i)$ (instead of $\mathbf{I}_f^r$), $\mathbf{I}_m = \mathcal{R}(V, \mathbf{T}_j)$ rendered from arbitrary viewing directions $\mathbf{T}_i$ and $\mathbf{T}_j$ respectively. The ground truth registration matrix $\hat{\mathbf{T}}$ is computed from the relative transformation between $\mathbf{T}_i$ and $\mathbf{T}_j$. Additionally, style augmented versions $\mathbf{I}_f^{sa}$, $\mathbf{I}_m^{sa}$ of bone projection style DRR $\mathbf{I}_f$ and $\mathbf{I}_m$ respectively are used for domain randomization. The style augmented version includes realistic projection style DRR and random style augmentations applied to DRR projections during training (Figure 2). We henceforth refer to the network trained with simulated data including domain randomization as simulated DIRN. The inference still needs to be performed on the real X-ray images $\mathbf{I}_f^r$. The simplifying assumptions used to render DRRs compared to real X-ray images [48], lead to variations in image properties between the DRRs and the real X-ray images. To ensure that the simulated DIRN can perform with minimal performance gap on $\mathbf{I}_f^r$, we perform unsupervised feature and pixel space domain adaptation. In feature space adaptation (Section 3.2.1), our goal is to minimize the distribution shift between the encoded feature maps of $\mathbf{I}_f^r$ and $\mathbf{I}_f$, while the features are optimal for DIRN. We perform the pixel space adaptation (Section 3.2.2) using unsupervised image-to-image style transfer network, where we transfer the input X-ray image $\mathbf{I}_f^r$ to the fixed bone projection style DRR $\mathbf{I}_f$. The training of the style transfer network is performed separately and coupled with feature adapted network during inference.

3.2.1 Feature Space Domain Adaptation

In feature space domain adaptation, the goal is to ensure the encoded feature map is well adapted for different style variations to the input image that can be encountered during the inference. Our feature space domain adaptation consists of Adversarial Feature Encoder (AFE) and Barlow Twins (BT) [56] module trained together with DIRN in an end-to-end manner. Figure 4 depicts both the feature space domain adaptation modules and how it is trained together with DIRN. In the following, we introduce the modules separately and describe how the end-to-end training is performed together with the simulated DIRN. We use the image encoder from the original RAFT architecture [44] as proposed in DIRN [20].

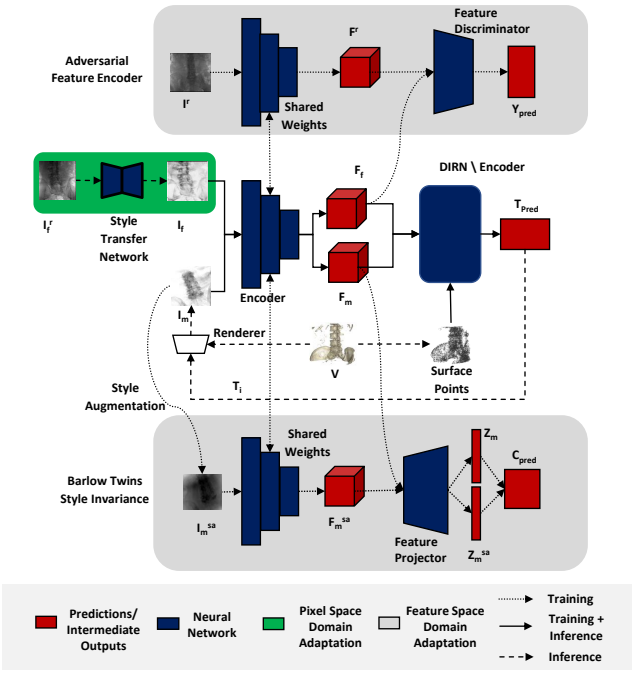

Figure 4: Our proposed registration framework with unsupervised pixel and feature space domain adaptation. We depict the encoder separately from the other DIRN modules for ease of visualization.

Adversarial Feature Encoder Adversarial feature adaptation as a standalone module has been previously proposed [7, 16] with application in unsupervised domain adaptation for medical images [22]. We use the adversarial feature loss (Eq. 4) computed using the encoded feature maps $\mathbf{F}_f$ and $\mathbf{F}^r$ from the bone style DRR $\mathbf{I}_f$ and unpaired real X-ray image $\mathbf{I}^r$. The feature discriminator D_f (based on patch GAN discriminator [18]) is used to distinguish between the feature representations of the real and simulated images. The encoder can be trained with adversarial training [10] which ensures that the encoded feature distribution produced from both the X-ray and DRR images matches closely for similar structural content in both image.

$$\begin{aligned}\mathcal{L}_{\text{afe}} = \mathbb{E}_{\mathbf{x}\sim\mathbf{I}^s}[logD_f(E(\mathbf{x}))] \\ + \mathbb{E}_{\mathbf{x}\sim\mathbf{I}^r}[logD_f(1-E(\mathbf{x}))]\end{aligned} \tag{4}$$

Barlow Twins for Style Invariance We use the Barlow Twins(BT) [56] module which was originally proposed for representation learning with recent advancements showing improvements in multi-modal representation learning [2]. We propose to use the original BT [56] for our feature adaptation imparting style invariance. The BT loss (Eq. 5) is computed using bone projection style DRR $\mathbf{I}_m$ and style augmented version $\mathbf{I}_m^{sa}$ of it. We compute the encoded feature maps $\mathbf{F}_m$, $\mathbf{F}_m^{sa}$ for both the images $\mathbf{I}_m$ and $\mathbf{I}_m^{sa}$ respectively. The encoded feature map is projected to a z-dim embedding vector $\mathbf{Z}_m$ and $\mathbf{Z}_m^{sa}$ using the feature projector as proposed in [56]. Cross-correlation $\mathbf{C}_{pred}$ is computed between the embedded feature vectors $\mathbf{Z}_m$ and $\mathbf{Z}_m^{sa}$. Barlow twins loss [56] (Eq. 5) which consists of invariance term and redundancy reduction term is used for training the encoder making it learn a feature representation that is invariant to the input style.

$$\mathcal{L}_{\text{bt}} = \sum_i (1-\mathbf{C}_{pred}^{ii})^2 + w_{\text{red}} \sum_i \sum_{i\neq j} (\mathbf{C}_{pred}^{ij})^2, \tag{5}$$

where w_{red} is the weighting factor for the redundancy loss term [56] which is set to 0.005.

End-to-end training with simulated DIRN Unsupervised feature adaptation for real X-ray images and different input styles can be achieved using the AFE and BT modules respectively. However, the features are not constrained to be optimal for the registration task (simulated DIRN). We propose to constrain the solution space for feature adaptation by training all the three modules (simulated DIRN, AFE, and BT) together in an end-to-end manner using shared weights for the encoder between the modules. Since AFE and BT does not require any paired registration data, we can perform task-specific unsupervised feature space domain adaption. We describe the training strategy used to train our network with Eq. 6 in Section 3.3.

$$\mathcal{L} = \mathcal{L}_{\text{dirn}} + w_{\text{afe}}\mathcal{L}_{\text{afe}} + w_{\text{dirn}}\mathcal{L}_{\text{bt}}, \tag{6}$$

where w_{afe} and w_{bt} are the weighting factors for the AFE and BT modules which are set to 0.2 and 0.05 respectively.

3.2.2 Pixel Space Domain Adaptation

In pixel space adaptation, our goal is to make the pixel distribution of real X-ray images match to the simulated images encountered during training. We train our pixel space domain adaptation component separately, where we use an unsupervised image-to-image style transfer network. A common method to perform such style transfer is to use CycleGAN [61]. However, CycleGANs need to learn both forward and inverse mapping without any specific structural loss to preserve the structural content of the style transferred images. We instead use the recently proposed Contrastive Unpaired Translation (CUT) [36] avoiding the need for learning forward and inverse mapping while also ensuring that the structural content is preserved with the patch NCE (PNCE) loss. We depict our X-ray to DRR style transfer using the CUT network in Figure 5. The PNCE loss is computed using the input X-ray image and generated DRR

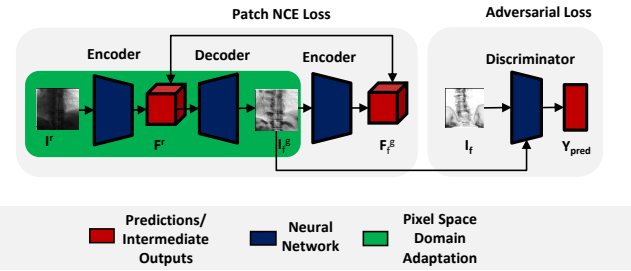

Figure 5: Unsupervised X-ray to DRR style transfer using Contrastive Unpaired Translation [36]

on multiple feature levels of the encoder [36]. A discriminator is used to distinguish between the real DRR images and the fake DRR images generated from input X-ray. The network is trained using Eq. 7 with adversarial learning, where the goal of the generator is to produce DRR-styled images conditioned on the input X-ray images.

$$\mathcal{L}_{cut} = \mathcal{L}_{gan} + \mathcal{L}_{pnce} + \mathcal{L}^{id}_{pnce} \tag{7}$$

The $\mathcal{L}_{gan}$ is the Generative Adversarial Network (GAN) loss [10] for matching the distribution for the generated DRR with real DRR images. The $\mathcal{L}_{pnce}$ loss is the contrastive loss computed between the X-ray and generated fake DRR image patches ensuring the content of the style transferred image is preserved [36]. The identity $\mathcal{L}_{pnce}$ is a learnable domain specific identity loss computed using the real DRR images [36]. The CUT network is trained separately for X-ray to DRR transfer and is applied as a style transfer module during the inference to transfer the real X-ray images to DRR-styled images for our self-supervised 2D/3D registration network.

3.3. Training and Inference

We pre-train simulated DIRN using the registration loss (Eq. 3) similar to [20]. We use simulated training with a fixed bone projection style DRR for 50 epochs and combine it with style augmented data for domain randomization for 50 epochs. We then fine-tune for feature space domain adaptation 3.2 using the combined loss (Eq. 6) in an end-to-end manner for 20 epochs. A cyclical learning rate between 1e-4 to 1e-6 is used during the simulated DIRN training phase. During fine-tuning for feature space adaptation a lower learning rate of 5e-6 is used for all DIRN modules except the image encoder. This is to make sure that our task performance is retained, while our encoder adapts for the different feature adaptations we perform using BT and AFE loss. The learning rate of other components (including the encoder) is set at 1e-4. We use a batch size of 16 for both pre-training and fine-tuning. We use Adam optimizer [23] to optimize all the network parameters. We update the feature discriminator of the AFE after each DIRN update. During inference except for the learned encoder weights, other feature adaptation modules are not used as depicted in Figure 4.

We train our pixel space domain adaptation network separately on unpaired X-ray and DRR image datasets. We train the CUT network for 50 epochs using a learning rate at 1e-4 and batch size of 2 with Adam optimizer [23]. The trained generator network from CUT is then used during the inference as depicted in Figure 4. We train our networks using PyTorch [37]. Our models are trained using a single NVIDIA V100 Tesla GPU with 16GB memory. The inference is run on NVIDIA Titan X GPU with 12 GB memory with 10 iterations of DIRN for a registration sample. Further implementation details are provided in the supplementary material.

4. Experiments and Results

4.1. Experimental Setup

We describe the dataset used for training and evaluation in Section 4.1.1 followed by the evaluation measures in Section 4.1.2 and the baseline methods against which we benchmark in Section 4.1.3.

4.1.1 Dataset

We use clinical Cone Beam CT (CBCT) reconstruction dataset consisting of the X-ray images with ground truth registration used for reconstruction along with the reconstructed 3D CBCT volumes. Our dataset is from the vertebra region, consisting of both thoracic and lumbar regions with 55 CBCT volumes acquired from 55 patients. The voxel spacing varies between 0.49 mm to 0.99 mm in all three dimensions. Due to the variations in the slice thickness, the number of X-ray images varies between 190 to 390 images per volume. The X-ray images have a resolution of 616×480 (width $\times$ height) with a pixel spacing of 0.616 mm. We split our dataset into 43 patients for training, 6 patients for validation, and 6 patients for testing. We report all the results on the held-out test data set. A visualization of samples from our dataset is provided in the supplementary material.

In the case of supervised scenario (used for training supervised DIRN [20] for comparison), we create random initial transformations from the ground truth registration of the X-ray images with the initial registration error in the range of $[0, 30]$ mm. A training sample consists of $\mathbf{I}^r_f$, $\mathbf{T}_{init}$, $\hat{\mathbf{T}}$, $\mathbf{I}_m = \mathcal{R}(\mathbf{V}, \mathbf{T}_{init})$ and $\mathbf{w}$. In the self-supervised scenario, we retain the same random start position from $\hat{\mathbf{T}}$ of $\mathbf{I}^r_f$ and just replace it with $\mathbf{I}_f$, which can be rendered using $\hat{\mathbf{T}}$. We additionally have style augmented version of DRR images (Figure 2) for domain randomization. In to-

tal, we have 80,000 samples for training and validation for both supervised and self-supervised scenarios. To train the pixel space domain adaptation network and adversarial feature encoder we use a dataset of 16000 unpaired real X-ray and bone projection style DRR images from the vertebra region for the same set of patients. Our test data set consists of 3600 samples from Anterior Posterior (AP) and lateral (LAT) views (as its the common views encountered during interventions [26, 31]) for the 6 patients with initial registration error in the range of $[0, 60]$ mm similar to [20]. We use real X-ray images for all evaluations.

4.1.2 Evaluation Measures

We use the standardized evaluation measure [24] of mean Re-Projection Distance (mRPD) and the success ratio (SR) of mRPD ≤ 5.0 mm [34]. The mRPD indicates the overlay misalignment error and SR indicates the ratio of the number of samples to the total samples for which the mRPD $\leq$ 5.0 mm. The initial registration error is measured using the mean Target Registration Error (mTRE) [39, 41] indicates the 3D euclidean distance between the start position and the ground truth registration varies between $[0, 60]$ mm.

4.1.3 Baseline methods

We compare our proposed technique with both optimization-based and learning-based 2D/3D registration. For the optimization-based technique, we use DPPC [51] as it showed state-of-the-art results compared to other optimization-based techniques [53, 51]. For learning-based 2D/3D registration, we directly compare with DIRN [20], as we build our self-supervised registration framework on top of it. As an annotation-free baseline, we use the simulated DIRN with domain randomization [11] which is commonly used in state-of-the-art registration networks that are trained only with simulated images [11]. We use the same training, validation, and test data for all the methods and the best hyper-parameter settings proposed in the respective original works.

4.2. Results

We validate our unsupervised pixel space domain adaptation performance against standard supervised style transfer using pix2pix [18] in Section 4.2.1. Following, we present the ablation of the pixel and feature space domain adaptation components of our proposed network in Section 4.2.2. We then compare our method with other state-of-the-art techniques in Section 4.2.3

4.2.1 Pixel space domain adaptation

We compare our pixel space adaptation using CUT network [36] with supervised pix2pix network [18] for registration performance (Table 1) and X-ray to DRR style transfer (Figure 6). Our proposed unsupervised pixel space domain adaptation matches closely to the supervised style transfer network both in registration performance (Table 1) and the image appearance (Figure 6) indicating we are close to the maximum performance achievable for domain adaptation using style transfer.

	mRPD [mm] ↓	SR [%] ↑
Simulated	2.97 ± 0.99	66.2
+ Unsupervised	2.35 ± 1.1	85.8
+ Supervised	**2.25 ± 1.0**	**86.8**

Table 1: Comparison of our unsupervised pixel space domain adaptation using CUT [36] with supervised style transfer using pix2pix [18].

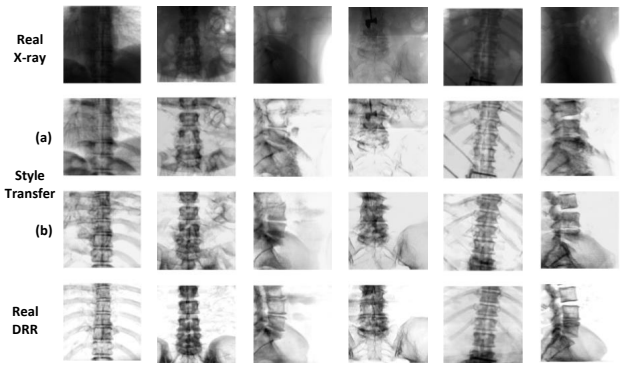

Figure 6: X-ray to DRR style transfer using (a) unsupervised CUT network (b) supervised pix2pix network along with real DRR. Each column indicates a test data sample.

	mRPD [mm] ↓	SR [%] ↑
Simulated	2.97 ± 0.99	66.2
+ Feature	2.30 ± 1.26	72.2
+ Pixel	2.35 ± 1.1	85.8
+ Feature + Pixel	**1.83 ± 1.16**	**90.1**

Table 2: Ablation of the different components of our self-supervised framework.

4.2.2 Ablation of domain adaptation components

We compare the different components of our self-supervised 2D/3D registration framework and how each component drives the performance. In Table 2, we show the quantitative results comparing the different components of our registration framework. We start with our baseline scenario, where we use simulated training of

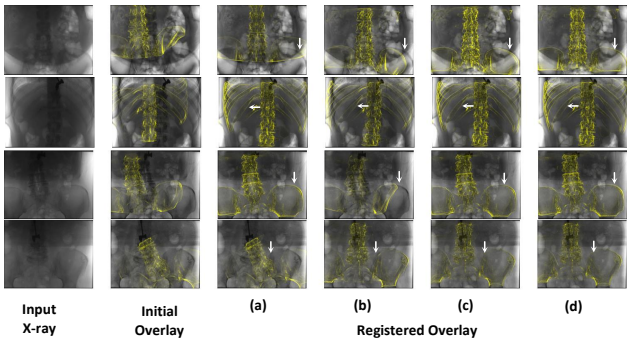

Figure 7: Visual comparison of overlays produced using (a) optimization-based [53, 51] (b) simulated DIRN (includes domain randomization [11]) (c) ours (d) supervised (DIRN) [20]. Each row indicates a test data sample.

	mRPD [mm] ↓	SR[%] ↑
Optimization-based [53]	**0.59 ± 0.26**	62.2
Simulated	2.97 ± 0.99	66.2
Ours	1.83 ± 1.16	90.1
Supervised [20]	0.65 ± 0.50	**99.4**

Table 3: Comparison of our proposed method with other state-of-the-art techniques.

DRR with domain randomization as proposed in [11]. We add the feature space domain adaptation technique to train our network and this already shows a 6% increase in SR compared to the simulated baseline. Adding the pixel space domain adaptation technique with the simulated baseline increases the SR by 19.6%. Our proposed self-supervised registration which consists of both pixel and feature space domain adaptation improves the SR by 23.9% from the simulated baseline. The registration error is also reduced from 2.97 ± 0.99 mm to 1.83 ± 1.16 mm. Further analysis of simulated baseline and the visualization of registration error distribution is provided in the supplementary material.

4.2.3 Comparison with state-of-the-art registration methods

We now compare with the different baseline methods which have shown state-of-the-art performance for different registration scenarios. Figure 7 shows the qualitative comparison of the overlays produced before and after registration. An arrow is marked on overlays produced using different methods to better illustrate the differences. The quantitative evaluation from Table 3 shows our proposed method achieves a SR of 90.1% and has minimal performance drop compared to supervised method. The optimization-based technique lacks robustness to large initial misalignment resulting in lower SR compared to our proposed method. Figure 7 shows that overlays produced with our method is more accurate than simulated baseline and matches closely to the supervised DIRN [20].

5. Discussion

Our novel unsupervised feature space domain adaptation couples Barlow Twins and Adversarial Feature Encoder, allowing us to increase the SR by 6% compared to the simulated baseline without any additional parameters during inference (Table 2). Our unsupervised pixel space domain adaptation using CUT network [36] closely matches the performance of the supervised pix2pix-based style transfer network. The additional overhead is minimal for the style transfer network as it requires only a single forward pass of the generator during application. Also, unlike the CycleGAN [61], we can directly learn the forward mapping between X-ray to DRR with structure consistency enforced by PatchNCE loss. Due to the strict design requirement of having structural consistency for style transfer, we skip the comparison with the CycleGAN as they do not satisfy it. CUT [36] also outperforms it significantly for unsupervised image-to-image translation in [36]. We compare against the state-of-the-art 2D/3D registration techniques and clearly show that our proposed method achieves significant improvements to the SR compared to other annotation-free methods [51] and closely matches the supervised learning-based technique [20]. Our proposed method also shows significant improvement in mRPD compared to the simulated baseline. However, one limitation of our proposed technique is the higher registration error compared to optimization-based and supervised registration methods. Optimization-based refinement step [51, 29, 31, 26] can be applied if lower registration error is necessary for the application.

6. Conclusion

Our self-supervised framework enables training learning-based 2D/3D registration without the need for annotated paired datasets. We are one of the first to propose self-supervised 2D/3D registration targeted for dense correspondence-based 2D/3D registration networks [20, 19, 41], which are highly sensitive to the small changes in the image content. We combine the novel techniques from domain adaptation, self-supervised representation learning, and image-to-image translation to build a complete framework for self-supervised learning of dense correspondence-based 2D/3D registration networks. We achieve a high SR of 90.1% on real X-ray images with a 23.9% increase in SR compared to annotation-free alternatives. The mRPD is also reduced from 2.97 ± 0.99 mm for the simulated baseline to 1.83 ± 1.16 mm for our proposed method.

References

[1] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684, 2020.

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[3] Bastian Bier, Mathias Unberath, Jan-Nico Zaech, Javad Fotouhi, Mehran Armand, Greg Osgood, Nassir Navab, and Andreas Maier. X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 55–63. Springer, 2018.

[4] Wolfgang Birkfellner, Rudolf Seemann, Michael Figl, Johann Hummel, Christopher Ede, Peter Homolka, Xinhui Yang, Peter Niederer, and Helmar Bergmann. Fast drr generation for 2d/3d registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 960–967. Springer, 2005.

[5] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, Ben Glocker, Xiahai Zhuang, and Pheng-Ann Heng. Pnp-adanet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation. *arXiv preprint arXiv:1812.07907*, 2018.

[6] Javier Esteban, Matthias Grimm, Mathias Unberath, Guillaume Zahnd, and Nassir Navab. Towards fully automatic x-ray to ct registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 631–639. Springer, 2019.

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[8] Cong Gao, Xingtong Liu, Wenhao Gu, Benjamin Killeen, Mehran Armand, Russell Taylor, and Mathias Unberath. Generalizing spatial transformers to projective geometry with applications to 2d/3d registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 329–339. Springer, 2020.

[9] Ren Hui Gong and Purang Abolmaesumi. 2d/3d registration with the cma-es method. In *Medical Imaging 2008: Visualization, Image-Guided Procedures, and Modeling*, volume 6918, pages 556–564. SPIE, 2008.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[11] 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.

[12] Martin Groher. *2D-3D registration of vascular images*. PhD thesis, Technische Universität München, 2008.

[13] 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.

[14] Wenhao Gu, Cong Gao, Robert Grupp, Javad Fotouhi, and Mathias Unberath. Extended capture range of rigid 2d/3d registration by estimating riemannian pose gradients. In *International Workshop on Machine Learning in Medical Imaging*, pages 281–291. Springer, 2020.

[15] Grant Haskins, Uwe Kruger, and Pingkun Yan. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31(1):1–30, 2020.

[16] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[17] Malte Hoffmann, Benjamin Billot, Douglas N Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. Synthmorph: learning contrast-invariant registration without acquired images. *IEEE transactions on medical imaging*, 41(3):543–558, 2021.

[18] 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.

[19] Srikrishna Jaganathan, Jian Wang, Anja Borsdorf, and Andreas Maier. Learning the update operator for 2d/3d image registration. In *Bildverarbeitung für die Medizin 2021*, pages 117–122. Springer, 2021.

[20] 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.

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[22] Mert Asim Karaoglu, Nikolas Brasch, Marijn Stollenga, Wolfgang Wein, Nassir Navab, Federico Tombari, and Alexander Ladikos. Adversarial domain feature adaptation for bronchoscopic depth estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 300–310. Springer, 2021.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] Everine B Kraats, Graeme P Penney, Dejan Tomaževič, Theo van Walsum, and Wiro J Niessen. Standardized evaluation of 2d-3d registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 574–581. Springer, 2004.

[25] Annkristin Lange and Stefan Heldmann. Multilevel 2d-3d intensity-based image registration. In *International Workshop on Biomedical Image Registration*, pages 57–66. Springer, 2020.

[26] Haofu Liao, Wei-An Lin, Jiarui Zhang, Jingdan Zhang, Jiebo Luo, and S Kevin Zhou. Multiview 2d/3d rigid registration via a point-of-interest network for tracking and triangulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12638–12647, 2019.

[27] Rui Liao, Li Zhang, Ying Sun, Shun Miao, and Christophe Chefd'Hotel. A review of recent advances in registration techniques applied to minimally invasive therapy. *IEEE transactions on multimedia*, 15(5):983–1000, 2013.

[28] Andreas K Maier, Christopher Syben, Bernhard Stimpel, Tobias Würfl, Mathis Hoffmann, Frank Schebesch, Weilin Fu, Leonid Mill, Lasse Kling, and Silke Christiansen. Learning with known operators reduces maximum error bounds. *Nature machine intelligence*, 1(8):373–380, 2019.

[29] 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.

[30] Stefan Matl, Richard Brosig, Maximilian Baust, Nassir Navab, and Stefanie Demirci. Vascular image registration techniques: A living review. *Medical Image Analysis*, 35:1–17, 2017.

[31] Shun Miao, Sebastien Piat, Peter Fischer, Ahmet Tuysuzoglu, Philip Mewes, Tommaso Mansi, and Rui Liao. Dilated FCN for multi-agent 2D/3D medical image registration. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 4694–4701, 2018.

[32] Shun Miao, Z. Jane Wang, and Rui Liao. A CNN Regression Approach for Real-Time 2D/3D Registration. *IEEE Transactions on Medical Imaging*, 35(5):1352–1363, 2016.

[33] Uroš Mitrović, Primož Markelj, Boštjan Likar, Zoran Miloševič, and Franjo Pernuš. Gradient-based 3d-2d registration of cerebral angiograms. In *Medical Imaging 2011: Image Processing*, volume 7962, pages 533–540. SPIE, 2011.

[34] Y Otake, S Schafer, JW Stayman, W Zbijewski, G Kleinszig, R Graumann, AJ Khanna, and JH Siewerdsen. Automatic localization of vertebral levels in x-ray fluoroscopy using 3d-2d registration: a tool to reduce wrong-site surgery. *Physics in Medicine & Biology*, 57(17):5485, 2012.

[35] Yoshito Otake, Adam S Wang, J Webster Stayman, Ali Uneri, Gerhard Kleinszig, Sebastian Vogt, A Jay Khanna, Ziya L Gokaslan, and Jeffrey H Siewerdsen. Robust 3d–2d image registration: application to spine interventions and vertebral labeling in the presence of anatomical deformation. *Physics in Medicine & Biology*, 58(23):8535, 2013.

[36] 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.

[37] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[38] 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.

[39] Roman Schaffert, Jian Wang, Peter Fischer, Anja Borsdorf, and Andreas Maier. Learning an Attention Model for Robust 2-D/3-D Registration Using Point-To-Plane Correspondences. *IEEE transactions on medical imaging*, 39(10):3159–3174, 2020.

[40] Roman Schaffert, Jian Wang, Peter Fischer, Andreas Maier, and Anja Borsdorf. Robust multi-view 2-d/3-d registration using point-to-plane correspondence model. *IEEE transactions on medical imaging*, 39(1):161–174, 2019.

[41] Roman Schaffert, Markus Weiß, Jian Wang, Anja Borsdorf, and Andreas Maier. Learning-based correspondence estimation for 2-d/3-d registration. In *Bildverarbeitung für die Medizin 2020*, pages 222–228. Springer, 2020.

[42] Žiga Špiclin, Boštjan Likar, and Franjo Pernuš. Fast and robust 3d to 2d image registration by backprojection of gradient covariances. In *International Workshop on Biomedical Image Registration*, pages 124–133. Springer, 2014.

[43] Richin Sukesh, Andreas Fieselmann, Srikrishna Jaganathan, Karthik Shetty, Rainer Kärgel, Florian Kordon, Steffen Kappler, and Andreas Maier. Training deep learning models for 2d spine x-rays using synthetic images and annotations created from 3d ct volumes. In *Bildverarbeitung für die Medizin 2022*, pages 63–68. Springer, 2022.

[44] 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.

[45] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[46] Dejan Tomazevic, Bostjan Likar, Tomaz Slivnik, and Franjo Pernus. 3-d/2-d registration of ct and mr to x-ray images. *IEEE transactions on medical imaging*, 22(11):1407–1416, 2003.

[47] Mathias Unberath, Cong Gao, Yicheng Hu, Max Judish, Russell H Taylor, Mehran Armand, and Robert Grupp. The impact of machine learning on 2d/3d registration for image-guided interventions: A systematic review and perspective. *Frontiers in Robotics and AI*, 8:716007, 2021.

[48] Mathias Unberath, Jan-Nico Zaech, Sing Chun Lee, Bastian Bier, Javad Fotouhi, Mehran Armand, and Nassir Navab. Deepdrr–a catalyst for machine learning in fluoroscopy-guided procedures. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 98–106. Springer, 2018.

[49] Moritz Venator, Fengyi Shen, Selcuk Aklanoglu, Erich Bruns, Klaus Diepold, and Andreas Maier. Dual-mode training with style control and quality enhancement for road image domain adaptation. *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, pages 1746–1755, 2020.

[50] Sulaiman Vesal, Mingxuan Gu, Ronak Kosti, Andreas Maier, and Nishant Ravikumar. Adapt everywhere: unsupervised adaptation of point-clouds and entropy minimization for multi-modal cardiac image segmentation. *IEEE Transactions on Medical Imaging*, 40(7):1838–1851, 2021.

[51] 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.

[52] Jian Wang, Anja Borsdorf, Benno Heigl, Thomas Köhler, and Joachim Hornegger. Gradient-based differential approach for 3-d motion compensation in interventional 2-d/3-d image fusion. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 293–300. IEEE, 2014.

[53] Jian Wang, Roman Schaffert, Anja Borsdorf, Benno Heigl, Xiaolin Huang, Joachim Hornegger, and Andreas Maier. Dynamic 2-D/3-D rigid registration framework using point-to-plane correspondence model. *IEEE Transactions on Medical Imaging*, 36(9):1939–1954, 9 2017.

[54] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021.

[55] Xingde Ying, Heng Guo, Kai Ma, Jian Wu, Zhengxin Weng, and Yefeng Zheng. X2ct-gan: reconstructing ct from biplanar x-rays with generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10628, 2019.

[56] 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.

[57] Yue Zhang, Shun Miao, Tommaso Mansi, and Rui Liao. Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 599–607. Springer, 2018.

[58] Jiannan Zheng, Shun Miao, and Rui Liao. Learning cnns with pairwise domain adaption for real-time 6dof ultrasound transducer detection and tracking from x-ray images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 646–654. Springer, 2017.

[59] Jiannan Zheng, Shun Miao, Z Jane Wang, and Rui Liao. Pairwise domain adaptation module for cnn-based 2-d/3-d registration. *Journal of Medical Imaging*, 5(2):021204, 2018.

[60] Shiqiang Zheng, Xin Yang, Yifan Wang, Mingyue Ding, and Wenguang Hou. Unsupervised cross-modality domain adaptation network for cnn-based x-ray to ct. *IEEE Journal of Biomedical and Health Informatics*, 2021.

[61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

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 11) and data samples from our clinical CBCT reconstruction dataset (Figure 12). 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 4, 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 4: 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

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

Table 5: 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.

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 5, we achieve similar results to DIRN [20] 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 8, 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.

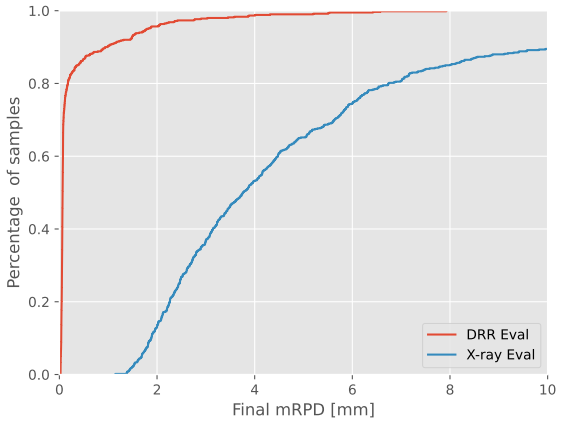

Figure 8: Comparison of final registration error using empirical cumulative distribution for DRR Eval and X-ray Eval of our simulated DIRN (includes DR), indicating the large domain gap that exists even after the application of domain randomization.

A.3. Ablation of Domain Adaptation Components

We visualize the cumulative distribution and kernel density distribution of the final registration error in Figure 9 and Figure 10 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.

B. Visualization

B.1. Comparison of Registered Overlays

Figure 11 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 12 shows example images from our clinical CBCT reconstruction dataset which includes the CBCT reconstructed volume along with the paired X-ray images.

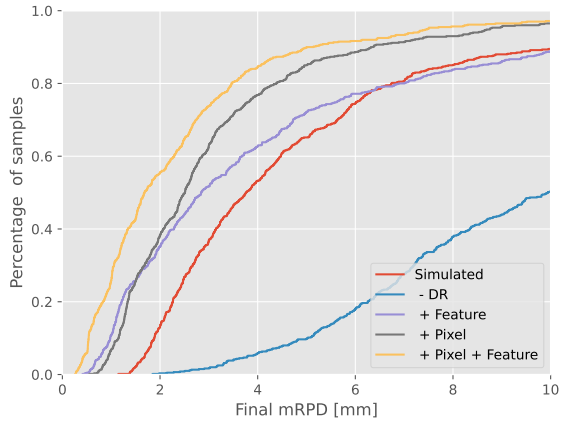

Figure 9: Comparison of final registration error using empirical cumulative distribution for different variations of our proposed framework.

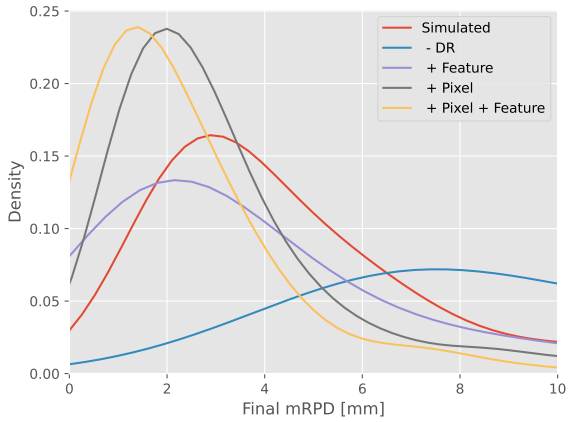

Figure 10: Comparison of final registration error using kernel density estimation for different variations of our proposed framework.

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 [20], feature adaptation components (Adversarial Feature Encoders and Barlow Twins [56]), and the unsupervised style transfer network [36]. We use the architecture proposed in the respective original works, with the specific configuration used for our framework described below. The registration network DIRN [20] consists of RAFT [44] 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 [44]. 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 [38] for correspondence weighting as proposed in DIRN [20]. 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 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 [18] 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 [36] uses a ResNet based generator consisting of 9 residual blocks [21] and patch GAN discriminator [18], with a patch size of 16.

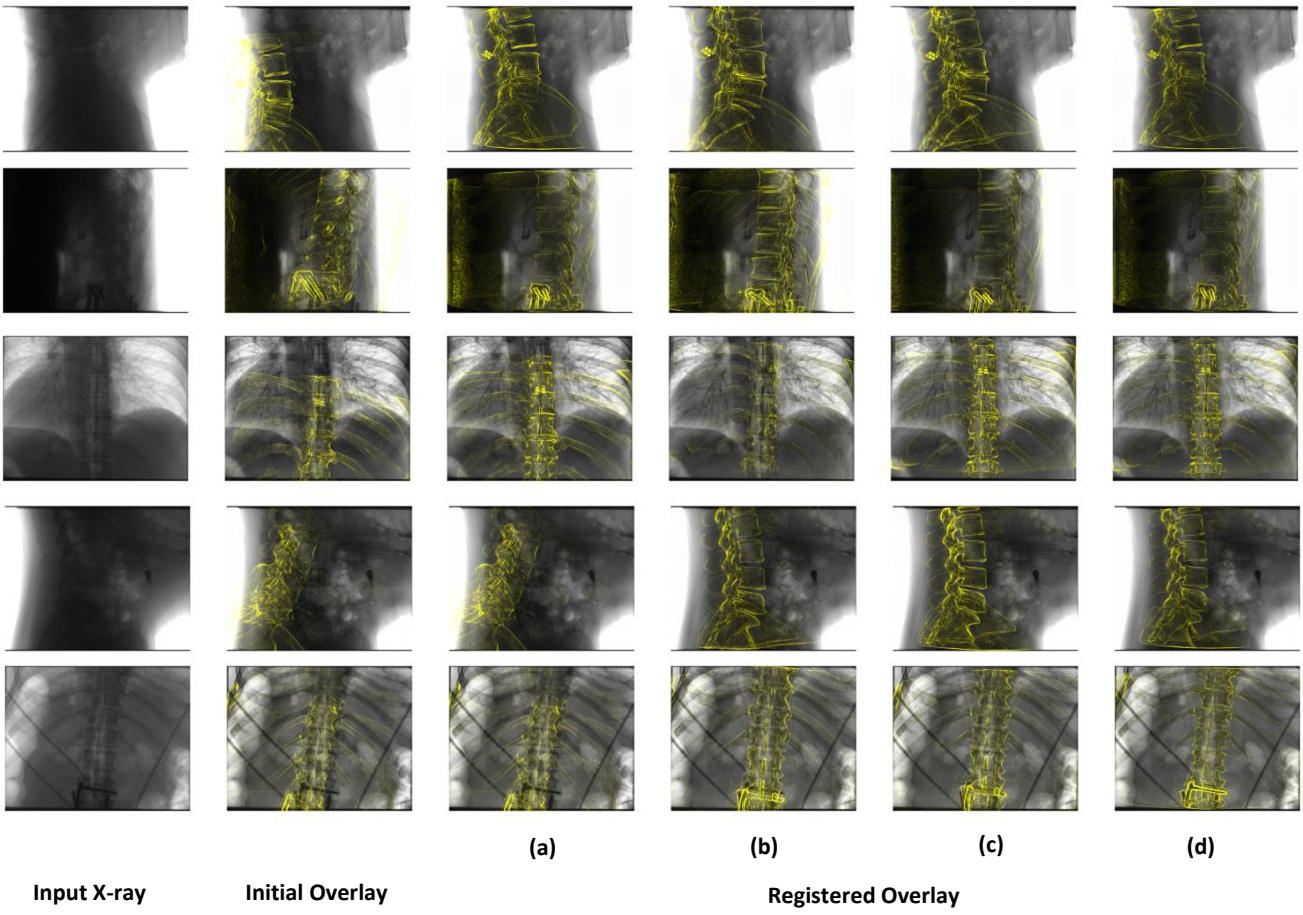

Figure 11: Additional samples from test dataset with comparison of overlays produced using (a) Optimization-based technique [51], (b) Simulated (with domain randomization [11]), (c) our proposed method, and (d) supervised [20]. Each row represents a data sample from the test dataset.

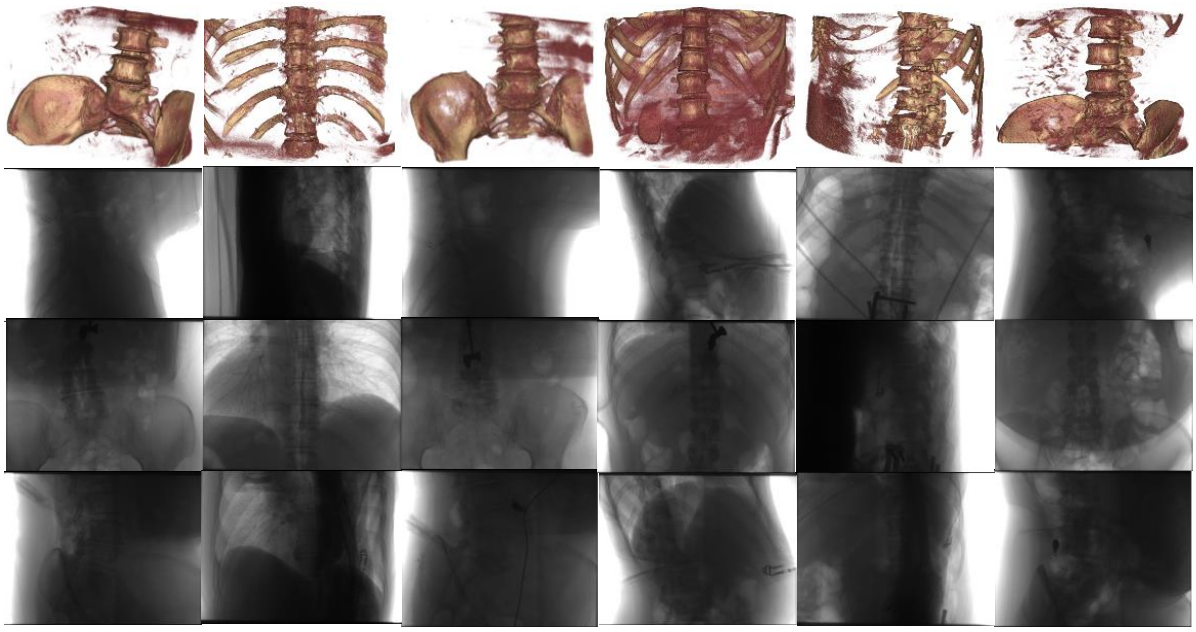

Figure 12: Exemplar data samples from our clinical CBCT dataset. The reconstructed volume is thresholded to better visualize the bone contours. Each column represents a reconstructed CBCT volume along with paired set of X-ray images used in reconstructing the CBCT volume.