A Patient-Specific Self-supervised Model for Automatic X-Ray/CT Registration

Baochang Zhang[1,2(✉)], Shahrooz Faghihroohi[1], Mohammad Farid Azampour[1], Shuting Liu[1], Reza Ghotbi[3], Heribert Schunkert[2,4], and Nassir Navab[1]

[1] Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany
baochang.zhang@tum.de
[2] German Heart Center Munich, Munich, Germany
[3] HELIOS Hospital west of Munich, Munich, Germany
[4] German Centre for Cardiovascular Research, Munich Heart Alliance, Munich, Germany

Abstract. The accurate estimation of X-ray source pose in relation to pre-operative images is crucial for minimally invasive procedures. However, existing deep learning-based automatic registration methods often have one or some limitations, including heavy reliance on subsequent conventional refinement steps, requiring manual annotation for training, or ignoring the patient's anatomical specificity. To address these limitations, we propose a patient-specific and self-supervised end-to-end framework. Our approach utilizes patient's preoperative CT to generate simulated X-rays that include patient-specific information. We propose a self-supervised regression neural network trained on the simulated patient-specific X-rays to predict six degrees of freedom pose of the X-ray source. In our proposed network, regularized autoencoder and multi-head self-attention mechanism are employed to encourage the model to automatically capture patient-specific salient information that supports accurate pose estimation, and Incremental Learning strategy is adopted for network training to avoid over-fitting and promote network performance. Meanwhile, an novel refinement model is proposed, which provides a way to obtain gradients with respect to the pose parameters to further refine the pose predicted by the regression network. Our method achieves a mean projection distance of 3.01 mm with a success rate of 100% on simulated X-rays, and a mean projection distance of 1.55 mm on X-rays. The code is available at github.com/BaochangZhang/PSSS_registration.

Keywords: X-ray/CT Registration · Self-supervised Learning · Patient-Specific Model

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_49.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 515–524, 2023.
https://doi.org/10.1007/978-3-031-43996-4_49

1 Introduction

Augmentation of intra-operative X-ray images using the pre-operative data (e.g., treatment plan) has the potential to reduce procedure time and improve patient outcomes in minimally invasive procedures. However, surgeons must rely on their clinical knowledge to perform a mental mapping between pre- and intra-operative information, since the pre- and intra-operative images are based on different coordinate systems. To utilize pre-operative data efficiently, accurate pose estimation of the X-ray source relative to pre-operative images (or called registration between pre- and intra-operative data) is necessary and beneficial to relieve surgeon's mental load and improve patient outcomes.

Although 2D/3D registration methods for medical images have been widely researched and systematically reviewed [10,17], developing an automatic end-to-end registration method remains an open issue. For conventional intensity-based registration, it is formulated as an iterative optimization problem based on similarity measures. A novel similarity measure called weighted local mutual information is proposed to perform solid vascular 2D-3D registration [11] but has a limited capture range, becoming inefficient and prone to local minima if initial registration error is large. A good approach that directly predicts the spatial mapping relationship between simulated X-rays and real X-rays using a neural network is put forward [12] but requires an initialization. Some supervised learning tasks, such as anatomical landmark detection [1,3], are defined to develop a robust initialization scheme. When used to initialize an optimizer [14] for refinement, they can lead to a fully automatic 2D/3D registration solution [3]. But extensive manual annotation [1,3] or pairwise clinical data [12] is needed for training, which is time- and labor-consuming when expanding to new anatomies, and the robustness of these methods might be challenged due to the neglect of patient's anatomical specificity [1]. A patient-specific landmark refinement scheme is then proposed, which contributes to model's robustness when applied intraoperatively [2]. Nevertheless, the final performance of this automatic registration method still relies on conventional refinement step based on derivative-free optimizer (i.e., BOBYQA), which limits the computational efficiency of deep learning-based registration. Meanwhile, some studies employ regression neural networks to directly predict slice's pose relative to pre-operative 3D image data [4,8,15]. While the simplicity of these approaches is appealing, the applicability of these methods is constrained by their performance and are more suitable as initialization.

In this paper, we propose a purely self-supervised and patient-specific end-to-end framework for fully automatic registration of single-view X-ray to preoperative CT. Our main contributions are as follows: (1) The proposed method eliminates the need for manual annotation, relying instead on self-supervision from simulated patient-specific X-rays and corresponding automatically labeled poses, which makes the registration method easier to extend to new medical applications. (2) Regularized autoencoder and multi-head self-attention mechanism are embedded to encourage the model to capture patient-specific salient information automatically, therefore improving the robustness of registration; and an

novel refinement model is proposed to further improve registration accuracy. (3) The proposed method has been successfully evaluated on X-rays, achieving an average run-time of around 2.5 s, which meets the requirements for clinical applications.

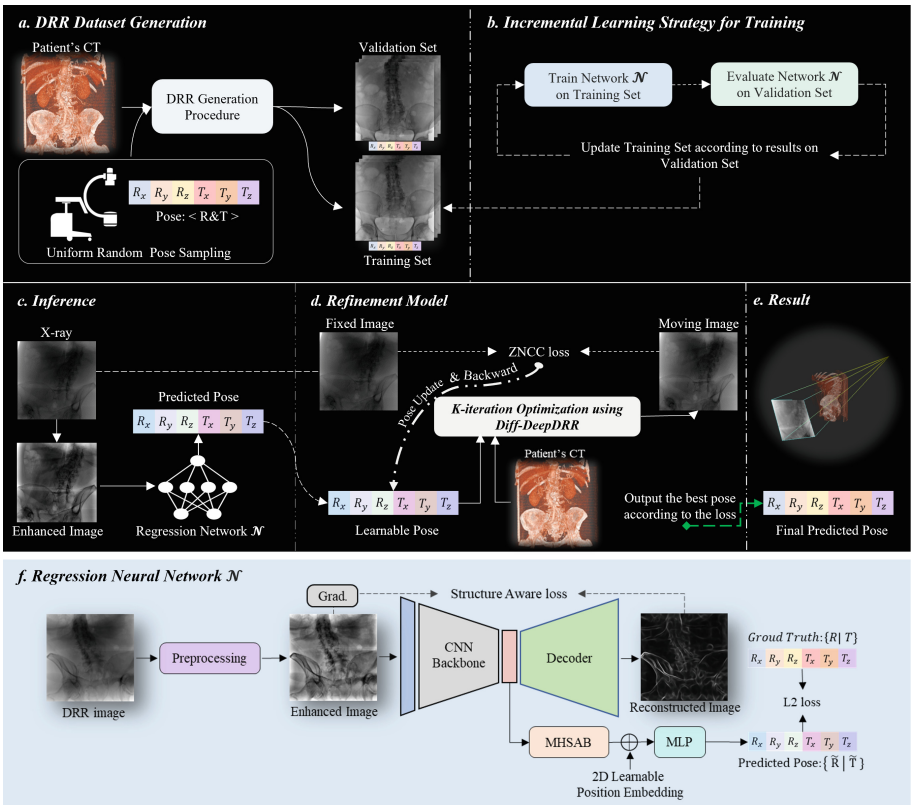

Fig. 1. An overview of our proposed framework (a-e) and the proposed regression neural network (f). The framework consists of five parts: (a) DRR generation, (b) Incremental Learning strategy for training, (c) inference phase, (d) refinement model, and (e) outputting the predicted pose.

2 Method

An overview of the proposed framework is illustrated in Fig. 1, which can be mainly divided into two parts that are introduced hereafter.

2.1 Pose Regression Model

Network Architecture. A regression neural network architecture is developed to estimate the six degrees of freedom (DOF) pose from an input X-ray

image, which consists of a regularized autoencoder, a multi-head self-attention block (MHSAB), a learnable 2D position embedding, and a multilayer perceptron (MLP), as shown in Fig. 1*f*.

Regularized Autoencoder. The autoencoder consists of a carefully selected backbone and a decoder. The first six layers of EfficientNet-b0 [16] are chosen as the backbone, where the first convolution layer is adapted by setting the input channel to 1 (shown as blue trapezoidal block in Fig. 1*f*). For image-based pose estimation, edge, as an important structural information, is more advantageous than intensity value. Hence, a structure aware loss function $\mathcal{L}_s$ is defined based on zero-normalized cross-correlation(ZNCC) to constrain the features extracted by encoder to contain necessary structure information, which is formulated as,

$$\mathcal{L}_s\left(G_I, Y\right) = \frac{1}{\sigma_{G_I}\sigma_Y}\frac{1}{N}\sum_{u,\ v}\left(G_I\left(u,v\right) - E\left(G_I\right)\right)\left(Y\left(u,v\right) - E(Y)\right) \tag{1}$$

$$G_I = \sqrt{\left(\frac{\partial I}{\partial u}\right)^2 + \left(\frac{\partial I}{\partial v}\right)^2} \tag{2}$$

Here, Y is the output of the decoder, I is the input image, G_I is the normalized gradient of the input image, and N is the number of pixels in the input image. $E()$ is the expectation operator, and σ is standard deviation operator.

Multi-head Self-attention Block and Positional Embedding. First, inspired by [19], a 3×1 convolutional layer, a 1×3 convolutional layer, and a 1×1 convolutional layer are used to generate the query ($q \in \mathcal{R}^{n\times hw\times C/n}$), key ($k \in \mathcal{R}^{n\times hw\times C/n}$), and value ($v \in \mathcal{R}^{n\times hw\times C/n}$) representations in the features ($f \in \mathcal{R}^{h\times w\times C}$) extracted by backbone respectively, which capture the edge information in the horizontal and vertical orientation. And the attention weight ($A \in \mathcal{R}^{n\times hw\times hw}$) is computed by measuring the similarity between q and k according to,

$$A = Softmax\left(\frac{qk^T}{\sqrt{C/n}}\right) \tag{3}$$

where n is the number of heads, C is number of channels, and h, w are the height and width of features. Using the computed attention weights, the output of MHSAB f_a is computed as,

$$f_a = Av + f \tag{4}$$

Second, in order to make the proposed method more sensitive to spatial transformation, 2D learnable positional embedding is employed to explicitly incorporate the object's position information.

Incremental Learning Strategy. Based on Incremental Learning strategy, the network is trained for 100 epochs with a batch size of 32 using the Adam optimizer (learning rate is 0.002, decay of 0.5 per 10 epochs). After training for 40 epochs, the training dataset will be automatically updated if the loss computed on the validation set does not change frequently (i.e., less than 10 percent of the maximum of loss) within 20 epochs, which allows the network to observe a wider range of poses while avoiding over-fitting. Final loss function is a weighted sum of the structure aware loss $\mathcal{L}_s$ for autoencoder and $L2$ Loss between the predicted pose and the ground truth, which is formulated as,

$$loss\,(I,p) = \|P(I) - p\|^2 + \alpha \mathcal{L}_s\,(G_I,\ D(I)) \tag{5}$$

where α is set as 5, I is the input DRR image, p is the ground truth of pose, $P\,(I)$ is predicted pose, and $D(I)$ is the output of autoencoder.

Pre-processing and Data Augmentation. In order to improve the contrast of the digitally reconstructed radiographs (DRRs) and reduce the gap to real X-rays, the DRR is first normalized by Z-score and then normalized using the sigmoid function, which maps the image into the interval [0, 1] with a mean of 0.5. Then contrast-limited adaptive histogram equalization (CLAHE) is employed to enhance the contrast. For data augmentation, random brightness adjustment, random adding offset, adding Gaussian noise, and adding Poisson noise are adopted.

2.2 Refinement Model

A novel refinement model is proposed to further refine the pose predicted by regression network as shown in Fig. 1(d). Specifically, inspired by DeepDRR [18], a differentiable DeepDRR (Diff-DeepDRR) generator is developed. Our proposed Diff-DeepDRR generator offers two main advantages compared with DeepDRR, including the ability to generate a 256^2 pixel DRR image in just 15.6 ms, making it fast enough for online refinement, and providing a way to obtain gradients with respect to the pose parameters. The refinement model takes the input X-ray image with an unknown pose as the fixed image and has six learnable pose parameters initialized by the prediction of pose regression network. Then the proposed Diff-DeepDRR generator utilizes the six learnable pose parameters to generate DRR image online, which is considered as the moving image. The ZNCC is then used as the loss function to minimize the distance between the fixed image and the moving image. With the powerful PyTorch auto-grad engine, the six pose parameters are learned iteratively. For each refinement process, the refinement model is online optimized for 100 iterations using an Adam optimizer with a learning rate of 5.0 for translational parameters and 0.05 for rotational parameters, and outputs the pose with the minimal loss score.

3 Experiments

3.1 Dataset

Our method is evaluated on six DRR datasets and one X-ray dataset. The six DRR datasets are generated from five common preoperative CTs and one specific CT with previous surgical implants, while the X-ray dataset is obtained from a Pelvis phantom containing a metal bead landmark inside. The X-ray dataset includes a CBCT volume and ten X-ray images with varying poses. For each CT or CBCT, 12800 DRR images with a reduced resolution of 256^2 pixels are generated using the DeepDRR method, which are divided into training (50%), validation (25%), and test (25%) sets. The DRR generation system is configured based on the geometry of a mobile C-arm, with a 432 mm detector, 0.3 mm pixel size, 742.5 mm source-isocenter distance, and 517.15 mm detector-isocenter distance. The center of the CT or CBCT volume is moved to the isocenter of the device. For pose sampling, each 6 DOF pose consists of three rotational and three translational parameters. The angular and orbital rotation are uniformly sampled from $[-40°, 40°]$, and the angle of in-plane rotation in the detector plane is uniformly sampled from $[-10°, 10°]$. Translations are uniformly sampled from [−70 mm, 70 mm].

3.2 Experimental Design and Evaluation Metrics

To better understand our work, we conduct a series of experiments, for example, 1) A typical pose regression model consists of a CNN backbone and an MLP. To find an efficient backbone for this task, EfficientNet, ResNet [5], and DenseNet [6] were studied. 2) A detailed ablation experiment was performed, where you can learn the evolution process and the superiority of our proposed method. 3) Through the experiment on X-ray data, you can know whether the proposed method trained only on DRRs can be generalized to X-ray applications. To validate the performance of our method on DRR datasets, we employed five measurements including 2D mean projection distance (mPD), 3D mean target registration error (mTRE) [7], Mean Absolute Error (MAE), and success rate (SR) [2], where success is defined as an mTRE of less than 10 mm. Cases with mTRE exceeding 10 mm were excluded from average mPD and mTRE measurements. For the validation on X-rays, projection distance (PD) is used to measure the positional difference of the bead between the X-ray and the final registered DRR. In addition, NCC [13], structural similarity index measure (SSIM), and contrast-structure similarity (CSS) [9] are employed to evaluate the similarity between the input X-ray and final registered DRR.

4 Results

4.1 The Choice of Backbone and Ablation Study

The performance of three models with different backbones were evaluated on the first DRR dataset as reported in Table 1. ***Res-backbone*** [15] means that the first

Table 1. The experimental results of different backbones. Rx, Ry, Rz, Tx, Ty and Tz are the mAE measured on three rotational parameters and three translational parameters.

backbone	mTRE↓ (mm)		mPD↓ (mm)		Rx↓ (degree)		Ry↓ (degree)		Rz↓ (degree)		Tx↓ (mm)		Ty↓ (mm)		Tz↓ (mm)		FLOPs (G)	Paras (M)	SR↑ (%)
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std			
Res-backbone [15]	7.68	3.08	7.63	2.05	1.39	1.13	1.31	1.05	0.99	0.88	3.06	2.63	3.03	2.45	4.84	3.70	17.2	9.05	67.13
Dense-backbone	6.63	2.57	6.94	2.10	1.11	0.89	1.28	0.93	0.90	0.72	2.48	2.22	2.46	2.14	4.33	3.28	13.9	4.78	81.94
Ef-backbone	6.71	2.75	6.83	2.05	0.98	0.78	0.93	0.76	0.71	0.56	2.44	2.05	2.57	1.99	3.92	3.05	1.05	0.91	**88.66**

Table 2. The results of ablation study & our method's results on 6 DRR datasets. # indicates previous surgical implants are included in this dataset.

–	mTRE↓ (mm)		mPD↓ (mm)		Rx↓ (degree)		Ry↓ (degree)		Rz↓ (degree)		Tx↓ (mm)		Ty↓ (mm)		Tz↓ (mm)		SR↑ (%)
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	
A	6.71	2.75	6.83	2.05	0.98	0.78	0.93	0.76	0.71	0.56	2.44	2.05	2.57	1.99	3.92	3.05	88.66
A+B	6.13	2.50	6.39	2.07	0.93	0.78	0.99	0.79	0.62	0.52	2.10	1.82	2.33	1.91	3.49	2.73	90.97
A+C	6.17	2.65	6.12	1.98	1.01	0.80	0.85	0.68	0.70	0.58	1.99	1.75	2.05	1.74	3.72	2.88	91.34
A+B+C	5.92	2.57	6.01	2.02	1.02	0.80	0.83	0.65	0.64	0.54	1.92	1.75	2.05	1.78	3.39	2.71	93.31
[A+B+C]*	5.43	2.18	5.80	2.05	0.90	0.73	0.82	0.65	0.57	0.47	1.80	1.53	1.80	1.50	2.89	2.26	95.25
[A+B+C]*+D proposed	**2.85**	**1.91**	**2.92**	**1.93**	**0.41**	**0.53**	**0.31**	**0.39**	**0.24**	**0.35**	**1.00**	**0.83**	**0.90**	**0.84**	**1.55**	**1.35**	**100.00**
dataset1	2.85	1.91	2.92	1.93	0.41	0.53	0.31	0.39	0.24	0.35	1.00	0.83	0.90	0.84	1.55	1.35	100.00
dataset2	2.54	1.28	3.08	1.72	0.38	0.55	0.35	0.38	0.22	0.29	0.65	0.60	0.90	0.51	1.38	0.88	100.00
dataset3	2.71	1.14	3.30	1.45	0.26	0.31	0.33	0.40	0.23	0.26	0.84	0.64	1.04	0.71	1.51	1.09	100.00
dataset4	2.72	1.32	3.12	1.53	0.27	0.48	0.26	0.33	0.23	0.32	0.90	0.86	1.04	0.66	1.58	1.10	100.00
dataset5	2.73	1.03	3.09	1.22	0.35	0.35	0.35	0.33	0.24	0.22	0.74	0.69	0.97	0.58	1.52	1.09	100.00
dataset6#	2.50	1.48	2.57	1.67	0.36	0.43	0.27	0.39	0.20	0.27	0.75	0.64	0.86	1.00	1.44	1.20	100.00

four layers of ResNet-50 as the backbone; ***Dense-backbone*** means that the first six layers of DenseNet-121 as the backbone; ***Ef-backbone*** means that the first six layers of EfficientNet-b0 as the backbone. Compared with the other two networks, ***Ef-backbone*** achieves a higher SR of 88.66%, and reduces parameter size and FLOPs by an order of magnitude. Therefore, ***Ef-backbone*** is chosen and regarded as baseline for further studies. Then, a second set of models was trained and evaluated for a detailed ablation study, and the experimental results are shown in Table 2. ***A*** means the aforementioned baseline; ***B*** means adding regularized autoencoder; ***C*** means adding multi-head self-attention block and position embedding; ***D*** means using refinement model; ******* means the network is trained via Incremental Learning strategy. It is clear that each module we proposed plays a positive role in this task, and our method makes arresting improvements, e.g., compared with baseline, the SR increased from 88.66% to 100%. In addition, observing the visualized self-attention map shown in Fig. 2(b), we find that the proposed network does automatically capture some salient anatomical regions, which pays more attention to bony structures. Meanwhile, the distribution of attention changes with the view pose, and it seems that the closer to the detector, the bone region gets more attention.

4.2 Experimental Results on DRR Datasets and X-Ray Dataset

The experimental results of our proposed method on six DRR datasets are shown in Table 2. It is worth noticing that the success rate of our method has achieved 100% on all datasets and the average of mTRE on six datasets achieves 2.67 mm.

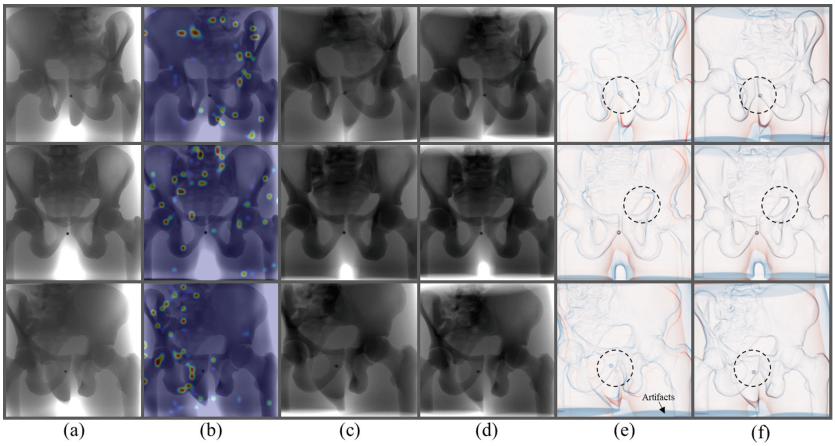

Fig. 2. Results on X-ray cases. (a) X-rays, (b) corresponding attention visualization, (c) initial registered DRRs from pose regression network, (d) final registered DRRs from refinement mode, (e) initial registration error map, (f) final registration error map. Red edges in (e&f) come from (a), green edge in (e) comes from (c), green edge in (f) comes from (d). (Color figure online)

Table 3. The results of our method on X-ray Cases

–	Case1	Case2	Case3	Case4	Case5	Case6	Case7	Case8	Case9	Case10	Mean	Std
PD↓	0.6762	1.8400	1.4251	2.8431	0.6770	1.3244	1.2777	1.4181	1.5259	2.5655	1.5573	0.6687
NCC↑	0.9885	0.9752	0.9867	0.9941	0.9858	0.9870	0.9888	0.9943	0.9913	0.9880	0.9880	0.0051
SSIM↑	0.9395	0.9220	0.9348	0.9736	0.9424	0.9456	0.9436	0.9616	0.9649	0.9412	0.9469	0.0146
CSS↑	0.9427	0.9321	0.9392	0.9750	0.9453	0.9481	0.9463	0.9630	0.9664	0.9448	0.9503	0.0127

For the X-ray dataset, the quantitative evaluation results on 10 X-ray cases are shown in Table 3, achieving a mPD of 1.55 mm, which demonstrates that our method can be successfully generalized to X-ray. More intuitive results are shown in Fig. 2. Observing the dotted circled areas on Fig. 2(e&f), we find that the network only makes coarse predictions for X-rays due to unavoidable artifacts on the CBCT boundaries, and the proposed refinement model facilitates accurate pose estimation, which can be confirmed by the overlapping of edges from the X-rays and the final registered DRRs.

5 Conclusion and Discussion

In this paper, we present a patient-specific and self-supervised end-to-end approach for automatic X-ray/CT rigid registration. Our method effectively addresses the primary limitations of existing methods, such as requirement of manual annotation, dependency on conventional derivative-free optimization, and patient-specific concerns. When field of view of CT is not smaller than that

of X-ray, which is often satisfied in clinical routines, our proposed method would perform very well without any additional post-process. The quantitative and qualitative evaluation results of our proposed method illustrates its superiority and its ability to generalize to X-rays even when trained solely on DRRs. For our experiments, the validation on X-ray of phantom cannot fully represent the performance on X-ray of real patients, but it shows that the proposed method has high potential. Meanwhile, domain randomization could reduce the gap between DRR and real X-ray images, which would allow methods validated on phantoms to perform better also on real X-ray. For the runtime aspect, our patient-specific regression network can complete the training phase within one hour using an NVIDIA GPU (Quadro RTX A6000), which meets the requirement of clinical application during pre-operative planning phase. Meanwhile, the proposed network achieves an average inference time of 6ms per image with a size of 256^2, when considering the run-time of the proposed refinement model, the total cost is approximately 2.5 s, which also fully satisfies the requirement for clinical application during intra-operative phase.

Acknowledgements. The project was supported by the Bavarian State Ministry of Science and Arts within the framework of the "Digitaler Herz-OP" project under the grant number 1530/891 02 and the China Scholarship Council (File No.202004910390). We also thank BrainLab AG for their partial support.

References

1. Bier, B., et al.: X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018, Part IV. LNCS, vol. 11073, pp. 55–63. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_7
2. Grimm, M., Esteban, J., Unberath, M., Navab, N.: Pose-dependent weights and domain randomization for fully automatic X-ray to CT registration. IEEE Trans. Med. Imaging **40**(9), 2221–2232 (2021)
3. Grupp, R.B., et al.: Automatic annotation of hip anatomy in fluoroscopy for robust and efficient 2D/3D registration. Int. J. Comput. Assist. Radiol. Surg. **15**, 759–769 (2020)
4. Guan, S., Meng, C., Sun, K., Wang, T.: Transfer learning for rigid 2D/3D cardiovascular images registration. In: Lin, Z., Wang, L., Yang, J., Shi, G., Tan, T., Zheng, N., Chen, X., Zhang, Y. (eds.) PRCV 2019, Part II. LNCS, vol. 11858, pp. 380–390. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-31723-2_32
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
7. Van de Kraats, E.B., Penney, G.P., Tomazevic, D., Van Walsum, T., Niessen, W.J.: Standardized evaluation methodology for 2-D-3-D registration. IEEE Trans. Med. Imaging **24**(9), 1177–1189 (2005)

8. Lee, B.C., et al.: Breathing-compensated neural networks for real time C-arm pose estimation in lung CT-fluoroscopy registration. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2022)
9. Liu, S., et al.: Unpaired stain transfer using pathology-consistent constrained generative adversarial networks. IEEE Trans. Med. Imaging **40**(8), 1977–1989 (2021)
10. Markelj, P., Tomaževič, D., Likar, B., Pernuš, F.: A review of 3D/2D registration methods for image-guided interventions. Med. Image Anal. **16**(3), 642–661 (2012)
11. Meng, C., Wang, Q., Guan, S., Sun, K., Liu, B.: 2D-3D registration with weighted local mutual information in vascular interventions. IEEE Access **7**, 162629–162638 (2019)
12. Miao, S., et al.: Dilated FCN for multi-agent 2D/3D medical image registration. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
13. Penney, G.P., Weese, J., Little, J.A., Desmedt, P., Hill, D.L., et al.: A comparison of similarity measures for use in 2-D-3-D medical image registration. IEEE Trans. Med. Imaging **17**(4), 586–595 (1998)
14. Powell, M.J.: The BOBYQA algorithm for bound constrained optimization without derivatives, vol. 26. Cambridge NA Report NA2009/06, University of Cambridge, Cambridge (2009)
15. Salehi, S.S.M., Khan, S., Erdogmus, D., Gholipour, A.: Real-time deep pose estimation with geodesic loss for image-to-template rigid registration. IEEE Trans. Med. Imaging **38**(2), 470–481 (2018)
16. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
17. Unberath, M., et al.: The impact of machine learning on 2D/3D registration for image-guided interventions: a systematic review and perspective. Front. Robot. AI **8**, 716007 (2021)
18. Unberath, M., et al.: DeepDRR – a catalyst for machine learning in fluoroscopy-guided procedures. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018, Part IV. LNCS, vol. 11073, pp. 98–106. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_12
19. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)