Training Deep Networks on Domain Randomized Synthetic X-ray Data for Cardiac Interventions

Daniel Toth[1,2] DANIEL.TOTH@KCL.AC.UK
[1] *Siemens Healthineers, Frimley, UK*
[2] *School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK*

Serkan Cimen[3]
Pascal Ceccaldi[3]
[3] *Siemens Healthineers, Medical Imaging Technologies, Princeton, NJ, USA*

Tanja Kurzendorfer[4]
[4] *Siemens Healthineers, Forchheim, Germany*

Kawal Rhode[*2]

Peter Mountney[*3]

Abstract

One of the most significant challenges of using machine learning to create practical clinical applications in medical imaging is the limited availability of training data and accurate annotations. This problem is acute in novel multi-modal image registration applications where complete datasets may not be collected in standard clinical practice, data may be collected at different times and deformation makes perfect annotations impossible. Training machine learning systems on fully synthetic data is becoming increasingly common in the research community. However, transferring to real world applications without compromising performance is highly challenging. Transfer learning methods adapt the training data, learned features, or the trained models to provide higher performance on the target domain. These methods are designed with the available samples, but if the samples used are not representative of the target domain, the method will overfit to the samples and will not generalize. This problem is exacerbated in medical imaging, where data of the target domain is extremely scarse. This paper proposes to use Domain Randomization (DR) to bridge the reality gap between the training and target domains, requiring no samples of the target domain. DR adds unrealistic perturbations to the training data, such that the target domain becomes just another variation. The effects of DR are demonstrated on a challenging task: 3D/2D cardiac model-to-X-ray registration, trained fully on synthetic data generated from 1711 clinical CT volumes. A thorough qualitative and quantitative evaluation of transfer to clinical data is performed. Results show that without DR training parameters have little influence on performance on the training domain of digitally reconstructed radiographs, but can cause substantial variation on the target domain (X-rays). DR results in a significantly more consistent transfer to the target domain.

Keywords: Domain Randomization, Imitation Learning, Cardiac Registration.

1. Introduction

In recent years, deep learning algorithms have been successfully applied to a wide variety of medical imaging tasks (Litjens et al., 2017). However, for many medical imaging problems, labeled

[*] Joint senior authors

training data is extremely scarce. Especially in novel clinical procedures, new, procedure-specific, retrospectively not available data might be required. Initially, data is available only in highly limited quantities and it can accumulate very slowly to usable amounts. Multimodal datasets, e.g., for image registration, can be even more challenging to acquire and previously collected data is highly unlikely to be multimodal. Additionally, Ground Truth (GT) annotations can be highly challenging to acquire. The annotation needs to be performed by experts in the respective domain, i.e., by clinicians. This can be time consuming and expensive for even smaller amounts of data, but might not be feasible for larger datasets. In certain applications, it might be impossible to acquire GT annotations, e.g., due to deformations in multimodal data acquired at different time points.

A system for registering 3D preoperative data to 2D X-rays (Toth et al., 2018) demonstrated the feasibility and benefits of training purely on synthetic data. The system generated synthetic images from a CT dataset, thus knowing the GT transformations. These synthetic X-ray images are commonly referred to as Digitally Reconstructed Radiographs (DRRs) (Russakoff et al., 2005). The registration system trained on DRR data has shown good accuracy on the same domain and acceptable performance on clinical X-ray data. However, transfer performance is highly affected by training parameters: different local minima might be reached, resulting in highly similar performance on the synthetic training domain, but highly variable results on the target domain. DRR generation algorithms do not model every aspect of the X-ray formation. This causes images to appear differently than clinically acquired X-rays. Algorithms used for image guidance have to be extremely robust against perturbations. In the case of the X-ray data, these could be intensity variations, due to different dose settings, variation of the collimation, or devices in the field of view. These variations can appear differently, or may not be present, in the artificially generated DRRs. The transfer, thus the robustness of the registration can suffer. The perfomance reduction after transferring to the target domain is called the reality gap.

Several approaches have addressed this issue including (Heimann et al., 2014), where a classifier is adapted to the target domain, without requiring labelled data on the target domain. However, it could fail, if the training and target distributions are far. Domain adaptation can also be performed through task-driven Generative Adversarial Networks (GANs) (Zhang et al., 2018). GANs, however, can be difficult to train, can be unstable, can show mode collapse. Advanced simulation methods can be used to generate higher quality DRRs that appear more similar to the target domain's images, such as (Unberath et al., 2018). Such an approach is still not a perfect X-ray simulator, important features might not be simulated. The collective weakness of these approaches is that they try to adapt/fit to samples of the target domain. If the samples are not representative of the target domain, the adaptation will overfit to the samples. If such a system has to process data from a new site, device, or of unexpected imaging parameters, performance can decrease significantly.

This paper addresses the problem of overfitting to the target domain, by using Domain Randomization (DR) (Tobin et al., 2017). In DR the training data is augmented with unrealistic transformations. The transformations introduce such a large variation to the data that the target domain will appear as just another variation to the learning system. This approach was shown to help transfer to the target domain in non medical applications, e.g., autonomous driving (Tremblay et al., 2018) or robotic control (Peng et al., 2018).

In this paper, the usability of DR in the medical imaging context is demonstrated for the first time. Its benefits are being shown on the challenging cardiac image registration task described in (Toth et al., 2018). It is shown that, if trained only on artificially generated data (DRRs), DR can greatly increase robustness against geometric perturbations. Furthermore, it is demonstrated that,

witouth DR, variation of training parameters can result in inconsistent results on the target domain. Performance with DR, however, appears to be agnostic to variation in training parameters, thus produces more consistent results.

2. Methods

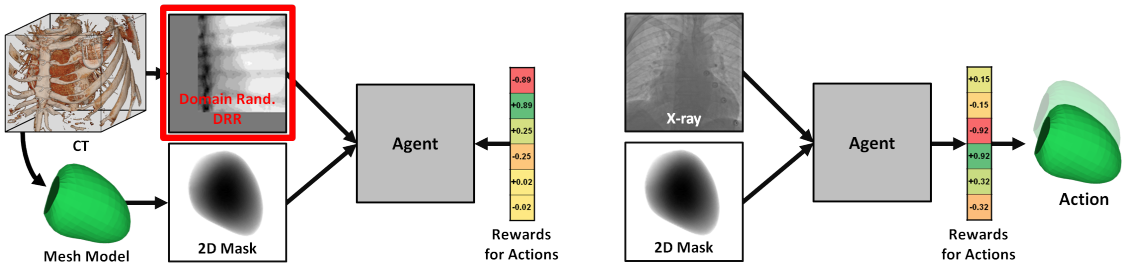
The registration framework with domain randomization is illustrated in Figure 1. The model-to-X-ray registration system from (Toth et al., 2018) is intended to be used for 3D/2D cardiac registration. It is trained purely with CT volumes, without any real X-rays from the target domain. The CTs are used for: 1) DR DRRs and 2) to segment, perturb, and project masks of the Left Ventricle (LV) into 2D. Since both images are generated, the GT transformation is known between them. This is used to compute rewards for each possible action of the three degrees of freedom given by the imaging plane: translations x, y and rotation about the z axis. The rewards and the two 2D images (DR DRR and mask image) are used to train an artificial agent represented by a Convolutional Neural Network (CNN), see Figure 1(*a*). For inference, the artificial agent is being shown the two images and is able to predict the rewards for each possible action. The action with the highest reward is chosen and is applied to the 3D model, as depicted in Figure 1(*b*).

2.1. Transfer Learning

The training of the artificial agent is performed on fully synthetic data. It has shown acceptable performance in terms of accuracy on the synthetic domain and robustness on the target domain. However, for interventional applications, highest levels of robustness are crucial. How well the imitation learning agent transfers to real data is influenced by three main factors: 1) the appearence of the training data (similarity to target data), 2) the weight initialization of the neural network, and 3) the training data order.

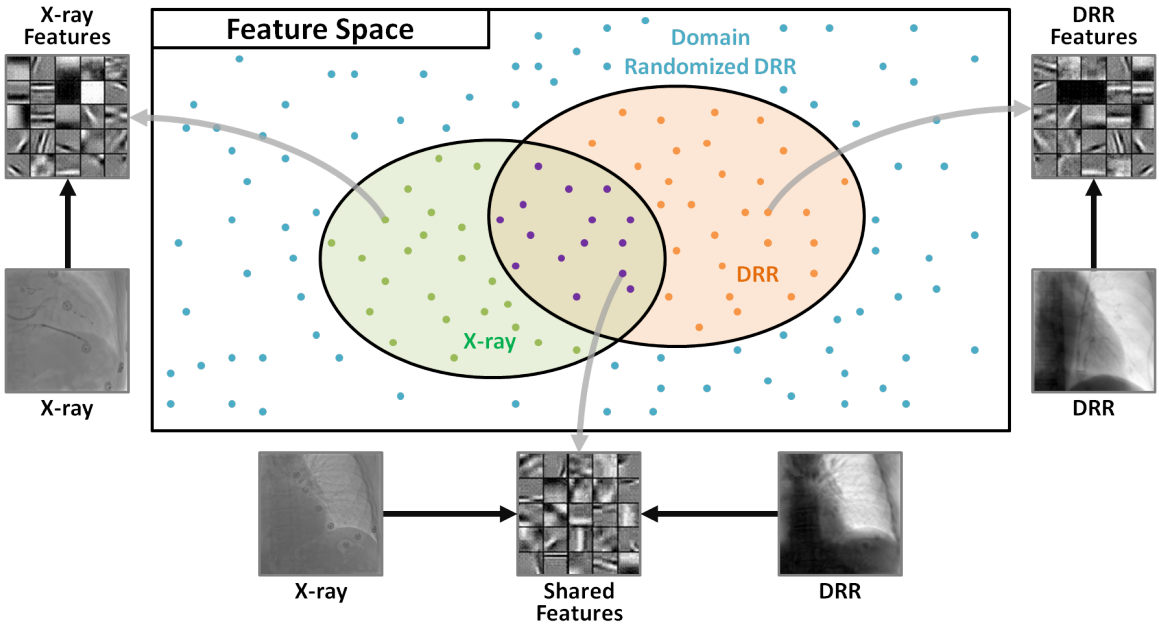
The similarity between the training domain and the target domain is the main factor in performance after transfer. If the training DRRs are sufficiently similar to the target X-ray images, the network will transfer with similar performance. This can be achieved to a certain degree by optimizing DRR generation parameters, but it is highly challenging and the learned network might not generalize well to unseen images.

The second factor is the initialization of the weights in the neural network representing the agent. The initial weights define the starting position in the optimization space of training. From

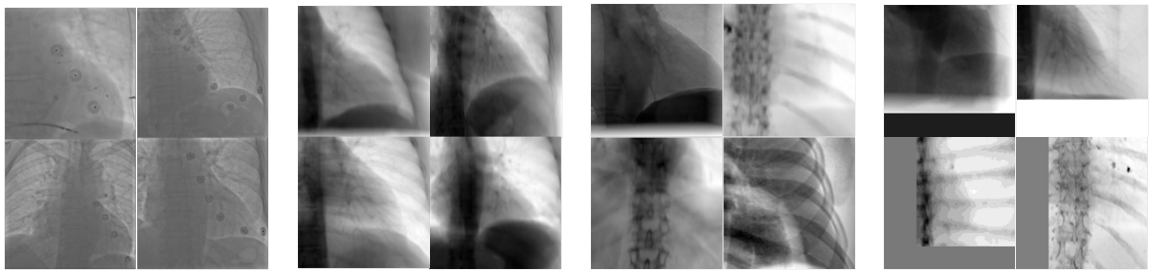


(*a*) Training workflow (*b*) Evaluation workflow

Figure 1: Overview of registration framework workflows.



(*a*) The concept of domain randomization in feature space. Every point represents a sample (a set of features) in the respective domain (green: X-ray, orange: DRR, purple: shared features, blue: DR DRR).



(*b*) X-rays (*c*) DRRs (*d*) DR intensity (*e*) DR int. + coll.

Figure 2: The concept of domain randomization in feature space and sample images.

different starting positions, different local minima can be reached that can result in highly similar results on the training, but inconsistent results on the target domain.

The effects of the variation of the training data order can have similar influence. If the same weights are used for initialization, but the network sees the training data in a different order, the optimization can take different steps and can end up in different local minima, resulting in different results on the target domain.

2.2. Domain Randomization

Domain randomization (Tobin et al., 2017) is an approach to transfer a deep neural network trained on only simulated images to the real world. Instead of high quality simulation images, domain randomization relies on the variability of the low quality simulation images. The main idea is to

randomize the simulator and generate a large variation of simulated images, such that the real images become just another variation.

For the present problem, instead of using a realistic X-ray simulator, it is proposed to use a simple ray tracing-based DRR renderer. The concept can be illustrated in feature space by a Venn diagram, see Figure 2(*a*). The sets of X-ray features and DRR features have a substantial intersection, because the method transfers (with limited performance) to the X-ray domain, if trained only on DRRs. If there were no shared features, no intersection of the sets, the approach would not transfer at all. To ensure a more successful transfer, the DRR generation is randomized that the resulting features cover a large portion of the whole feature space (blue in Figure 2(*a*)). If the DR images cover most of the feature space, the approach will generalize to the X-rays too, even if no X-ray similar images are generated.

The DRR generation is performed with randomized 1) Hounsfield Unit (HU) values of the 3D CT data and 2) collimation added to the 2D projections. Before computing the ray tracing, an intensity mapping to the HU values of the CT data was applied either globally or locally. For global randomization, a non-linear function (cumulative distribution function of a beta distribution) was applied to transform the voxel values, where the parameters α and β are randomly sampled from a uniform distribution of $[0.5;5.0]$. For the local randomization, the intensity range of the CT data was subdivided into 5 to 15 non-overlapping ranges. The start and end of the intervals was defined by evenly spacing the whole range, then a random perturbation was added to the start/end values sampled from a Gaussian distribution (mean = 80 and standard deviation = 40). For each intensity range, one of three randomization options was chosen: 1) the non-linear mapping described above; 2) invert the intensity range; 3) shift the intensity values by adding a random value of a uniform distribution on the interval $[-100;100]$. This results in images shown in Figure 2(*d*).

The acquired X-ray images of the target domain are often collimated. Thus, in addition to the intensity randomization, bands of random intensity were added to the generated images, see Figure 2(*e*). The intensity was varied in the range of possible intensity values and the size of the borders was maximally 25 % of the image size in the respective direction.

3. Experiments and Results

The effects of domain randomization are demonstrated on the image registration task with an imitation learning agent, described in Section 2. The agent was trained only on synthetic data generated from 1711 clinical CT volumes of 799 patients. The evaluation was performed in two steps, on two datasets: 1) synthetic data from the same domain as the training data originates from; 2) clinical patient data, from the target domain.

The effect of DR is presented for a certain set of seed values for 1) random weight initialization and 2) random data shuffling. Then the seeds are varied to demonstrate the inconsistency of the transfer from the synthetic training domain to the real target domain and to show the consistency of the domain randomization results. An ablation study was also performed to investigate the influence of each DR.

3.1. Evaluation Data

The synthetic test dataset consists of 100 CT volumes of 100 patients. There is no overlap between training and test patients. The 3D mesh model of each patient was perturbed 10 times sampled from a uniform distribution, resulting in 1000 artificial test cases. Due to the known perturbation, the GT

registration is available. Accuracy is measured by the points of a 3D landmark at the center of the LV model.

The clinical data consists of 21 patient datasets. Each patient dataset has one MR volume of that the mesh model was extracted and an X-ray image to register to. The mesh model is perturbed 169 times in a rectangular grid manner about an approximate, manual registration. There is no GT registration available for this dataset. The registration is evaluated quantitatively for robustness, i.e., that it provides the same result from different perturbations. The robustness error measure was defined as described in (Toth et al., 2018):

$$e_f = \|\boldsymbol{x}_f - \tilde{\boldsymbol{x}}_f\|_2, \tag{1}$$

where $\tilde{\boldsymbol{x}}_f$ is the median final position of the center of the cross landmark of all perturbations and $\boldsymbol{x}_f$ is the final position for a single perturbation. Rotational robustness ε_f was computed in the same manner. Accuracy is only verified qualitatively, by comparing the heart shadow in the X-rays and the edges of the projected LV mask.

Table 1: Accuracy in mm on synthetic DRR data of the standard (SN) and domain randomized networks (DRN) for varying weight initialization W_i and data order D_i.

(a) Weight initialization varied

		SN/DRN				
	Start	W_1, D_1	W_2, D_1	W_3, D_1	W_4, D_1	W_5, D_1
Mean	22.41	2.80/2.65	2.66/2.66	2.70/2.65	2.63/2.64	2.77/2.68
StD.	10.60	2.17/2.05	2.27/2.12	2.18/2.10	2.23/2.07	2.24/2.19
Median	21.27	2.15/2.16	1.99/2.14	2.15/2.14	1.96/2.08	2.23/2.14
90 %	37.12	5.54/5.00	5.50/5.34	5.31/5.33	5.26/5.07	5.52/5.24

(b) Data order varied

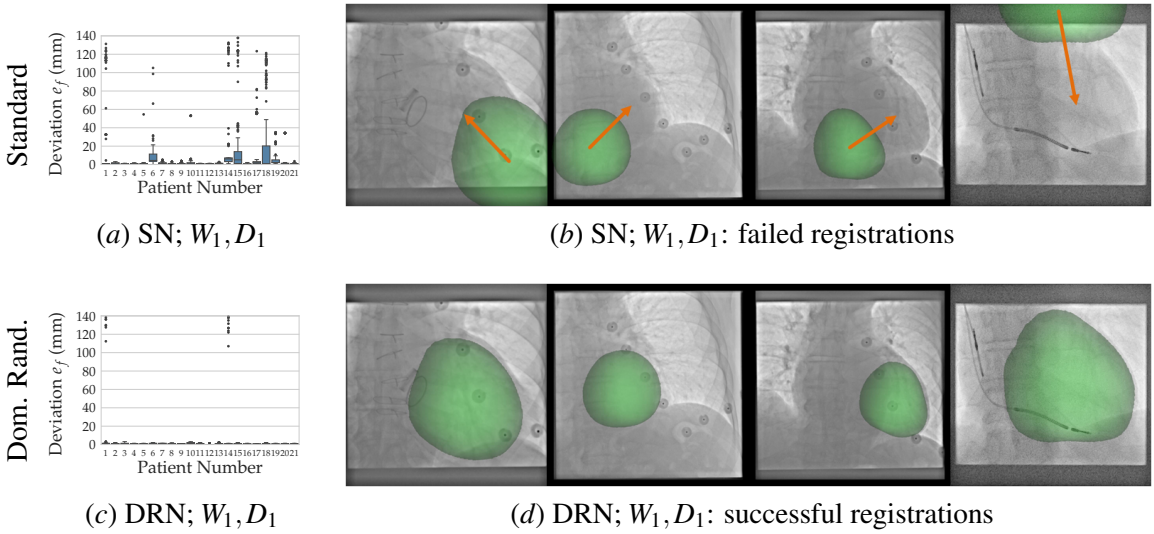
		SN/DRN				
	Start	W_1, D_1	W_1, D_2	W_1, D_3	W_1, D_4	W_1, D_5
Mean	22.41	2.80/2.65	2.78/2.58	2.64/2.59	2.80/2.78	2.70/2.62
StD.	10.60	2.17/2.05	2.19/1.98	2.17/2.05	2.21/2.23	2.19/2.07
Median	21.27	2.15/2.16	2.26/2.03	2.02/2.06	2.25/2.19	2.08/2.10
90 %	37.12	5.54/5.00	5.31/5.11	5.21/5.23	5.39/5.50	5.14/5.05

3.2. Baseline

The Standard Network (SN) was trained with a selection of seeds for data shuffling (D_1) and weight initialization (W_1). The training curves are marginally closer to each other for DR, see Appendix A. The synthetic results are slightly better, see Table 1, but the clinical data results are slightly worse than in (Toth et al., 2018), see Figure 3(a). This is due to training on significantly more data, causing the network to overfit more to the synthetic domain. The network with domain randomization (DRN) has the same architecture and was trained with the same parameters. On synthetic data, as expected, the results improve only slightly, because no transfer is required. The clinical data experiment has shown, however, that the distributions are compressed and the number of outliers is greatly reduced, see Figures 3(a) and 3(c). Most individual cases have improved greatly in the final alignment, for samples see Figures 3(b) and 3(d). The network has learned to handle intensity and collimation variations, although, it has never seen realistic transformations.

3.3. Parameter Variation

Two experiments were performed: 1) the seed for random network weight initialization was varied, with a fixed data order (W_i, D_1) and 2) the data shuffling seed was varied, with fixed weight



(*a*) SN; W_1, D_1

(*b*) SN; W_1, D_1: failed registrations

(*c*) DRN; W_1, D_1

(*d*) DRN; W_1, D_1: successful registrations

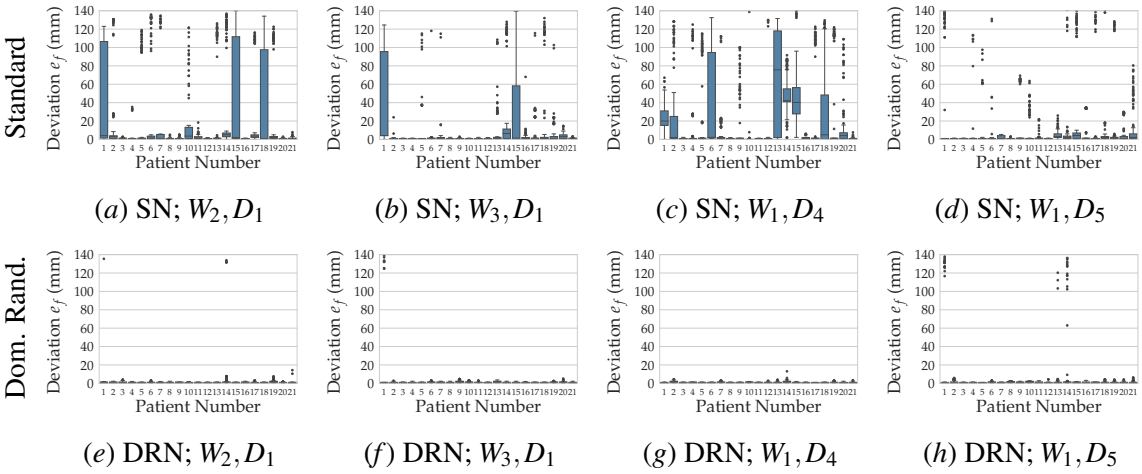
Figure 3: Effects of domain randomization on the target domain (clinical patient data). Baseline quantitative and qualitative results with fixed seeds (W_1 and D_1) of the (a-b) standard network (SN) and the (c-d) domain randomized network (DRN).

initialization (W_1, D_i). Both experiments had similar outcomes: Results on the synthetic data show that, as expected, all training parameters with both networks (SN and DRN) result in highly similar performance, because no domain transfer is required, see Table 1. In the case of the clinical X-ray data, the SN produces acceptable, but highly variable results. The DRN produces more consistent results, the distributions are compressed and the number of outliers is greatly reduced, see Figure 4 for a selection of setups and Figures 9 and 10 for all results.

If looking at individual cases, such as patients 3 and 8, good robustness and no major variation was present with the SN. The features of these images must be shared between the training and target domains, thus they can be registered without an explicit transfer. Cases such as 12 or 15 vary greatly with varying training parameters. The trained SN networks are in different minima of the optimization space, and the features required for performing well on these cases are learned only in some of the setups. A quantitative summary of the translational and rotational robustness results is displayed in Appendix B.

3.4. Ablation Study

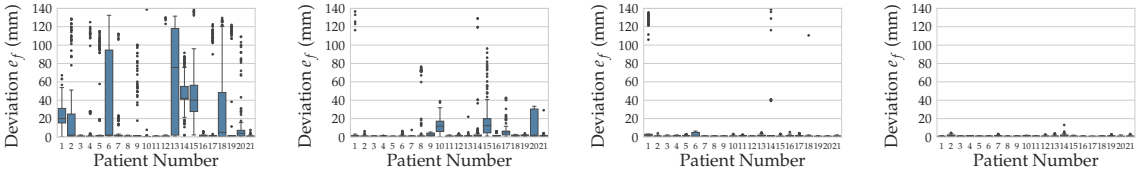
To investigate the effects of each DR, they were applied individually in a specific setup (W_1, D_4), see Figure 5 and Table 3. The translational robustness distributions have improved most significantly through the DR of the intensity. The DR of the collimation mainly suppresses extreme outliers. Combined DR is the most effective.



(*a*) SN; W_2, D_1 (*b*) SN; W_3, D_1 (*c*) SN; W_1, D_4 (*d*) SN; W_1, D_5

(*e*) DRN; W_2, D_1 (*f*) DRN; W_3, D_1 (*g*) DRN; W_1, D_4 (*h*) DRN; W_1, D_5

Figure 4: Translational robustness results e_f of weight initialization (W_i) and training data order (D_i) variation on clinical patient data. (a–d) Standard network (SN). (e–h) Domain randomized network (DRN).



(*a*) No DR (*b*) Collimation DR (*c*) Intensity DR (*d*) Coll. + int. DR

Figure 5: Effects of individual DRs on translational robustness e_f on clinical data for a single setup (W_1, D_4).

4. Conclusion

This paper addresses the performance reduction on clinical data of a cardiac image registration system, if trained only on artificially generated data. Through DR, DRRs are being generated by: 1) varying the intensity transfer function and 2) adding artificial image borders (representing collimation). It was shown, that the trained registration system has higher robustness against geometric perturbations in the test data. Furthermore, it was demonstrated that training with different parameters, i.e., weight initialization or data order, can result in varying results on the target (real) domain, if trained purely on synthetic data. If trained with DR, however, the results are more consistent, there is substantially less variation.

Acknowledgments

Concepts and information presented are based on research and are not commercially available. Due to regulatory reasons, the future availability cannot be guaranteed. This work was supported by the Wellcome EPSRC Centre for Medical Engineering at KCL (WT 203148/Z/16/Z) and the NIHR Biomedical Research Centre based at GSTT and KCL. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Dept. of Health.

References

Tobias Heimann, Peter Mountney, Matthias John, and Razvan Ionasec. Real-time ultrasound transducer localization in fluoroscopy images by transfer learning from synthetic training data. *Medical Image Analysis*, 18(8):1320–1328, 2014. ISSN 13618423. doi: 10.1016/j.media.2014.04.007.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42 (1995):60–88, 2017. ISSN 13618423. doi: 10.1016/j.media.2017.07.005.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. *arXiv*, page 1, 2018. ISSN 0364152X. doi: 10.1007/s00267-013-0043-7.

Daniel B. Russakoff, Torsten Rohlfing, Kensaku Mori, Daniel Rueckert, Anthony Ho, John R. Adler, and Calvin R. Maurer. Fast Generation of Digitally Reconstructed Radiographs Using Attenuation Fields With Application to 2D-3D Image Registration. *IEEE Transactions on Medical Imaging*, 24(11):1441–1454, 2005. doi: 10.1109/TMI.2005.856749.

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 *IEEE International Conference on Intelligent Robots and Systems*, pages 23–30, 2017. ISBN 1367-8868\n1469-8374. doi: 10.15607/RSS.2017.XIII.034. URL http://arxiv.org/abs/1611.04201.

Daniel Toth, Shun Miao, Tanja Kurzendorfer, Christopher A. Rinaldi, Rui Liao, Tommaso Mansi, Kawal Rhode, and Peter Mountney. 3D/2D model-to-image registration by imitation learning for cardiac procedures. *International Journal of Computer Assisted Radiology and Surgery*, 13(8): 1141–1149, 2018. ISSN 18616429. doi: 10.1007/s11548-018-1774-y. URL https://doi.org/10.1007/s11548-018-1774-y.

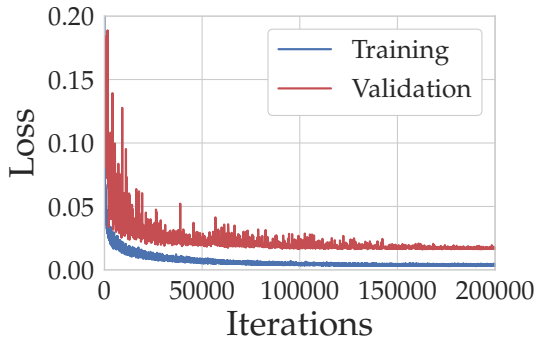
Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization. In *arXiv*, pages 1–9, 2018. doi: 10.1109/CVPRW.2018.00143. URL http://arxiv.org/abs/1804.06516.

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 *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial*

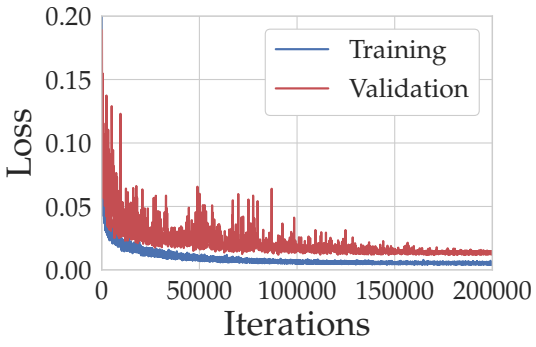
Intelligence and Lecture Notes in Bioinformatics), volume 11073 LNCS, pages 98–106, 2018. ISBN 9783030009366. doi: 10.1007/978-3-030-00937-3_12.

Yue Zhang, Shun Miao, Tommaso Mansi, and Rui Liao. Task driven generative modeling for unsupervised domain adaptation: Application to X-ray image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11071 LNCS, pages 599–607, 2018. ISBN 9783030009335. doi: 10.1007/978-3-030-00934-2_67.

Appendix A. Training Curves

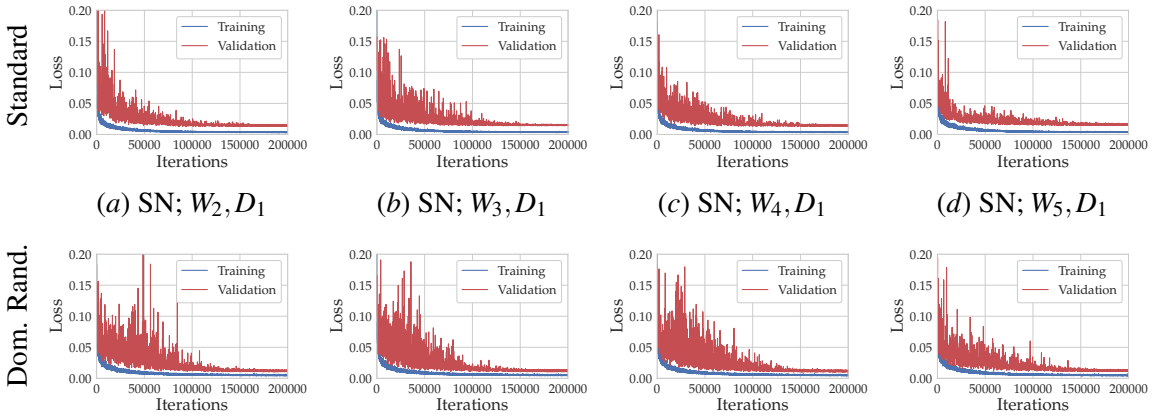


(*a*) SN; W_1, D_1



(*b*) DRN; W_1, D_1

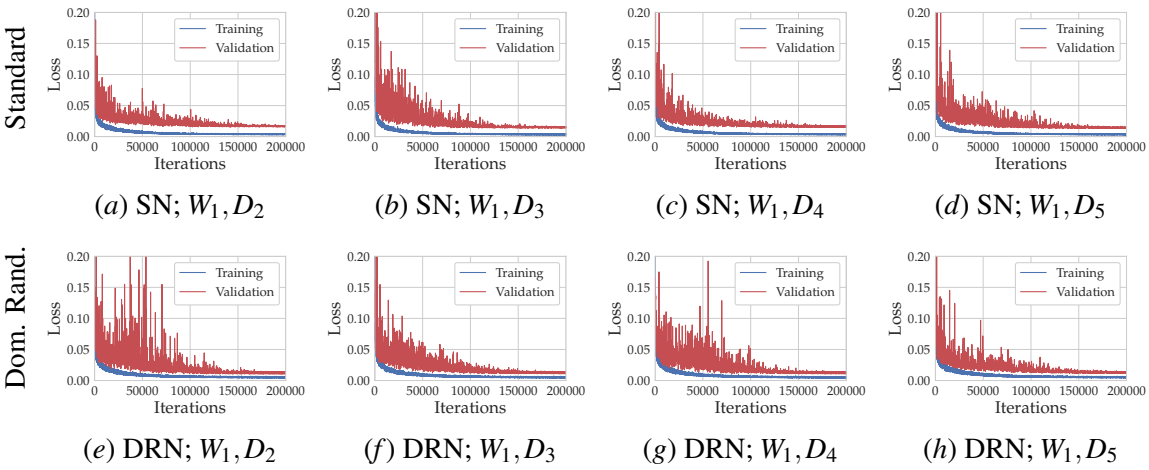
Figure 6: Training and validation curves of the standard (SN) and the domain randomized networks (DRN), with weight initialization and data order seeds, W_1 and D_1 respectively.



(*a*) SN; W_2, D_1 (*b*) SN; W_3, D_1 (*c*) SN; W_4, D_1 (*d*) SN; W_5, D_1

(*e*) DRN; W_2, D_1 (*f*) DRN; W_3, D_1 (*g*) DRN; W_4, D_1 (*h*) DRN; W_5, D_1

Figure 7: Training and validation curves of the standard (SN) and the domain randomized networks (DRN), with weight initialization varied W_i and a fixed data order seed D_1.



(*a*) SN; W_1, D_2 (*b*) SN; W_1, D_3 (*c*) SN; W_1, D_4 (*d*) SN; W_1, D_5

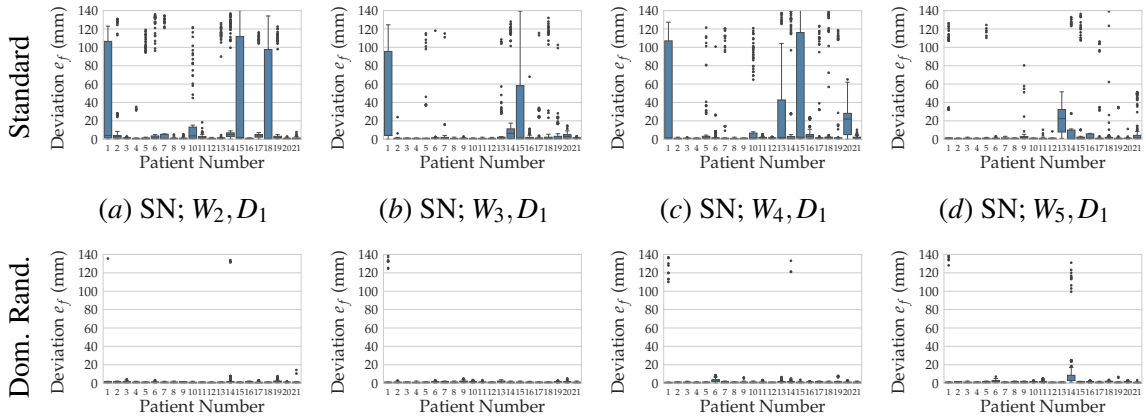
(*e*) DRN; W_1, D_2 (*f*) DRN; W_1, D_3 (*g*) DRN; W_1, D_4 (*h*) DRN; W_1, D_5

Figure 8: Training and validation curves of the standard (SN) and the domain randomized networks (DRN), with data order varied D_i and a fixed weight initialization seed W_1.

Appendix B. Quantitative Results on Clinical Data

Table 2: Robustness on clinical X-ray data of the standard (SN) and domain randomized networks (DRN) for varying weight initialization W_i and data order D_i for translation e_f and rotation ε_f.

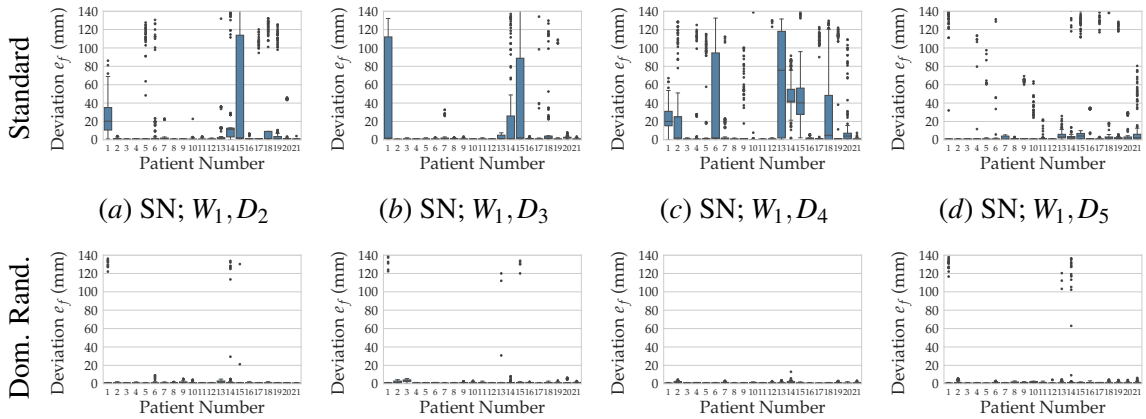
	SN/DRN								
	W_1, D_1	W_2, D_1	W_3, D_1	W_4, D_1	W_5, D_1	W_1, D_2	W_1, D_3	W_1, D_4	W_1, D_5
Median e_f (mm)	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
Mean e_f (mm)	5.77/1.38	14.07/1.01	7.30/1.15	10.78/1.34	5.05/1.71	9.42/1.43	6.65/1.41	18.45/0.75	7.60/2.26
75 % e_f (mm)	2.00/1.00	3.00/1.00	1.41/1.00	2.24/1.00	1.41/1.41	1.42/1.00	1.41/1.01	16.55/1.00	2.00/1.41
90 % e_f (mm)	6.32/1.41	97.64/1.41	8.06/1.42	27.00/1.41	5.83/2.24	19.72/1.41	3.16/2.00	85.79/1.41	7.00/2.00
Outliers e_f	237/80	263/60	172/28	263/59	154/53	274/54	98/69	342/30	338/114
Median ε_f (°)	0.46/0.39	0.55/0.50	0.67/0.53	0.60/0.46	0.50/0.49	0.51/0.49	0.52/0.46	0.71/0.46	0.54/0.49
Mean ε_f (°)	0.99/0.56	3.48/0.82	2.11/0.93	2.35/1.15	1.09/0.71	1.68/0.79	2.01/1.68	2.26/1.06	2.67/1.23
90 % ε_f (°)	1.69/1.19	7.88/1.55	3.48/1.70	3.57/2.08	1.97/1.47	3.38/1.61	3.37/2.29	5.69/1.48	3.39/1.70



(*a*) SN; W_2, D_1 (*b*) SN; W_3, D_1 (*c*) SN; W_4, D_1 (*d*) SN; W_5, D_1

(*e*) DRN; W_2, D_1 (*f*) DRN; W_3, D_1 (*g*) DRN; W_4, D_1 (*h*) DRN; W_5, D_1

Figure 9: Translational robustness results e_f on clinical patient data for weight initialization variation (W_i) with a fixed training data order seed (D_1). (a–d) Standard network (SN). (e–h) Domain randomized network (DRN).



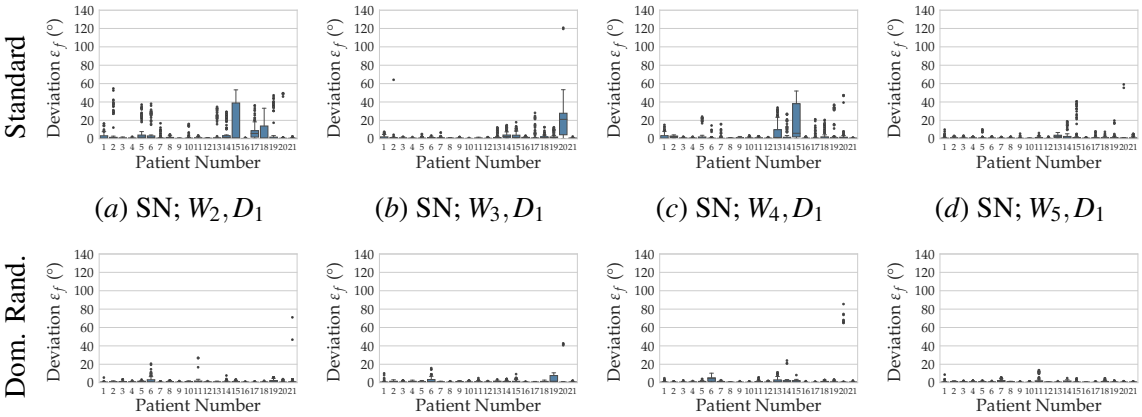
(*a*) SN; W_1, D_2 (*b*) SN; W_1, D_3 (*c*) SN; W_1, D_4 (*d*) SN; W_1, D_5

(*e*) DRN; W_1, D_2 (*f*) DRN; W_1, D_3 (*g*) DRN; W_1, D_4 (*h*) DRN; W_1, D_5

Figure 10: Translational robustness results e_f on clinical patient data for training data order variation (D_i) with a fixed weight initialization seed (W_1). (a–d) Standard network (SN). (e–h) Domain randomized network (DRN).

Table 3: Effects of individual domain randomizations on translational e_f and rotational ε_f robustness for a single setup (W_1, D_4) on clinical data.

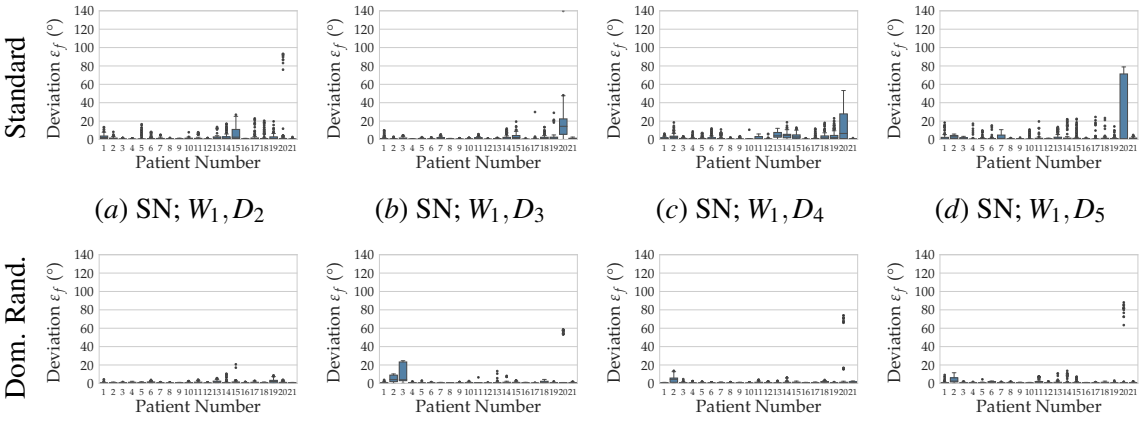
	Translation e_f							**Angle ε_f (°)**		
	Median (mm)	**Mean** (mm)	75 % (mm)	90 % (mm)	**Outliers**	< 5 mm (%)	< 3 mm (%)	**Median**	**Mean**	90 %
No DR	1.00	18.45	16.55	85.79	342	69.57	66.78	0.71	2.26	5.69
Collimation DR	1.00	4.03	2.23	7.81	169	84.76	81.09	0.48	0.95	1.78
Intensity DR	1.00	2.42	1.00	1.42	77	97.49	95.10	0.52	0.75	1.54
Coll. + Int. DR	1.00	0.75	1.00	1.41	30	99.94	98.96	0.46	1.06	1.48



(*a*) SN; W_2, D_1 (*b*) SN; W_3, D_1 (*c*) SN; W_4, D_1 (*d*) SN; W_5, D_1

(*e*) DRN; W_2, D_1 (*f*) DRN; W_3, D_1 (*g*) DRN; W_4, D_1 (*h*) DRN; W_5, D_1

Figure 11: Rotational robustness results ε_f on clinical patient data for weight initialization variation (W_i) with a fixed training data order seed (D_1). (a–d) Standard network (SN). (e–h) Domain randomized network (DRN).



(*a*) SN; W_1, D_2 (*b*) SN; W_1, D_3 (*c*) SN; W_1, D_4 (*d*) SN; W_1, D_5

(*e*) DRN; W_1, D_2 (*f*) DRN; W_1, D_3 (*g*) DRN; W_1, D_4 (*h*) DRN; W_1, D_5

Figure 12: Rotational robustness results ε_f on clinical patient data for training data order variation (D_i) with a fixed weight initialization seed (W_1). (a–d) Standard network (SN). (e–h) Domain randomized network (DRN).