Summer Internship Project Report

Duration: May. 2020 - July. 2020

Topic: CycleGAN for Interpretable Online EMT Compensation

Under the guidance of :

TECHNISCHE UNIVERSITÄT DARMSTADT

Dr. Anirban Mukhopadhyay
Head of Medical and Environmental Computing (MEC-Lab), Informatics,
TU Darmstadt, Germany

Submitted by :

Dhritimaan Das
Department of Electrical Engineering
Indian Institute of Technology(IIT-BHU), Varanasi
Roll No- 17085092
Part-4

Acknowledgments

I am grateful to Dr. Anirban Mukhopadhyay, Head of Medical and Environmental Computing (MEC-Lab), Informatics,TU Darmstadt, Germany who gave me this opportunity to pursue this project under his guidance during such adversity. I would also like to express my sincere gratitude to Henry John Krumb with whom I have been working in this project, who has been a guide to me at every step and obstacle that I encountered in this project. I am also thankful to my department(Department of Electrical Engineering,IIT-BHU), and my college(Indian Institute of Technology(IIT-BHU), Varanasi) who supported me to pursue this internship. Last but not the least I am particularly grateful to my family, friends and teachers without whose support pursuing this internship at such hard times would have been impossible for me.

Contents

1 Abstract

ElectroMagnetic Tracking (EMT) can partially replace X-ray guidance in minimally invasive procedures, reducing radiation in the OR. However, in this hybrid setting, EMT is disturbed by metallic distortion caused by the X-ray device. We plan to make hybrid navigation clinical reality to reduce radiation exposure for patients and surgeons, by compensating EMT error. Our online compensation strategy exploits cycle-consistent generative adversarial neural networks (CycleGAN). Points are translated from various bedside environments to their bench equivalents. Domain-translated points are fine-tuned to reduce error in the bench domain. We evaluate our compensation approach in a phantom experiment. Since the domain-translation approach maps distorted points to their lab equivalents, predictions are consistent among different C-arm environments. Error is successfully reduced in all evaluation environments. Our qualitative phantom experiment demonstrates that our approach generalizes well to an unseen C-arm environment. Adversarial, cycle-consistent training is an explicable, consistent and thus interpretable approach for online error compensation. Qualitative assessment of EMT error compensation gives a glimpse to the potential of our method for rotational error compensation.

2 Introduction

In minimally invasive surgery, EMT has the potential to partially replace continuous X-ray navigation [6], reducing the radiation exposure to both patients and surgeons. Such procedures are traditionally performed under X-ray only (for example laparoscopy [1], Endovascular Aneurysm Repair (EVAR) [2]). Our vision is to enable hybrid navigation in the clinical setting, where EMT can replace X-ray as the primary continuous tracker, and X-ray snapshots are only acquired intermittently. However, in current practice EMT is susceptible to metallic distortion caused by the C-arm, such that the surgeon can put little trust in EMT in between snapshots.

Traditional error-compensating algorithms to increase trust in EMT are offline in nature, resulting in tedious calibration and impractical clinical translation. We thus advocate learning-based online error compensation where a general purpose learning model is trained *only once*. Online compensation of EMT error can be realized by implementing data-driven models, which generalize among data from different environments. In previous works, we investigated learning-based online EMT compensation by employing a series of increasingly complex learning models (polynomial regression [7], Artificial Neural Networks (ANNs) [6]).

Despite giving increasingly better results, the interpretation of the failure modes for these complex models is a growing concern in learning literature [11]. Our major focus is to develop an *interpretable* error compensation technique, which imposes two more constraints beyond mere error reduction: *explicability* and *consistency*. Firstly, predictions need to be *explicable*; that is we need to know *why* a certain point from C-arm domain is mapped to a specific compensated point. Otherwise, compensation results could be arbitrary and still fulfill topological constraints, giving a false sense of reliability.

Our second constraint is *consistency* among distorted environments. Consistency is important for *online* error compensation, where training data stems from environments with varying distortion characteristics. If output is inconsistent among input environments, changing distortion characteristics of the electromagnetic field (e.g. by moving the C-arm in the course of a

surgical procedure) affect the coordinate frame of compensated points. Such changes jeopardize the X-ray-to-EMT registration, rendering our envisioned hybrid setting infeasible.

To impose these two constraints, we approach the problem as domain adaptation with cyclic-consistent Generative Adversarial Networks (GANs) [3, 12]. The main purpose of our approach is to learn a mapping from points of domain C, which is the domain of C-arm-distorted points from environments similar to bedside, to domain L, which is the domain of laboratory bench settings. Our approach is both explicable and consistent by design, and thus more interpretable than our ANN [6] approach. While domain adaptation using GANs is quite popular in the medical imaging setting [4], it has never been studied before in the context of EMT error compensation.

3 Materials

Our data-driven compensation approach uses measurements we acquired during our previous work [6], and were collected using an Ascension 3D Guidance trakSTAR system, a 180-type sensor and mid-range field generator. Custom C++ software is used for data collection with the trakSTAR system. We use a calibrated Lego board to collect positional data in three Degrees of Freedom (DoF), as proposed by us earlier [7]. For each measuring point, 500 samples are collected and averaged to reduce random noise. Displacement distances between points on the Lego board are then used as ground truth to calculate displacement error. This metric, based on relative displacements, eliminates the need for an additional measurement standard (e.g. a ruler). Positional error of displacements we use for training, validation and evaluation are listed in table. Similarly to other EMT systems, the trakSTAR provides a quality estimate with each measuring point, indicating the amount of metallic distortion. Each measuring point is constituted by $(x, y, z, q, \phi_x, \phi_y, \phi_z)$, where q denotes the system's quality estimate and ϕ_x, ϕ_y and ϕ_z denote rotation around x, y and z axes.

In our phantom study , we place the sensor at different positions inside an acrylic glass phantom, resembling a human aorta in life size. A 3D printed Lego adapter holds the sensor cable in place. Measurements for the phantom study and other bedside measurements are performed in vicinity of a Ziehm Vision RFD C-arm device.

Our learning models are implemented in Python using the PyTorch [10] framework.

	scenario	#points	displacement RMSE [mm]	displacement std. dev. [mm]
training	laboratory[1]	60	**0.367**	**0.202**
	c-arm 8 cm	60	1.292	1.264
	c-arm 11 cm	60	1.064	0.917
	c-arm[2] 50 cm	60	0.639	0.309
validation	c-arm 10 cm	60	1.101	0.989
	c-arm[3] 30 cm	60	0.743	0.372
evaluation	c-arm 7 cm	60	1.386	1.389
	c-arm 9 cm	60	1.192	1.139
	c-arm 12 cm	60	1.025	0.833

Table 1: Datasets collected in varying distances to c-arm and in a laboratory setup. Number of displacements, RMSE and standard deviation are noted for each dataset. Distances to c-arm are measured from x-ray source to base board center. [1]: only 2 of 3 z-layers used for training, c-arm[2]: gantry rotated at 60°, c-arm[3]: gantry rotated at 30°

4 Methods

To fulfill the explicability constraints we identified for interpretable online error compensation, we modify a domain adaptation approach that is originally used for image translation tasks. Instead of translating between two image domains (e.g. photorealistic vs. abstract), our goal is to translate EMT measurements from the bedside (high error) to the bench (low error) domain. Since the objective of this translation task is intuitive, we expect this compensation to be *explicable*.

In the following, we detail the modified domain adaptation architecture, the protocol and parameters we use for training. Subsequently, we describe how error in laboratory domain is compensated by a post-processing step, which uses a simple linear regression model.

4.1 Domain Adaptation by Adversarial Training

We employ cycle-consistent adversarial training for interpretable EMT compensation. Similar to the work of Zhu and Park et al. [12], we make use of two different GAN models, one for each direction of domain translation (C-arm to lab, lab to C-arm). As illustrated in Figure 1, the training process connects both GANs to achieve cyclic consistency. Since input data from laboratory and C-arm (Table 1) are unpaired and translation is thus under-constrained, adversarial training benefits from this additional cycle-consistency constraint.

Each of the two GANs consists of a generator network and a discriminator neural network. The generator receives an input point and generates a domain-adapted point, whereas the discriminator judges whether the generated point stems from the target domain. The generator's objective is to trick the discriminator by generating points close to its target domain, given an input point from the original domain. For instance, G_{CL} takes a point from C-arm environment and tries to generate a corresponding laboratory point, and D_{CL} judges whether this point actually is a *valid* laboratory

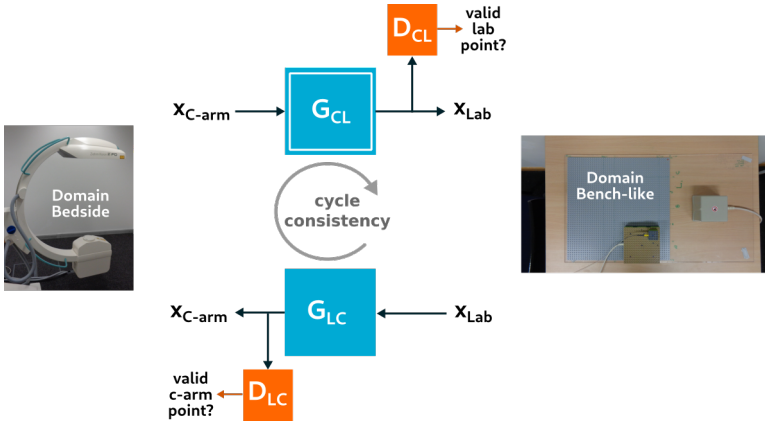

Figure 1: Cycle-consistent GAN architecture for unsupervised domain adaptation. G_{CL} translates points from C-arm to lab domain and, once trained, is used for compensation.

point. Ideally, both parties in this adversarial two-player game achieve the Nash equilibrium [3, 9].

Our two generator models (and discriminator models respectively) share an identical structure. Generators receive a $(x, y, z, q, \phi_x, \phi_y, \phi_z)$ point (normalized to $[0, 1]$) as input and produce a vector $(\hat{z}, \hat{q}, \hat{\phi_x}, \hat{\phi_y}, \hat{\phi_z})$, where $\hat{z}$, $\hat{q}$, $\hat{\phi_x}$, $\hat{\phi_y}$, $\hat{\phi_z}$ denote domain-translated values for z, quality and orientation.

To ensure that points from C-arm domain are not translated to *arbitrary* points in the lab domain, the generators do not alter the x and y positional components. We focus only on compensating the z component, since of all positional components, it is the most susceptible to error (for the trakSTAR system). Error in the x-y-plane is compensated in a fine tuning step further described in subsequent section.

4.2 Training Protocol

Our training protocol is similar to that of CycleGAN, but includes additional loss terms tailored to the problem of EMT error compensation. In particular, we compute the generator loss as weighted sum of individual penalties, which are described in the following:

Adversarial Loss L_{adv} is a binary-cross-entropy (BCE) loss term, which reflects how well each generator can fool its corresponding discriminator. It is computed as

$$L_{adv} = BCE\big(D_{CL}(G_{CL}(x_C)), l_{valid}\big) + BCE\big(D_{LC}(G_{LC}(x_L)), l_{valid}\big)$$

Where l_{valid} denotes the discriminator label we assign to valid points, that is the label the discriminator would assign to points that truly stem from the target domain.

Cycle Loss As described in the original CycleGAN paper [12], we enforce cycle-consistency by adding a loss term L_{cycle}:

$$\begin{aligned} L_{recov,L} &= |G_{CL}(G_{LC}(x_L)) - x_L| \\ L_{recov,C} &= |G_{LC}(G_{CL}(x_C)) - x_C| \\ L_{cycle} &= L_{recov,L} + L_{recov,C} \end{aligned}$$

$L_{recov,L}$ indicates how well an input point is recovered after translating it from domain L to domain C and back to L ($C \to L \to C$ in $L_{recov,C}$ respectively).

Compensation Loss In addition to the CycleGAN losses, we also penalize distance error only in the laboratory setting (we cannot enforce error to be low in generated C-arm samples) as a means of regularization. Compensation loss is computed as $L_{comp} = MSE(d_{G_{CL}}, d_{true})$ where $d_{G_{CL}}$ and d_{true} are distances between pairs of points, which stem from G_{CL} or ground truth data respectively.

The total generator loss is $L_{total} = \lambda_{adv} \cdot L_{adv} + \lambda_{cycle} \cdot L_{cycle} + q_{CL}^2$ with coefficients $\lambda_{adv} = 0.5$, $\lambda_{cycle} = 10$ and $\lambda_{comp} = 10^{-5}$, which we determine empirically. By penalizing high values of q_{CL} we enforce generated laboratory points with low distortion estimates and thus lower error.

4.3 Prediction Uncertainty

Although neural networks are known to generalize well to unseen data, predictions made under a lack of knowledge are uncertain. Fortunately, this uncertainty can be approximated by training the same architecture multiple times, but initialized with different random seeds (deep ensembling [8]). Computing the standard deviation among the resulting predictions yields an approximation of model-inherent (epistemic) uncertainty, and averaging the predictions is expected to yield better prediction accuracy, since we combine the knowledge of multiple models. We choose to sequentially train 10 different initializations in an ensemble as a compromise between training time

and accuracy.

4.4 Network and Training Parameters

Our discriminator models have three layers each, with 16 nodes per layer. All layers use LeakyReLU activations with a leak of 0.2, except for the last layer, which is Sigmoid-activated. The discriminators are trained with soft labels (uniform distribution of 0.0..0.2 and 0.8..1.0 respectively).

The generators also have four layers each, with 16 nodes per layer. Similar to the discriminators, the generator's layers use LeakyReLU activations, but with a leak of 0.01. The last layer uses linear activation.

Generators and discriminators are optimized under the use of Adam optimizer [5] both with a learning rate of 0.0005, which linearly decays to 0 after 100 epochs. Whereas the generators share a common optimizer, the discriminators are both trained by individual optimizers. During training, we use mini batches with a batch size of 16. The whole training for each model in the ensemble lasts 200 epochs. This training protocol is similar to that of the original CycleGAN implementation [12].

4.5 Fine-Tuning

As the cycle-consistent GAN model does not affect the x and y components of input points, there still exists positional error on the x-y-plane. Assuming that the points compensated by G_{CL} always lie in the laboratory domain, we can apply a compensation approach tailored to the laboratory domain. For the sake of simplicity, we choose to fit a linear regression model that compensates distance error similarly to our previous work [7], using input features (x, y, z, q).

5 Results

We first perform a bedside phantom evaluation of online error compensation.

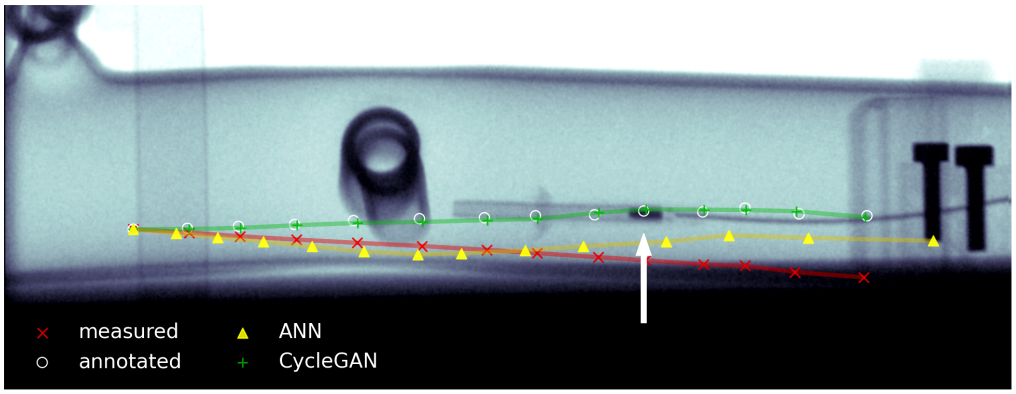

Figure 2: X-ray snapshot of aortic phantom with sensor inside. (Please find an animated video in supplementary material.) Arrow points to EMT sensor tip (dark rectangle). White circles are manually annotated sensor center points. Red, yellow and green overlays show uncompensated, ANN-compensated and domain-adapted EMT points respectively.

Afterwards, we compare our domain adaptation approach to the ANN we have previously proposed [6].

5.1 Bedside Evaluation on Aortic Phantom

To assess the quality of our online compensation method in a realistic hybrid bedside setting, we combine EMT and X-ray imaging in a pilot phantom study. Figure 2 shows the measuring setup used in this experiment, in which the EMT sensor is placed inside an aortic phantom which is positioned close to the C-arm. Since the C-arm gantry is rotated to 90°, this setup is different from anything the CycleGAN has seen during training. Using a custom 3D printed Lego fixture, we pull out the sensor in 13 steps of 8 mm.

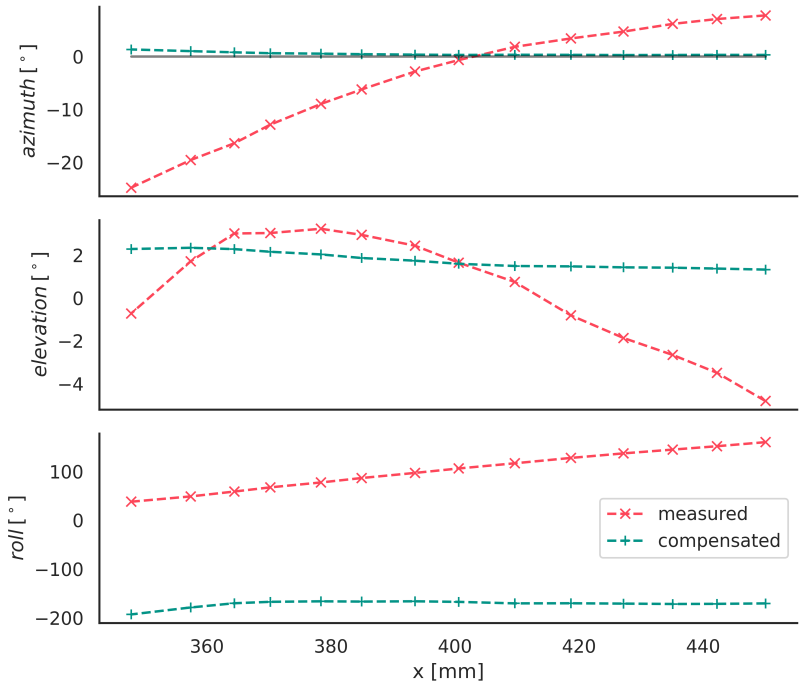

Figure 3: Measured and CycleGAN-compensated rotation angle (azimuth, elevation, roll) over x position, during sensor retraction in aorta phantom study.

For each individual step, an X-ray snapshot is created in the median plane, which corresponds to the x-z-plane of the EMT coordinate system.

The EMT sensor can be clearly distinguished from its background on the X-ray. We could therefore annotate the EMT sensor's center points in all 13 snapshots by hand. Points measured by EMT are scaled to pixel dimensions, and translated to match the annotated point (leftmost in Fig. 2). Rotation angle of the whole trajectory is estimated from ϕ_y at the starting annotation. We use the same transform for all three sets of points (uncompensated, ANN and CycleGAN) and omit the fine tuning step, to allow for better comparison between uncompensated and compensated trajectories. Fig. 2 shows that our compensated points are close to the annotations, indicating that our domain-adaptation model generalizes well to the unseen environment.

Since our compensations alters rotation angles ϕ_x, ϕ_y and ϕ_z as well, the phantom study in the C-arm environment also allows for a pilot qualitative assessment of rotational error compensation. Although we cannot directly measure the actual orientation of the sensor inside the phantom, we can make three assumptions:

1. The sensor tip is heading in positive direction of the tracker's x-axis throughout the whole experiment. We expect azimuth angle to be constant and close to zero.

2. Elevation is almost constant and near zero. Since the aortic phantom is slightly curved, elevation should decrease with higher x.

3. Roll angle is hard to determine absolutely (only relatively). However, we know that it does not change substantially during the experiment, as the cable is not twisted and is rigidly attached to the Lego block.

Figure 3 shows orientations over positions on the x-axis, which are measured in our phantom study. It illustrates that all three assumptions hold for compensated values: 1) Azimuth is constant and close to zero degrees (gray line), 2) elevation is almost constant at about 2° and 3) roll is nearly constant. Contrary to this, raw measurements violate all three assumptions.

5.2 Quantitative Comparison

In Figure 4, we see that the domain adaptation approach is more consistent among distorted environments and yields results close to laboratory points. Even C-arm points without a corresponding lab point in the training set are matched to their corresponding point in the laboratory domain, indicating that our method generalizes well.

On the other hand, tolopogy-based compensation by our previously proposed ANN [6] does not yield consistent output among input environments. Actually, ANN compensated points are still close to the input points on the z-axis and several millimeters away from corresponding lab points.

5.3 Abalation Experiment

Adversarial training with cycle-consistency is beneficial for compensation performance, compared to training a single GAN translating C-arm to laboratory points. Vanilla GAN without cycle loss worsens overall compensation performance, as we show in results table.

Our fine-tuning step brings an additional boost in accuracy to the proposed CycleGAN setup (Table 2). However, this step comes with the cost of introducing another source of predictive uncertainty.

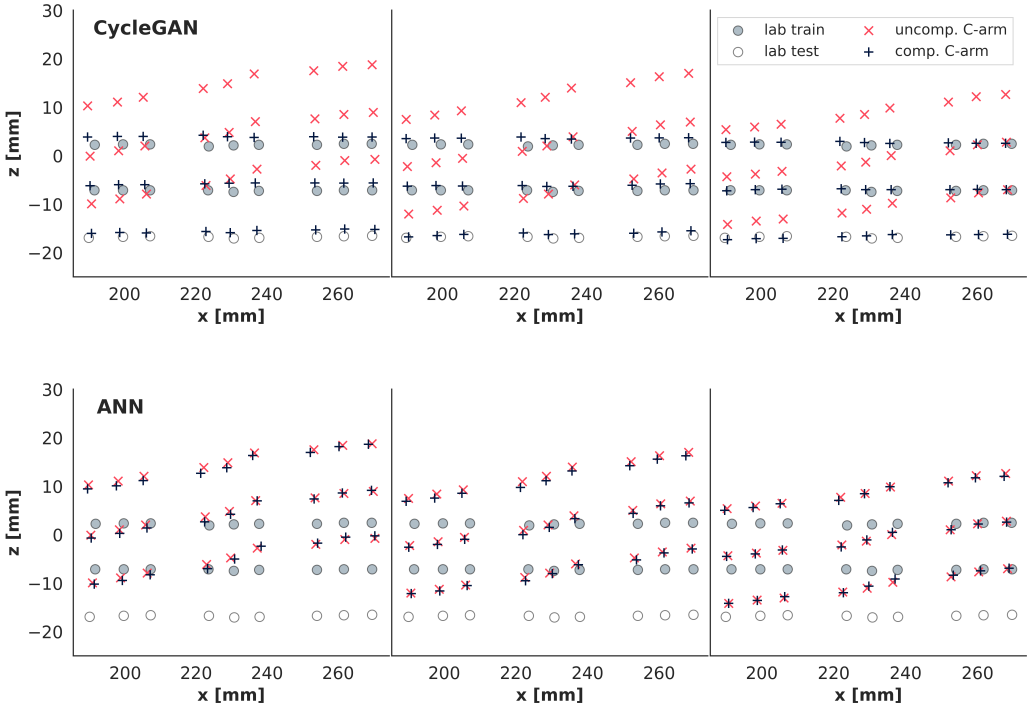

Figure 4: CycleGAN-adapted (top) and ANN-compensated (bottom) measuring points from C-arm at 7 cm (left), 9 cm (center), 12 cm (right), compared to corresponding lab points. White lab points are used for testing only.

Architecture	Dataset	RMSE [mm] ↓	σ_{error} [mm] ↓	σ_{pred} [mm] ↓
CycleGAN	7 cm	1.295	1.264	**0.370**
	9 cm	1.090	0.949	
	12 cm	1.007	0.722	
CycleGAN + Fine Tuning	7 cm	**1.100**	**1.117**	0.768[1]
	9 cm	**0.811**	**0.759**	
	12 cm	**0.622**	**0.530**	
GAN	7 cm	1.400	2.744	0.654
	9 cm	1.176	1.890	
	12 cm	1.030	1.256	

Table 2: Comparison of tracking error (RMSE & standard deviation) and prediction uncertainty σ_{pred} for different online architectures in evaluation setups from Table 1.
[1] Linear regression only: $\sigma_{pred} = 0.673\,mm$

6 Conclusion

Comparing our domain adaptation approach to previously proposed topology-based compensation [6], we see that online compensation performance is not only a matter of RMSE, but needs to be assessed with interpretability in mind. Predictions made by the topology-based method are hardly explicable. We can only hypothesize that the ANN tries to fulfill topological constraints to minimize distance error, without developing an understanding of what makes a plausible compensated point. Figure 5 illustrates how domain adaptation is explicable and consistent by design, whereas topological compensation is not.

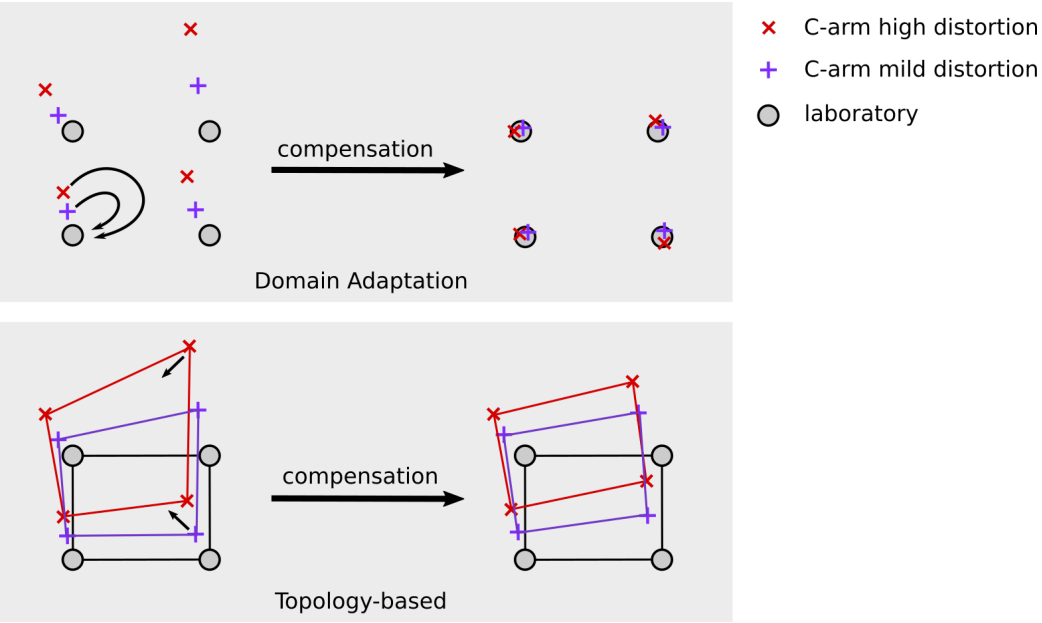

Figure 5: Schematic comparison of domain adaptation (top) versus topological (bottom) compensation schemes.

We train a modified CycleGAN on positions from various environments, achieving a translation of distorted (C-arm) to undistorted (laboratory) points.

The CycleGAN-based setup is capable of reducing error in different unseen C-arm scenarios, while still producing consistent results. Since we can show that points are always mapped to laboratory counterparts – regardless of these being directly present in the training set – our approach is also explicable. Hence, our domain adaptation approach is interpretable by design.

Bibliography

[1] Aoki, T., Mansour, D.A., Koizumi, T., Wada, Y., Enami, Y., Fujimori, A., Kusano, T., Matsuda, K., Nogaki, K., Tashiro, Y., et al.: Laparoscopic liver surgery guided by virtual real-time ct-guided volume navigation. Journal of Gastrointestinal Surgery pp. 1–8 (2020)

[2] Dijkstra, M.L., Eagleton, M.J., Greenberg, R.K., Mastracci, T., Hernandez, A.: Intraoperative c-arm cone-beam computed tomography in fenestrated/branched aortic endografting. Journal of vascular surgery **53**(3), 583–590 (2011)

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NEURIPS. pp. 2672–2680 (2014)

[4] Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A.: Gans for medical image analysis. Artificial Intelligence in Medicine **109**, 101938 (2020). https://doi.org/https://doi.org/10.1016/j.artmed.2020.101938, http://www.sciencedirect.com/science/article/pii/S0933365719311510

[5] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[6] Krumb, H., Hofmann, S., Kügler, D., Ghazy, A., Dorweiler, B., Bredemann, J., Schmitt, R., Sakas, G., Mukhopadhyay, A.: Leveraging spatial uncertainty for online error compensation in emt. IJCARS pp. 1–9 (2020)

[7] Kügler, D., Krumb, H., Bredemann, J., Stenin, I., Kristin, J., Klenzner, T., Schipper, J., Schmitt, R., Sakas, G., Mukhopadhyay, A.: High-precision evaluation of electromagnetic tracking. IJCARS **14**(7), 1127–1135 (2019)

[8] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in neural information processing systems. pp. 6402–6413 (2017)

[9] Nash, J.: Non-cooperative games. Annals of mathematics pp. 286–295 (1951)

[10] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (eds.) NEURIPS 32, pp. 8024–8035. Curran Associates, Inc. (2019)

[11] Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (xai): Towards medical xai. IEEE Trans. Neural Netw. Learn. Syst **PP** (2020)

[12] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)