ORIGINAL ARTICLE

Generation of annotated multimodal ground truth datasets for abdominal medical image registration

Dominik F. Bauer[1] · **Tom Russ**[1] · **Barbara I. Waldkirch**[1] · **Christian Tönnes**[1] · **William P. Segars**[2] · **Lothar R. Schad**[1] · **Frank G. Zöllner**[1] · **Alena-Kathrin Golla**[1]

Abstract

Purpose Sparsity of annotated data is a major limitation in medical image processing tasks such as registration. Registered multimodal image data are essential for the diagnosis of medical conditions and the success of interventional medical procedures. To overcome the shortage of data, we present a method that allows the generation of annotated multimodal 4D datasets.

Methods We use a CycleGAN network architecture to generate multimodal synthetic data from the 4D extended cardiac–torso (XCAT) phantom and real patient data. Organ masks are provided by the XCAT phantom; therefore, the generated dataset can serve as ground truth for image segmentation and registration. Realistic simulation of respiration and heartbeat is possible within the XCAT framework. To underline the usability as a registration ground truth, a proof of principle registration is performed.

Results Compared to real patient data, the synthetic data showed good agreement regarding the image voxel intensity distribution and the noise characteristics. The generated T1-weighted magnetic resonance imaging, computed tomography (CT), and cone beam CT images are inherently co-registered. Thus, the synthetic dataset allowed us to optimize registration parameters of a multimodal non-rigid registration, utilizing liver organ masks for evaluation.

Conclusion Our proposed framework provides not only annotated but also multimodal synthetic data which can serve as a ground truth for various tasks in medical imaging processing. We demonstrated the applicability of synthetic data for the development of multimodal medical image registration algorithms.

Keywords CycleGAN · Image registration · Image synthesis · Liver · Multimodal imaging

Introduction

Background

Multimodal imaging plays an important part in the diagnosis of cancers, such as liver cancer [19]. A variety of treatment options are available for hepatocellular carcinoma (HCC), the sixth most common malignancy worldwide and the third

leading cause of cancer-related deaths [10]. These include interventional procedures such as transarterial chemoembolizations (TACE) [21] or radioembolization [17]. The treatment planning benefits from using multimodal registration to combine pre- and intrainterventional data. Each imaging modality has strengths and weaknesses. Image registration enables the fusion of complementary information of each modality.

The lack of convenient ground truth data is a major limitation in the field of medical image segmentation and registration [26]. The generation of organ masks for segmentation requires labor-intensive manual annotation. For the development of image registration algorithms (especially for non-rigid image registration methods) and validation of registration accuracy, the ground truth is generally not available [26]. This is because the patient positioning in-between scans usually cannot be reproduced, particularly in the case

---

Frank G. Zöllner and Alena-Kathrin Golla share senior authorship.

✉ Dominik F. Bauer
dominik.bauer@medma.uni-heidelberg.de

[1] Computer Assisted Clinical Medicine, Mannheim Institute for Intelligent Systems in Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[2] Carl E. Ravin Advanced Imaging Labs, Department of Radiology, Duke University, Durham, USA

of multimodal imaging. In the abdomen, the variable content of the bladder and bowel and additional patient motion like respiration and heartbeat further exacerbate the problem.

Related work

To evaluate registration results or to train deep learning registration approaches, either anatomical multi-label organ masks or landmarks are required. However, generating these labels is labor-intensive, subjective or even impractical for large datasets. Pluim et al. introduced a semi-automatic framework that allows the creation of a large number of high-quality landmarks with minimal user effort [13]. Established ground truth datasets for the validation of image registration are usually only available for brain imaging. For instance, the simulated brain database (BrainWeb) provides simulated MRI imaging sequences (T1-weighted, T2-weighted, and proton density) [5]. The images are perfectly aligned, since they are calculated from the same model.

Image synthesis is able to reduce multimodal registration problems to monomodal problems by first converting one modality into the other. Modality reduction has shown improvements in registration accuracy for the brain [14] and the pelvis [3].

For MRI-only radiotherapy planning Wolterink et al. demonstrated feasible results using a CycleGAN approach for MRI-to-CT translation and showed that training with unpaired images is superior to training with paired images [24].

Using the digital 4D extended cardiac–torso (XCAT) phantom [16] instead of patient images for image synthesis is beneficial, because organ masks and motion displacement fields are provided by the phantom. Tmenova et al. presented a CycleGAN to synthesize X-ray angiograms from the XCAT phantom, which proved to be useful as a data augmentation strategy [20]. Russ et al. synthesized abdominal CT images using a CycleGAN and the XCAT phantom [15]. They showed that a vessel segmentation network trained on a combination of real CT and synthetic CT images achieved a superior performance compared to a network trained only on real data. Analytical models [12,23] and a GAN approach [1] to transform the CT XCAT phantom into cardiac or abdominal MRI images have been developed. To our knowledge, no multimodal registration ground truth dataset of the abdomen created from the same XCAT or digital phantom has been reported in the literature.

Contribution

Our approach to bypass the lack of ground truth data in image registration and segmentation is the generation of a mulit-modal synthetic dataset from the XCAT phantom. In this work, we focus on multimodal image registration. The syn-thesis is performed via CycleGAN networks, which were separately trained for each modality. To improve the preservation of high-contrast structures, we extend the CycleGAN generator loss with an intensity loss and a gradient difference loss.

Interventions are often monitored via cone beam CT (CBCT), whereas CT and MRI images are taken for diagnosis beforehand to assist the navigation during intervention [21]. Thus, our multimodal dataset consists of T1-weighted MRI, CT, and CBCT images. We use XCAT data in the inhaled and exhaled motion state to synthesize images. Since the same XCAT phantom is used as the starting point for all modalities, the resulting multimodal synthetic data is perfectly co-registered. Displacement fields for respiratory movements and segmentation masks for all organs are provided by the XCAT phantom. Therefore, it serves as a ground truth dataset for registration.

To demonstrate the utility of the multimodal dataset for the optimization of registration algorithms, we evaluate a multimodal non-rigid registration for varying parameter settings. We focus on the registration of the liver; however, the registration quality can be assessed for any other organ.

Materials and methods

Image synthesis framework

A schematic of our simulation framework is shown in Fig. 1. Starting from the CT XCAT phantom, CBCT and MRI XCAT versions are generated by applying a FOV Mask or by simulating the volume interpolated breathold exam (VIBE) signal equation [12], respectively. Organ masks for each modality are extracted from the phantoms. Images are synthesized via CycleGAN networks using the XCAT phantom as input. CycleGANs learn the mapping between two domains X and Y given unpaired training samples $x \in X$ and $y \in Y$ [25]. The mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$ are called generators. Two discriminators D_X and D_Y aim to distinguish between real images and generated images. Figure 2 shows the complete CycleGAN network architecture for the XCAT and CT image domain. Synthetic CT, CBCT, and MRI images are created via separately trained CycleGAN networks. The cycle consistency loss $L_{\text{cyc}}(G, F)$ enforces forward and backward consistency for the generators, i.e., $F(G(x)) \approx x$ and $G(F(y)) \approx y$. With a least squares generative adversarial loss $L_{\text{adv}}(G, F, D_X, D_Y)$, the generators are trained to generate images that cannot be distinguished from real images by the discriminator. The discriminators are 70×70 PatchGANs, which are trained with a least squares generative adversarial loss function. For the generators, we use a Res-Net architecture with an encoding stage, nine residual blocks and a decoding stage.
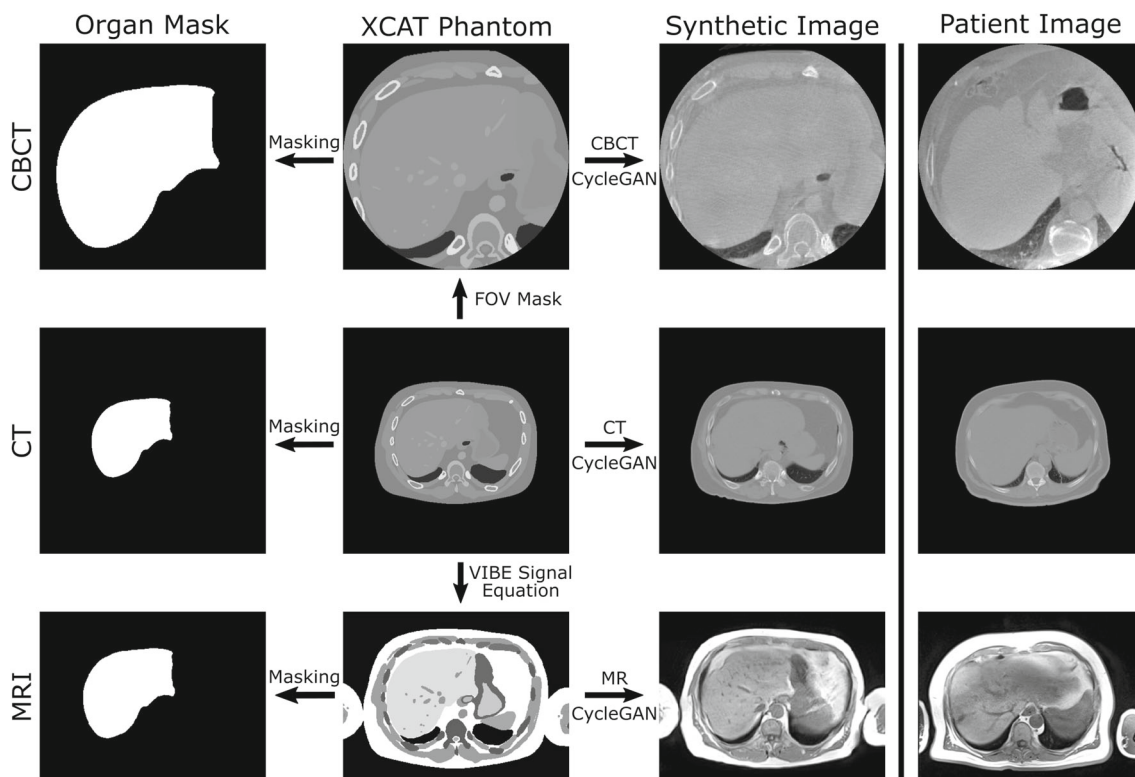
Fig. 1 Schematic of the simulation framework. Starting point is the CT XCAT, from which CBCT and MRI versions are derived. Synthetic CT, CBCT, and MRI images are created via separately trained CycleGAN networks. Organ masks can be obtained from the XCAT phantoms. Patient images that are used to train the CycleGANs are shown on the right-hand side

Fig. 2 CycleGAN network architecture: The generators $G_{XCAT \rightarrow CT}$ and $F_{CT \rightarrow XCAT}$ map images from the XCAT domain to the CT domain and vice versa. CycleGAN networks for MRI and CBCT images are trained analogously
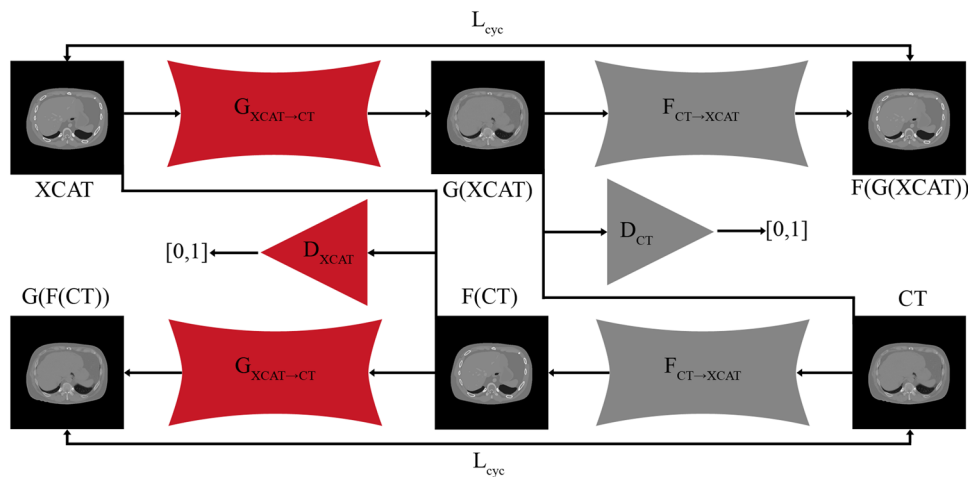
Training and loss functions

Training is performed using the Adam optimizer with a learning rate of 0.0002. The network is trained with 256×256 pixel image patches and a batch size of 4. For each axial slice, one random patch is extracted. We train the networks for 150.000 steps.

To enhance the preservation of high-contrast structures, we extend the generator loss with an intensity loss and a gradient difference loss:

$$L_{int}(G, F) = ||(G(x) - x)||_1 + ||(F(y) - y)||_1, \quad (1)$$

$$L_{gdl}(G, x) = \sum_{i,j} ||x_{i,j} - x_{i-1,j}| - |G(x)_{i,j} - G(x)_{i-1,j}||^2$$
$$+ ||x_{i,j} - x_{i,j-1}| - |G(x)_{i,j} - G(x)_{i,j-1}||^2. \quad (2)$$

The intensity loss preserves the signal intensity of the organs provided by the XCAT phantom. As shown by Nie et al., the gradient difference loss prevents blurring and therefore sharpens the synthesized images [11]. The total generator loss is a combination of the previously defined losses with different weights:

$$
\begin{aligned}
L_{\text{gen}}(G, F, D_{\text{X}}, D_{\text{Y}}) \\
= L_{\text{adv}}(G, F, D_{\text{X}}, D_{\text{Y}}) + \lambda_{\text{cyc}} \, L_{\text{cyc}}(G, F) \\
+ \lambda_{\text{int}} \, L_{\text{int}}(G, F) + \lambda_{\text{gdl}} \, (L_{\text{gdl}}(G, x) + L_{\text{gdl}}(F, y)).
\end{aligned}
\tag{3}
$$

We train three CycleGAN networks for CBCT, CT and MRI with empirically chosen combinations of weights $\lambda_{\text{cyc}}/\lambda_{\text{int}}/\lambda_{\text{gdl}}$, of 10/10/5, 10/10/5 and 10/0.4/0.4, respectively. As shown in our previous study, a combination of the gradient loss and the intensity loss yields the best results [2]. Further increasing the weighting factors leads to excessive regularization and thus the networks learns an identity mapping. For the MRI networks, lower over-regularization thresholds are found.

Data

We train our CycleGAN network to map between XCAT phantom data and real patient data. The goal is to obtain networks that generate realistic looking synthetic data using the XCAT phantoms as input. In the following paragraphs, we will address the real patient and the XCAT training data separately.

Patient data

The patient training data is retrospectively extracted from our Picture Archiving and Communication System (PACS). CT and $T1$-weighted MRI scans are acquired as part of routine clinical practice before TACE patients undergo a CBCT-guided TACE intervention. All scans are acquired on whole body clinical devices (Siemens Healthineers, Forchheim, Germany; CT: Somatom Emotion 16; CBCT: Artis Zeego; MRI: Magnetom Tim Trio). The MRI images are acquired at 3 Tesla with the VIBE sequence. For each modality the patient images are resampled to a unified voxel spacing given in Table 1. All scans include the whole liver and the narrow field of view of the CBCT scan is focused on the liver. The MRI images include arms, the CT and CBCT images do not. As the XCAT phantom does not include the patient couch, we removed the patient couch from the CT patient volumes. For CBCT and MRI, no couch is visible in the patient images.

The image intensities are windowed to the ranges given in Table 1. For CT and CBCT, a fixed window was used. Since

MRI intensities vary widely from image to image, the 10th and 90th percentile of each volume (whole 3D matrix) was used for windowing. For training, a linear intensity transformation is applied to transform the intensities from the windowing interval to $[-1,1]$. Normalization of training data is a crucial step in improving training performance, regardless of the normalization method used [8].

XCAT phantom data

The XCAT model provides highly detailed whole-body anatomies. Organ masks can be easily obtained within the XCAT framework. Since CycleGANs maintain the geometry provided by the XCAT, the organ masks can be used as segmentation masks in the synthesized images. The phantom includes female and male models for varying ages. The heart beat and respiratory motions can be simulated and displacement fields of these motions can be generated. The anatomy and motion can be adapted by various parameters. This allows the creation of highly individual patient geometries. For the XCAT training data we generate one XCAT volume per XCAT model for each modality with 56 different models of varying ages. The XCATs include the whole liver and are generated with the same voxel spacing, windowing and normalization as the resampled patient data. Arms are included only in the MRI XCATs.

The XCAT phantom provides attenuation coefficients for all organs. We vary the simulated tube energy of the CBCT and CT phantoms from 90–120 keV in steps of 5 keV. This leads to a variation of attenuation coefficients in the phantoms. Afterward, those are transformed into Hounsfield Units. To obtain CBCT and MRI XCAT data, we need to convert the CT XCAT. For the CBCT XCAT, we apply a field of view mask obtained from the patient CBCTs, which is centered on the liver. For the MRI phantoms, we replace the attenuation coefficients for each organ with simulated MRI values using the signal equation for the VIBE sequence. It ensures that the MRI signal is initialized with realistic values matching the MRI training data. This enables us to use the aforementioned intensity and gradient loss for the generation of synthetic MRI images, since the transformation with the CycleGAN is now monomodal. The signal intensities (SI) for the VIBE sequence in terms of acquisition parameters repetition time TR, echo time TE, and flip angle α and tissue-specific $T1$, $T2$ relaxation times, and proton density ρ is given by:

$$
\text{SI} = \frac{\rho \sin \alpha (1 - \exp{-\frac{\text{TR}}{T1}})}{(1 - \cos \alpha \exp{\frac{-\text{TR}}{T1}})} \exp{\frac{-\text{TE}}{T2}}.
\tag{4}
$$

We calculate the MRI intensity for all 44 abdominal organs present in the XCAT. The imaging parameters TE $= 4.54$ ms, TR $= 7.25$ ms, and $\alpha = 10°$ are obtained from the patient

Table 1 Training data statistics

	Parameter	Resolution $(x/y/z)$ [mm]	Windowing	Volumes	Volume size	Arms	Age
CT patient	100–130kVp	1/1/2	[−1024, 1500]HU	22	$512 \times 512 \times [52, 151]$	No	66 ± 9
CT phantom	90–120kVp	1/1/2	[−1024, 1500]HU	56	$512 \times 512 \times [80, 124]$	No	51 ± 14
CBCT patient	93–124.7kVp	0.486/0.486/0.486	[−1024, 2000]HU	24	$512 \times 512 \times 386$	No	67 ± 10
CBCT phantom	90–120kVp	0.486/0.486/0.486	[−1024, 2000]HU	56	$512 \times 512 \times 386$	No	51 ± 14
MRI patient	3 T	1/1/3	[10th, 90th] percentile	24	$330 \times 450 \times [48, 93]$	Yes	67 ± 10
MRI phantom	3 T	1/1/3	[10th, 90th] percentile	56	$330 \times 450 \times [59, 88]$	Yes	51 ± 14

For CT and MRI, the number of slices per image vary in the given interval

VIBE scans. The values for the proton density ρ are taken from [23]. T1 and T2 relaxation times for 3 T for blood and the spinal cord are obtained from [18] and the rest from [6]. For organs with no available $T1$, $T2$ or ρ, we use values of similar organs. To simulate some organ variability, we randomly vary $T1$, $T2$, and ρ by $\pm 5\%$ using a uniform distribution.

Evaluation metrics

Quantification of the synthetic image quality is difficult, since there are no corresponding real images for comparison [20]. Therefore, metrics that require a one-to-one correspondence like the mean absolute error (MAE) cannot be calculated between synthetic and real images. Instead, we calculate one-to-one corresponding metrics between the synthetic images and the XCATs, to investigate the magnitude of change from the XCAT phantoms. Real patient images and synthetic images are then compared by assessing their noise characteristics and voxel intensity distributions.

Synthetic versus XCAT

The axial slices of the synthetic CT volumes are compared to the corresponding axial slices of the XCAT volumes with respect to anatomical accuracy. The MAE is calculated to assess the change of the intensity values. We exclude the background for the calculation of the MAE. The similarity of structure and features is evaluated using structural similarity index measure (SSIM) and feature similarity index measure (FSIM) [9,22]. Additionally, we calculate the edge preservation ratio (EPR) and edge generation ratio (EGR) [4,15].

Synthetic versus real patient

Regarding realistic noise characteristics and intensity distribution, the 3D synthetic volumes are compared to the 3D patient volumes. For the noise characteristics, only liver voxels are considered. Limiting the noise considerations to the liver is reasonable, since the liver is a large and mostly homogeneous organ. We manually segmented the liver in four patients for each modality. The liver segmentations for the 56 synthetic images are provided by the XCAT phantom. The noise texture is evaluated using an estimation of the radial noise power spectrum (NPS). The radial NPS of the synthetic and patient images is compared by calculating the Pearson correlation coefficient, further called the NPS correlation coefficient (NCC) [15]. In addition to noise texture, we calculate the noise magnitude (NM), i.e., the standard deviation of the liver voxel intensities.

Furthermore, intensity distribution histograms of patient and synthetic images are calculated. To quantify their sim-

ilarity, the Pearson correlation coefficient between them is calculated (HistCC).

Proof of principle registration evaluation

We perform a proof of principle image registration to demonstrate the feasibility of the multimodal dataset for evaluation and thus development of registration algorithms. Our goal is to investigate different parameter settings to optimize the registration result. We implement the registration in Python 3.5 with SimpleITK 1.2.4. A non-rigid B-spline transform with a gradient descent optimizer, a learning rate of 1 and a maximum of 300 iterations is used. Three different registration metrics are considered, namely Mattes Mutual Information (MMI), Normalized Correlation (NC), and Mean Squares (MS). For the MMI, 50 histogram bins are used. The MS metric is only used for the monomodal CT to CT registration, since it is not suited for multimodal images. Additionally, we vary the spacing of the B-spline control points from 50 mm to 150 mm in steps of 20 mm. For the multimodal and monomodal registrations this results in 12 and 18 different parameter settings, respectively.

The registration is performed on the synthetic data from all 56 XCAT models. We registered the CT, MRI and CBCT images in the inhaled state to the CT image in the exhaled state. To evaluate the registration, we take advantage of the liver organ masks obtained from the XCAT phantoms. The veins and arteries inside the liver are included in the liver mask by using a morphological closing operation. We apply the registration transform to the liver masks in the inhaled state and compare the result to the CT liver mask in the exhaled state. The overlap of the two masks is assessed by calculating the Dice similarity coefficient (DSC).

Results

Synthetic images

First, we consider the metrics that compare the synthetic images with the XCAT phantoms shown in the upper half of Table 2. The FSIM and SSIM indicate that image structures and features were well-preserved in the CT and CBCT images, whereas the synthetic MRIs showed little structural and feature similarity to the XCATs. Regarding edges, the EPR is similar for all modalities, whereas the EGR is largest for the CBCT images. The MAE is slightly larger than the NM (synthetic) for every modality. The MAE for CBCT is more than twice as high as the MAE for CT.

Secondly, we compare the synthetic images to the patient images. The two right columns of Fig. 1 show axial synthetic and patient slices of each modality. Qualitatively, the style of the synthesized images is in good agreement with the real

patient images. To quantify this observation, we compared the noise characteristics and voxel intensity distribution of the synthetic images to the patient images, and the results are listed in the lower half of Table 2. A high NCC for all modalities indicates that the noise texture was emulated realistically, albeit the NCC is slightly smaller for the synthetic MRI images. For all modalities, the NM (synthetic) is in excellent agreement with the NM (patient). In Fig. 3, the intensity histograms are shown. In general, the synthetic intensity distributions match the patient intensity distributions nicely. This is underlined by the overall high HistCC values in Table 2. However, for CT and CBCT, the soft tissue peaks are modeled a bit too narrowly. The lung tissue peak is shifted toward higher CT numbers for the CT. In the MRI, the soft tissues is slightly underrepresented.

Proof of principle registration

The DSC for the evaluation of the proof of principle registration is shown in Fig. 4. The monomodal CT to CT registration yielded good results for all registration metrics and grid point spacings, with the best result for MMI with 50 mm grid point spacing. For CBCT, the MMI again worked well, whereas the registrations using the NC mostly failed. The best results were again obtained with MMI and a grid point spacing of 50 mm. For MRI, the registrations with MMI and NC yielded similar results with the best result obtained for NC with a grid point spacing of 150 mm. Overall, the monomodal CT to CT registration achieved the best results.

Coronal views of the registration results for the best settings of each modality are visualized in Fig. 5. The registered images in the middle row show a large similarity to the ground truth. This observation is further supported by the overlaid liver contours. The post-registration liver contour (yellow) is in high agreement with the ground truth liver contour (red).

Discussion

The image quality metrics in Table 2 demonstrate that our framework provides reaslistic multimodal image data. Low SSIM and FSIM for the MRI images indicate that image values of the homogeneous organs in the MRI XCAT phantoms needed to be altered more strongly by the networks in comparison with the CT XCATs. The lower SSIM, FSIM, and EPR for MRI are likely a result of lower weighting of the gradient loss and intensity loss.

The ratio of MAE to NM (synthetic) is 2.1, 1.3, and 1.5 for CBCT, CT, and MR, respectively. Assuming normally distributed noise, the ratio of MAE to NM is only about 0.8 [7]. This means that the MAE cannot be attributed to noise alone. The large MAE for the CBCT images compared to the CT images could be due to the introduction of metal

Table 2 Image quality metrics for the evaluation of the synthetic images

	CBCT	CT	MRI
SSIM	0.85 ± 0.05	0.94 ± 0.02	0.59 ± 0.04
FSIM	0.82 ± 0.03	0.82 ± 0.02	0.51 ± 0.02
EPR	0.47 ± 0.06	0.43 ± 0.04	0.40 ± 0.03
EGR	3.0 ± 0.7	1.9 ± 0.3	1.7 ± 0.2
MAE	109 ± 14	51 ± 16	37 ± 6
NCC	0.997 ± 0.001	0.980 ± 0.010	0.86 ± 0.04
NM (Synthetic)	52 ± 13	39 ± 5	25 ± 3
NM (Patient)	60 ± 16	39 ± 19	22 ± 5
HistCC	0.994 ± 0.003	0.999 ± 0.002	0.94 ± 0.03
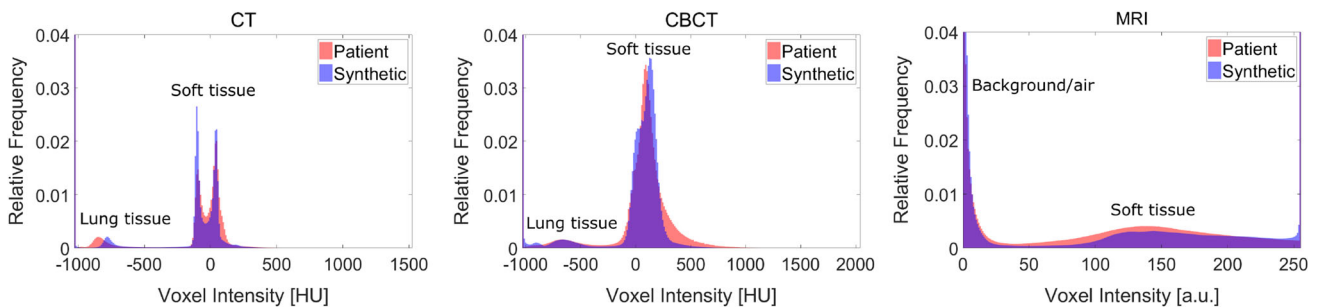
Fig. 3 Intensity histograms of the patient and synthetic images averaged over all volumes. Note that the background peaks are cropped
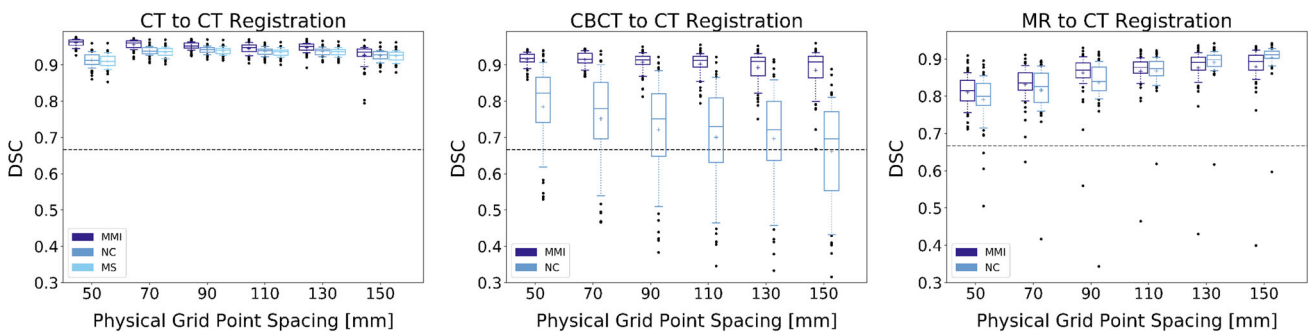
Fig. 4 DSC for the proof of principle registrations with 56 data points each. The mean is marked as a "+" and the whiskers indicate the 10th and 90th percentile. The dashed horizontal line shows the mean pre-registration DSC

Fig. 5 Pre- and Post-registered images and the corresponding ground truth. Red contours indicate the ground-truth boundaries of the liver (target). Blue and yellow contours represent the boundaries of the liver before and after deformation, respectively
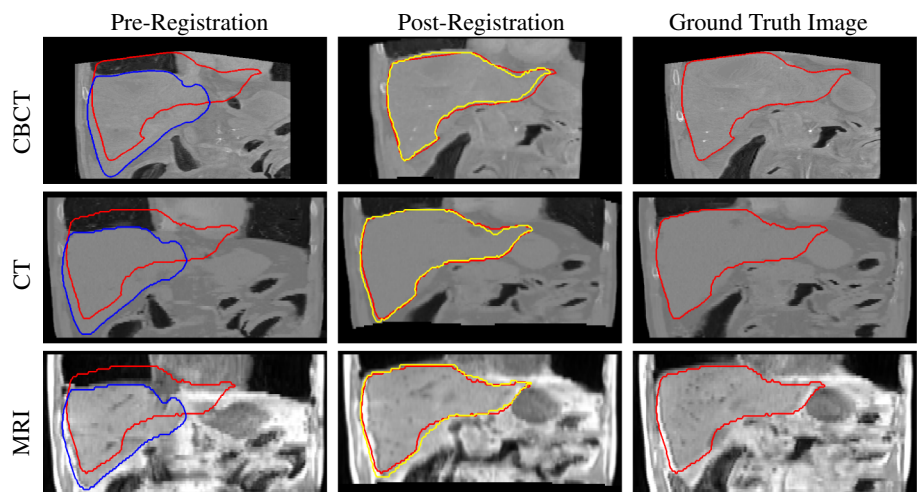
artifacts, as the patient CBCTs showed metal artifacts in the liver caused by medical instruments. This is supported by a large EGR for CBCT.

For all modalities realistic noise texture and magnitude was achieved. Additionally, the voxel intensity distribution was modeled adequately. Most of the discrepancies between the patient and synthetic histograms in Fig. 3 can be explained by inspecting the XCAT phantoms. The deviation of the CT lung peaks (synthetic −780 HU, patient −835 HU) can be explained by an overestimated initial lung value of −760 HU given by the XCAT. The narrow soft tissue peaks for CT and CBCT could be due to insufficient variation in organ attenuation coefficients. The under representation of soft tissue in the synthetic MRI is due to the body size of the patients and XCATs. We found that in the MRI patient dataset 66.5 % of the image voxels show the body, whereas for the MRI XCAT dataset, it is only 46.5 %. A rather large HistCC of 0.94 ± 0.03 was still achieved, since this under representation has only a minor effect on the correlation between the histograms. We prepared the XCAT data such that it matches the patient dataset as good as possible, see Table 1. In the future, we will consider the patient body size beforehand and adjust the XCAT body size accordingly.

An important requirement for using synthetic data to evaluate or train registration algorithms is a realistic respiratory deformation model. The XCAT framework allows the respiratory rate, the amount of diaphragmatic motion, the amount of chest expansion, and the amount of cardiac motion due to respiration to be varied. In addition, curves controlling the diaphragm motion and chest expansion can be individualized. However, as the breathing patterns of patients are complex and diverse, the XCAT model may still lack generalizability.

The results of the proof of principle registration demonstrate that the synthetic dataset can be used to evaluate different registration algorithms. We were able to evaluate the performance of different registration algorithms and to fine-tune parameter settings. It is noticeable that the registration for CT and CBCT works better for small B-spline grid point spacings, while it is the opposite for MRI. Further studies are needed to assess whether this is due to the smaller image size of the MRI data and thus the total number of B-spline control points or to other imaging characteristics of MRI and CT. Computation time can be measured and taken into consideration. For example, registrations with smaller grid point spacings take much longer. Thus, choosing MMI with 150 mm for CT and CBCT registrations might be a reasonable trade-off, as the registration quality is only slightly lower, while the registration time is substantially reduced. The availability of organ masks enabled a rather simple registration evaluation.

Conclusion

The presented simulation framework can be used to extend small datasets by transferring the style of the dataset onto the geometry given by the XCAT phantom. The obtained datasets can serve as a ground truth for image registration. A multimodal dataset consisting of T1-weighted MRI, CT and CBCT images was created and used to demonstrate the refinement and evaluation of multimodal image registration algorithms. In the future, the framework will be extended to other modalities, such as T2-weighted MRI or PET, which can further boost the performance of multimodal methods. An extension to other body regions, such as the thorax or pelvis, is also possible. Synthetic images over larger body regions are especially interesting for whole body segmentation. Expansion of datasets using this method provides a promising tool to overcome the dearth of medical training data.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval Ethical approval for this study was obtained from the ethics committee II of the Medical Faculty Mannheim, Heidelberg University (2016-863R-MA).

Human and animal rights We are in compliance with ethical standards.

References

1. Abbasi-Sureshjani, S., Amirrajab, S., Lorenz, C., Weese, J., Pluim, J., Breeuwer, M.: 4d semantic cardiac magnetic resonance image synthesis on xcat anatomical model. arXiv preprint arXiv:2002.07089 (2020)

2. Bauer, D.F., Schnurr, A.K., Russ, T., Goerttler, S., Schad, L.R., Zoellner, F.G., Chung, K.: Synthesis of ct images using cyclegans: Enhancement of anatomical accuracy. In: Proc. Int. Conf. Med. Imag. Deep Learning (2019)

3. Cao X, Yang J, Gao Y, Guo Y, Wu G, Shen D (2017) Dual-core steered non-rigid registration for multi-modal images via bi-directional image synthesis. Med. Image Anal. 41:18–31

4. Chen, L., Jiang, F., Zhang, H., Wu, S., Yu, S., Xie, Y.: Edge preservation ratio for image sharpness assessment. In: 2016 12th World Congress on Intelligent Control and Automation (WCICA), pp. 1377–1381. IEEE (2016)

5. Cocosco CA, Kollokian V, Kwan RKS, Pike GB, Evans AC (1997) Brainweb: Online interface to a 3d mri simulated brain database. NeuroImage 5:425

6. De Bazelaire, C.M., Duhamel, G.D., Rofsky, N.M., Alsop, D.C.: Mr imaging relaxation times of abdominal and pelvic tissues measured in vivo at 3.0 t: preliminary results. Radiology **230**(3), 652–659 (2004)

7. Geary RC (1935) The ratio of the mean deviation to the standard deviation as a test of normality. Biometrika 27(3/4):310–332

8. Jacobsen N, Deistung A, Timmann D, Goericke SL, Reichenbach JR, Güllmar D (2019) Analysis of intensity normalization for optimal segmentation performance of a fully convolutional neural network. Z. Med. Phys. 29(2):128–138

9. Lin Zhang, Lei Zhang, Xuanqin Mou, Zhang, D.: FSIM: A Feature Similarity Index for Image Quality Assessment. IEEE Trans. Image Processing **20**(8), 2378–2386 (2011)

10. Memon K, Lewandowski RJ, Kulik L, Riaz A, Mulcahy MF, Salem R (2011) Radioembolization for primary and metastatic liver cancer. Semin. Radiat. Oncol. 21(4):294–302

11. Nie D, Trullo R, Lian J, Wang L, Petitjean C, Ruan S, Wang Q, Shen D (2018) Medical image synthesis with deep convolutional adversarial networks. IEEE Trans. Biomed. Eng. 65(12):2720–2730

12. Paganelli C, Summers P, Gianoli C, Bellomi M, Baroni G, Riboldi M (2017) A tool for validating mri-guided strategies: a digital breathing ct/mri phantom of the abdominal site. MBEC 55(11):2001–2014

13. Pluim, J.P., Muenzing, S.E., Eppenhof, K.A., Murphy, K.: The truth is hard to make: Validation of medical image registration. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 2294–2300. IEEE (2016)

14. Roy, S., Carass, A., Jog, A., Prince, J.L., Lee, J.: Mr to ct registration of brains using image synthesis. In: Proc. SPIE Med. Imag., vol. 9034, p. 903419. SPIE (2014)

15. Russ T, Goerttler S, Schnurr AK, Bauer DF, Hatamikia S, Schad LR, Zöllner FG, Chung K (2019) Synthesis of ct images from digital body phantoms using cyclegan. Int. J. Comput. Assist. Radiol. Surg. 14(10):1741–1750

16. Segars WP, Sturgeon G, Mendonca S, Grimes J, Tsui BMW (2010) 4D XCAT phantom for multimodality imaging research. Med. Phys. 37(9):4902–4915

17. Spahr N, Thoduka S, Abolmaali N, Kikinis R, Schenk A (2019) Multimodal image registration for liver radioembolization planning and patient assessment. Int. J. Comput. Assist. Radiol. Surg. 14(2):215–225

18. Stanisz GJ, Odrobina EE, Pun J, Escaravage M, Graham SJ, Bronskill MJ, Henkelman RM (2005) T1, t2 relaxation and magnetization transfer in tissue at 3t. Magn. Reson. Med. 54(3):507–512

19. Tempany CM, Jayender J, Kapur T, Bueno R, Golby A, Agar N, Jolesz FA (2015) Multimodal imaging for improved diagnosis and treatment of cancers. Cancer 121(6):817–827

20. Tmenova O, Martin R, Duong L (2019) Cyclegan for style transfer in x-ray angiography. Int. J. Comput. Assist. Radiol. Surg. 14(10):1785–1794

21. Waldkirch, B., Engelhardt, S., Zöllner, F.G., Schad, L.R., Wolf, I.: Multimodal image registration of pre-and intra-interventional data for surgical planning of transarterial chemoembolisation. In: Proc. SPIE Med. Imag., vol. 10951, p. 109512U. SPIE (2019)

22. Wang Z, Bovik AC, Sheikh HR (2004) Image quality assessment: From error measurement to structural similarity. IEEE Trans. Image Processing 13(4):600–612

23. Wissmann L, Santelli C, Segars WP, Kozerke S (2014) Mrxcat: Realistic numerical phantoms for cardiovascular magnetic resonance. J. Cardiovasc. Magn. Reson. 16(1):63

24. Wolterink, J.M., Dinkla, A.M., Savenije, M.H.F., Seevinck, P.R., van den Berg, C.A.T., Išgum, I.: Deep mr to ct synthesis using unpaired data. In: Proc. Int. Workshop Simul. Synth. Med. Imag., pp. 14–23 (2017)

25. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. IEEE Int. Conf. Comput. Vis., pp. 2223–2232 (2017)

26. Zöllner FG, Šerifović-Trbalić A, Kabelitz G, Kociński M, Materka A, Rogelj P (2020) Image registration in dynamic renal mri-current status and prospects. Magn. Reson. Mater. Phy. 33:33–48