Optimizing CycleGAN Design for CBCT-to-CT Translation

Optimizing CycleGAN Design for CBCT-to-CT Translation: Insights into 2D vs 3D Modeling, Patch Size, and the Need for Tailored Evaluation Metrics

Ibrahim Hadzic[*,a,1], Suraj Pai[a,1], Vicki Trier Taasti[a], Dennis Bontempi[a], Ivan Zhovannik[a,b], Richard Canters[a], Jan Jakob Sonke[b], Andre Dekker[a], Jonas Teuwen[b,2], Alberto Traverso[a,2]
[a]Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Center+, Maastricht, the Netherlands; [b]Department of Radiation Oncology, Netherlands Cancer Institute, Amsterdam, the Netherlands

ABSTRACT

This paper explores the impact of different design choices on the performance of cycle-consistent Generative Adversarial Networks (GANs) in the context of CBCT-to-CT image translation. We examined the implications of choosing between 2D and 3D models, the size of 3D patches, and the integration of the Structural Similarity Index Measure (SSIM) into the cycle-consistency loss. Additionally, we introduced a partially-invertible VNet architecture into the RevGAN framework, enabling the use of 3D UNet-like architectures with minimal memory footprint.
Our evaluation methods included standard image similarity metrics (MAE, PSNR, SSIM) and visual inspection of both patient scans (in-distribution data) and phantom scans (out-of-distribution data). Our findings suggest that 3D models, despite requiring a longer training time to converge due to the number of parameters, produce fewer image perturbations compared to 2D models. Training with larger patches also improved stability and significantly reduced artifacts, but increased the training time, while the SSIM-L1 cycle-consistency loss function enhanced performance.
Interestingly, our study revealed a discrepancy between standard image similarity metrics and visual evaluation, with the former failing to adequately penalize visually evident artifacts. This underscores the need for tailored and standardized evaluation metrics for medical image translation, which would facilitate more accurate comparisons across studies. To further the clinical applicability of image-to-image translation, we have open-sourced our methods and experiments, available at github.com/ganslate-team.

Keywords: CycleGAN, CBCT, Synthetic CT, Image-to-image translation, Adaptive radiotherapy, 2D vs 3D, Image similarity metrics

1. DESCRIPTION OF PURPOSE

In image-guided adaptive radiotherapy, medical imaging enhances the precision of radiation treatment by refining dose delivery [1]. CBCT machines, commonly used in this process, provide a 3D view of a patient's treatment position for better setup and adaptability [2]. However, CBCT images have limitations, including a lower signal-to-noise ratio, and multiple artifacts like X-ray scattering and beam hardening compared to traditional CT scans. These drawbacks in CBCT hinder tasks like accurate dose recalculations and on-the-fly adjustments before dose delivery. While one approach has been to align the original planning CT with the CBCT using deformable registration [3], its accuracy is limited by artifacts and anatomical differences between the planning CT and daily CBCT.

To address these issues, creating a synthetic CT (sCT) from the routinely taken CBCT that matches the quality of the planning CT is a promising avenue. This synthetic CT would serve dual purposes: patient setup as well as dose recalculation for adaptive treatment planning. Deep learning techniques, specifically image-to-image translation tools like CycleGAN [4], have accelerated this transition. By inputting a CBCT scan, these models are trained to modify it, aiming to mimic the quality and details of planning and repeated CTs taken during treatment. Leveraging deep learning ensures superior anatomical precision by acting directly on the CBCT, bypassing the need to deform the CT. Furthermore, it offers a more effective correction than simply removing scattering/artifacts.

*i.hadzic@maastrichtuniversity.nl

However, effectively crafting deep learning models that cater to these needs remains a challenge. Ensuring the performance is both robust and of high quality is essential for its clinical viability. In our research, we tackle these challenges by focusing on unpaired adversarial networks, which are particularly suited for this application as they do not need a pixel-to-pixel match of the data, making them a more versatile solution for CBCT to CT translation in adaptive radiotherapy.

2. METHODS

2.1 Datasets

Two patient radiotherapy cohorts were utilized: *LungProton* and *CervixPhoton*, named after their respective anatomical sites and type of radiotherapy. The datasets differed in CBCT and CT scanner models and acquisition centers. Both datasets consisted of planning CT scans (pCT) and CBCT scans for radiotherapy planning and patient positioning during treatment. Notably, the *LungProton* set also contained repeated CT scans (rCT) to validate and potentially adjust the treatment plan for dose differences. Quality variations were observed, with *LungProton*'s CBCT scans displaying more artifacts and variations in CT numbers compared to the *CervixPhoton* set.
Phantoms, objects with known properties used in calibration and quality control, were incorporated to gauge the model's adaptability to out-of-distribution data. The *LungProton* phantom replicates a thorax and was scanned using the same devices as patient scans in the *LungProton* dataset. In contrast, the *CervixPhoton* phantom was scanned only with CBCT, using ideal values and known locations of the phantom insert plugs as ground truth.

2.2 Preprocessing

All scans were resampled to an optimal voxel spacing of 1.2965x1.26953 mm^3, windowed to a (-1000, 2000) HU range and normalized to the (-1, 1) range.
While training, a pair of CBCT and CT scans is randomly selected, followed by the truncation of CT field of view (FOV) to the CBCT's FOV, application of a body-masking algorithm, and sampling of CBCT and CT patch pairs that contain similar anatomy.
For validation and testing, matching CBCT-CT pairs from the same patient was required to ensure anatomical coherence. The pairing process involved selecting scans with minimal time difference between acquisitions. CT was then deformed to match the CBCT scan. All target and organ-at-risk (OAR) contours were subsequently transferred to the deformed CT.

2.3 Models

For our experiments, we investigated two unpaired generative adversarial frameworks - CycleGAN [4] and RevGAN [5]. RevGAN introduces invertible layers [6] to the CycleGAN framework to reduce its memory requirements, which is of particular interest in 3D applications. Specifically, invertible layers allow 1) recalculation of the intermediate activations instead of keeping them in the memory, and 2) performing forward passes in both directions. As a result, RevGAN does not only save memory through activation recalculation but also through sharing of the invertible layers between its two generators, reducing the number of parameters to train. In our study, we introduced partially-invertible VNet [7] generators to replace the partially-invertible ResNet [8] generators that RevGAN uses. We also introduce SSIM-cycle-consistency, which combines Structural Similarity Index Measurement (SSIM) [9] with the default L1 cycle-consistency loss. In Table 1, we list the configuration of each model we trained.

Table 1 Model configurations for all runs. The table shows the name of a model as referred to in-text, its architecture, patch size, and the number of iterations during training. Memory consumption during training is also shown, where AMP refers to memory usage with automatic mixed precision.

Model	Architecture	Patch size	Iterations	Memory/AMP
2DCyc	2D CycleGAN	384x384	200k	2.9/2.4 GB
3DCyc	3D CycleGAN	24x384x384	30k	45.9/24.9 GB

3DRev	3D RevGAN	32x384x384	30k	42/23.2 GB
3DRevSSIM	3D RevGAN + SSIM-cycle-consistency	32x384x384	30k	42/23.2 GB
3DRevSSIM Medium	3D RevGAN + SSIM-cycle-consistency	24x224x224	60k	11.7/6.9 GB
3DRevSSIM Small	3D RevGAN + SSIM-cycle-consistency	16x128x128	120k	3.4/2.9 GB

2.4 Evaluation

We employed the validation set for the best checkpoint selection and subsequently assessed the chosen model on a test set. For a quantitative understanding, we relied on metrics like MAE, PSNR, SSIM, MSE, and NMSE. For each training run we saved a total of thirty checkpoints at which we compared synthetic CT images against true CT, taking both full-image and body-masked views into account. The best checkpoint was determined through a combination of metric performance and visual inspection of axial mid-slices. During testing, we harnessed image-similarity metrics on the test set, with out-of-distribution evaluations using specific phantoms and qualitative views spanning axial, sagittal, and coronal orientations.

3. RESULTS

Visual evaluations (Figure 1) revealed that 2DCyc synthetic CTs closely resembled real CTs, although it had several drawbacks like drastic filling of air pockets and checkerboard patterns. Conversely, 3DCyc produced blurred CTs. 2DCyc also handled body motion CBCT artifacts effectively, yet occasionally altered the diaphragm shape. Importantly,

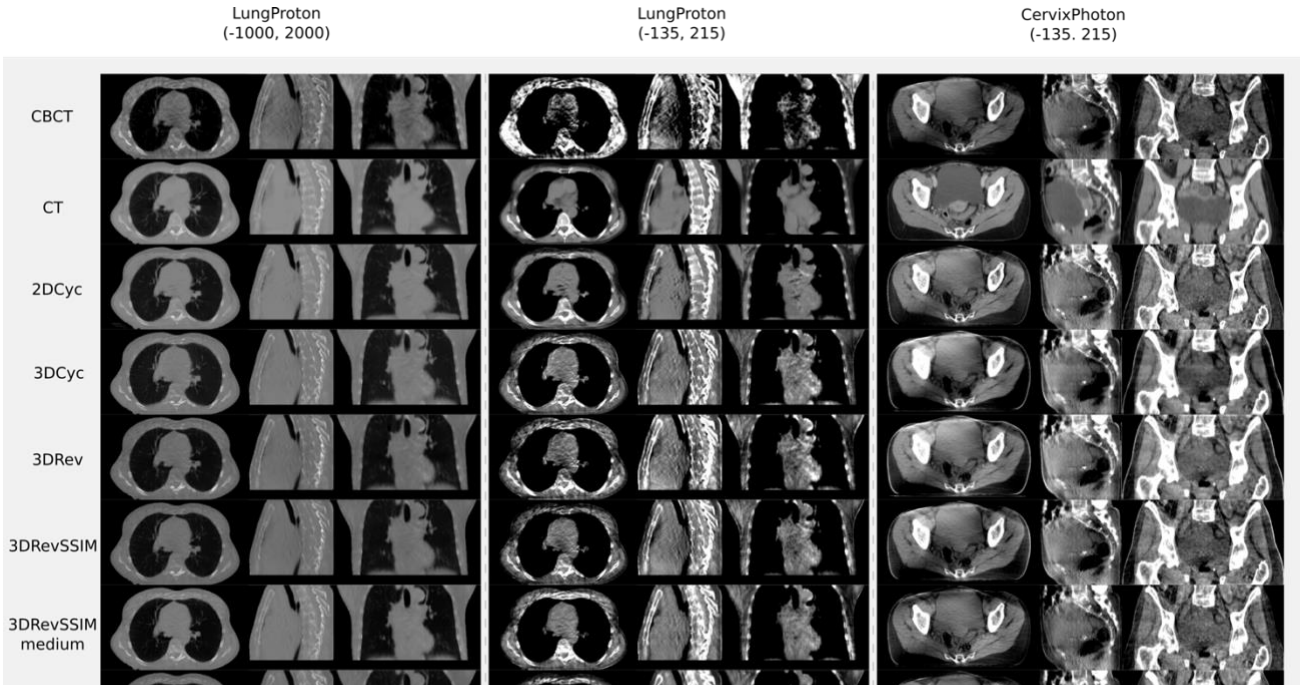

Figure 1. Qualitative results of translation for different models, shown on a patient chosen randomly from the test set, for both the LungProton and CervixPhoton datasets. The images are categorized based on the dataset and HU viewing. For each of these categories, the input CBCT, target CT, and translations from different models are shown (labeled on the left).

we suspect the underperformance of 3D models in artifact removal could stem from their under-training, due to computational constraints.

Regarding SSIM + L1 cycle-consistency, 3DRevSSIM generally outperformed 3DRev, correcting motion-related artifacts and presenting fewer overall artifacts. Notably, metrics and visual evaluations diverged, especially for *CervixPhoton*, highlighting the limited ability of metrics to detect certain visual artifacts (Table 2).

In the patch size study, 3DRevSSIMsmall showcased superior MAE, SSIM, and PSNR results. Visually, however, it generated artifacts missing in other versions, suggesting that reducing patch size might compromise model robustness. The inconsistencies between metrics and visual outcomes emphasize the potential pitfalls of solely relying on metrics.

When assessing performance on out-of-distribution phantoms, 3DRevSSIM demonstrated the best results for *LungProton*, while outcomes varied for *CervixPhoton* phantoms. Notably, artifacts observed in patient scans also manifested in phantom scans. Out-of-distribution data emphasized differences undetectable through metrics alone.

Table 2. Image similarity metrics (MAE, PSNR, and SSIM) for all the investigated models, as described in Table 1, on the *LungProton* and *CervixPhoton* test sets. The metrics for the best-performing models are shown in bold

Model	LungProton			CervixPhoton		
	MAE (HU)	PSNR (dB)	SSIM	MAE (HU)	PSNR (dB)	SSIM
No translation	140.03 (± 22.05)	27.75 (± 1.68)	0.902 (± 0.027)	93.45 (± 16.52)	30.21 (± 1.80)	0.9397 (± 0.015)
2DCyc	75.24 (± 22.31)	30.24 (± 2.44)	0.9351 (± 0.019)	58.30 (± 12.3)	32.54 (± 1.86)	0.9351 (± 0.014)
3DCyc	78.24 (± 23.6)	31.16 (± 1.99)	0.9401 (± 0.018)	75.6 (± 18.55)	31.79 (± 1.80)	**0.9470 (± 0.010)**
3DRev	87.18 (± 21.47)	30.48 (± 1.77)	0.9371 (± 0.019)	63.55 (± 13.07)	32.48 (± 1.85)	**0.9470 (± 0.011)**
3DRevSSIM	74.76 (± 28.87)	31.93 (± 2.24)	0.9453 (± 0.017)	66.09 (± 14.77)	32.38 (± 1.84)	0.9355 (± 0.0163)
3DRevSSIM medium	71.87 (± 22.38)	32.39 (± 1.9)	0.9482 (± 0.015)	62.91 (± 16.43)	**32.60 (± 2.15)**	0.9368 (± 0.0153)
3DRevSSIM small	**67.48 (± 17.73)**	**32.73 (± 1.97)**	**0.9488 (± 0.017)**	**52.97 (± 14.39)**	32.55 (± 2.06)	0.9429 (± 0.0130)

4. NEW OR BREAKTHROUGH WORK TO BE PRESENTED

2D vs. 3D Model Performance: We systematically compared 2D and 3D CycleGAN models for medical imaging. Our results highlight the trade-offs between these models: 3D models better preserve 3D structures but require longer training, while 2D models retain image content more accurately.

SSIM Cycle-Consistency Loss Function: We introduced the SSIM cycle-consistency loss function to CycleGANs. This metric offers a comprehensive assessment of image quality, ensuring the preservation of both image content and structure, thereby enhancing synthetic image quality.

Impact of Patch Size: We evaluated how patch size affects CycleGAN performance. Smaller patches lead to better metric performance and shorter training time but more artifacts, emphasizing the need to choose patch sizes thoughtfully to strike a balance between accuracy and artifact generation.

Evaluation Metrics Importance: We stressed the necessity for diverse evaluation metrics for translation models. Standard metrics might not fully capture image quality aspects, like artifacts, indicating a need for more sophisticated evaluation methods.

5. CONCLUSIONS

We have found a notable gap between standard image metrics and human visual evaluations, especially in medical image translation. Our exploration with CycleGANs for CT synthesis from CBCT scans highlighted:

1. Trade-offs between 2D and 3D models in terms of structure and content preservation.
2. The benefit of the SSIM cycle-consistency loss function in enhancing translation quality.
3. The critical balance between performance metrics, training time, and artifact generation based on patch size.
4. The necessity for holistic evaluation metrics beyond traditional ones.

Our study significantly advances the image-to-image translation field, stressing both the potential and challenges in CT synthesis.

REFERENCES

[1] Dawson LA, Sharpe MB. Image-guided radiotherapy: rationale, benefits, and limitations. The Lancet. Oncology. 2006 Oct;7(10):848-858. DOI: 10.1016/s1470-2045(06)70904-4. PMID: 17012047.

[2] Boda-Heggemann, J., Lohr, F., Wenz, F., Flentje, M., Guckenberger, M., 2011. kv cone-beam ct-based igrt. Strahlentherapie und Onkologie 187, 284–291

[3] Wang, H., Dong, L., Lii, M.F., Lee, A.L., De Crevoisier, R., Mohan, R., Cox, J.D., Kuban, D.A., Cheung, R., 2005. Implementation and validation of a three-dimensional deformable registration algorithm for targeted prostate cancer radiotherapy. International Journal of Radiation Oncology* Biology* Physics 61, 725–735.

[4] Zhu, J., Park, T., Isola, P., Efros, A.A., 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. CoRR abs/1703.10593. URL: http://arxiv.org/abs/1703.10593, arXiv:1703.10593.

[5] van der Ouderaa, T.F.A., Worrall, D.E., 2019. Reversible gans for memory-efficient image-to-image translation. arXiv:1902.02729.

[6] Gomez, A.N., Ren, M., Urtasun, R., Grosse, R.B., 2017. The reversible residual network: Backpropagation without storing activations. arXiv:1707.04585.

[7] Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. arXiv:1606.04797.

[8] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[9] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13, 600–612