Adapted nnU-Net: A Robust Baseline for Cross-Modality Synthesis and Medical Image Inpainting

Arthur Longuefosse, Edern Le Bot, Baudouin Denis de Senneville, Rémi Giraud, Boris Mansencal, Pierrick Coupé, Pascal Desbarats, Fabien Baldacci

Adapted nnU-Net: A Robust Baseline for Cross-Modality Synthesis and Medical Image Inpainting

Arthur Longuefosse[1], Edern Le Bot[1], Baudouin Denis De Senneville[2], Rémi Giraud[3], Boris Mansencal[1], Pierrick Coupé[1], Pascal Desbarats[1], and Fabien Baldacci[1]

[1] Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
[2] Univ. Bordeaux, CNRS, Bordeaux INP, IMB, UMR 5251, F-33400 Talence, France
[3] Univ. Bordeaux, CNRS, Bordeaux INP, IMS, UMR 5218, F-33400 Talence, France

Abstract. In medical image synthesis, the development of robust and reliable baseline methods is crucial due to the complexity and variability of existing techniques. Despite advances with architectures such as GANs and diffusion models, a clear state-of-the-art has yet to be established. This paper introduces a versatile adaptation of the nnU-Net framework as a robust baseline for both cross-modality synthesis and image inpainting tasks. Known for its superior performance in segmentation challenges, nnU-Net's automatic configuration and parameter optimization capabilities have been adapted for these new applications. We evaluate this method on two use cases: pelvis MR to CT translation using the Synthrad2023 challenge dataset and local synthesis using the BraTs 2023 inpainting challenge dataset. Standard synthesis metrics - MAE, MSE, SSIM and PSNR- demonstrate that our adapted nnU-Net outperforms GAN-based methods like pix2pixHD and ranks among the best methods for both challenges. We recommend this adapted nnU-Net as a new benchmark for medical image translation and inpainting tasks, and provide our implementations for public use on GitHub.

Keywords: Medical Image Synthesis · Medical Image Inpainting

1 Introduction

nnU-Net (or no-new-U-Net) [1] is a widely recognized method for the self-configuration of deep learning networks in medical imaging. Due to the numerous design decisions that can impact performance when selecting hyperparameters for training, nnU-Net was originally introduced to provide a reliable and consistent baseline for segmentation models by appropriately configuring settings for specific segmentation problems. It has been reported as the winning solution for numerous segmentation challenges and has achieved leading performance on several popular medical segmentation benchmarks. Recent reports have further solidified nnU-Net's position as a competitive model [2]. Comparisons between

the nnU-Net-configured baseline and various proposed architectural improvements, including state-of-the-art transformer-based architectures, have demonstrated that CNNs, given the same computational resources, can surprisingly match or even outperform these complex models.

Over the past decade, medical image synthesis has emerged as a significant trend in healthcare. This technique addresses the shortfall in medical datasets by producing synthetic data, thereby enhancing data diversity and improving the robustness of machine learning models. It also plays a crucial role in translating between different imaging modalities, which is beneficial for applications such as attenuation correction and radiation therapy planning. Additionally, medical image synthesis can significantly enhance patient comfort by reducing the need for repeated imaging procedures and minimizing exposure to harmful radiations. The significant advancement in this field is mainly attributed to the evolution of generative methods, such as generative adversarial networks [3] (GANs) and diffusion models [10] (DMs) . Numerous recent studies have explored synthetic data generation and cross-modality translation, addressing reconstruction challenges in anatomical areas such as the brain[11], pelvis [12], lungs [13].

In the specific area of image reconstruction, inpainting methods are also gaining interest for restoring missing or damaged parts of images in a plausible manner. This is especially relevant in brain MRI analysis, where traditional methods fail due to a lack of contextual information about disease-related anomalies. The introduction of challenges such as the BraTS Local Synthesis of Healthy Tissues [14] in 2023 underscores the growing emphasis on using inpainting to synthesize medical data. In practice, most state-of-the-art inpainting methods on natural images rely on diffusion models [17]. However, the heavy computational requirements of diffusion models and the processing of large 3D images necessitate custom strategies to manage memory requirements. For example, some propose using a 2D denoising model applied slice-wise on the axial plane [15]. Despite their effectiveness, diffusion models are hindered by the slow forward diffusion process, making them less suitable for time-constrained applications. Consequently, the primary approach for medical image inpainting remains Generative Adversarial Networks (GANs). For example, a 3D pix2pix [4] model was proposed as a baseline for the inpainting task in the BraTS2023 challenge.

Despite these advances, there remains a need for a robust and adaptable baseline in the fields of cross-modality translation and medical image synthesis. nnU-Net has already demonstrated its efficacy in the SynthRad2023 challenge [8] for medical image translation, ranking among the top methods. In this paper, we present a versatile adaptation of the nnU-Net framework to serve as a robust baseline for both medical translation and image inpainting tasks. We aim to extend nnU-Net's applications by enhancing its capabilities and adapting it to new tasks. We evaluate our adapted nnU-Net on two specific use cases: Pelvis MR to CT translation and brain local inpainting with healthy reconstruction, utilizing publicly available datasets from recent challenges. Furthermore, to facilitate broader use and collaboration, we enhance the frame-

work's capabilities and make the adapted code freely available on GitHub : `https://github.com/Phyrise/nnUNet_translation`

2 Methods

2.1 Adapted nnU-Net

In this section, we detail the adaptation of the nnU-Net framework for the tasks of image synthesis, focusing on preprocessing, architectural modifications, loss functions, validation techniques, and volume reconstruction.

Preprocessing We adapted the nnU-Net preprocessing procedures to support (image, image) inputs instead of the traditional (image, segmentation) inputs. Specifically, we disabled the cropping step in preprocessing because, by default, the cropping settings are not uniform across different modalities. For the remaining steps, we adhered to the nnU-Net guidelines [1]. This includes z-score normalization for MRI images and CT normalization for CT images, where intensity values are collected from the foreground classes, the mean and standard deviation are computed, values are clipped to the 0.5 and 99.5 percentiles, and then normalized by subtracting the mean and dividing by the standard deviation. Data augmentation was disabled to enable a fair comparison with state-of-the-art methods.

Architecture We retain the same architecture as nnU-Net, with minor adaptations to ensure the output is treated as an image instead of a segmentation. Specifically, we ensure that the output is retained in a floating point format, rather than being cast to integers. Additionally, we have added support for both nearest-neighbor and linear upsampling coupled with convolution layers as a replacement for transposed convolutions. This modification helps in reducing checkerboard artifacts when using a perceptual loss [6].

Losses and validation We adapt the loss function to support image synthesis instead of image segmentation. Specifically, we replace the Dice loss used in training and validation with a mean absolute error (L1) loss. Additionally, we support the use of mean squared error (L2) loss and 3D perceptual loss, leveraging the MedicalNet pre-trained model [7].

Volume reconstruction By default, nnU-Net operates using patches with a 50% overlap. For volume reconstruction, we replace the initial majority voting method with a mean reconstruction of patches. Support for median reconstruction has also been added. While median reconstruction enables a visually more accurate and detailed final output by preserving high-frequency details, it generally yields poorer results on global intensity metrics compared to mean reconstruction.

2.2 Comparison with GAN-based approaches

pix2pixHD [5], originally developed for synthesizing 2D natural images, is a Generative Adversarial Network [3] (GAN) that employs a multi-scale approach. It features a global generator and local generators, which work together to enhance image quality across various resolutions. This GAN architecture is characterized by its advanced adversarial loss, denoted feature-matching loss, which uses discriminator embeddings to ensure consistency in intermediate layer features. For our study, we implemented a 3D version of pix2pixHD, employing a patch-based approach with median reconstruction, similar to our modifications for nnU-Net.

3 Experiments

3.1 Datasets

Pelvis dataset For the synthesis of pelvic images, we use the open-access dataset from Task 1 of the SynthRAD2023 challenge. This dataset includes paired MR and CT images from 180 patients used for training, with an additional 30 patients designated for validation, with no ground truth CT publicly available. The dataset provides high-quality, well-aligned MR and CT image pairs, which eliminates the need for additional registration processes. The voxel sizes in this dataset are standardized to 2.5 $\times$ 1 $\times$ 1 mm, ensuring consistent spatial resolution across all images. The results of this MR to CT translation task were obtained using the SynthRad2023 challenge online validation server.

Brain dataset The 2023 BraTS Local Synthesis of Healthy Brain Tissues Training dataset is comprised of 1251 T1w images pre-processed with bias field correction, spatial resampling to 1 $\times$ 1 $\times$ 1 mm, and cropped to a final spatial resolution of 155 x 240 x 240. For each image, a corresponding healthy and non-healthy tissue mask is provided. Non-healthy tissue mask, delineates areas affected by tumoral invasion, while healthy tissue mask are generated according to the process described in [14]. The dataset originally included a number of images for validation and testing, hosted on the synapse servers, however, evaluation on those images is no longer available since the closure of submission queues. As such, 251 subjects were randomly sampled and kept for testing. Remaining subjects were used for training.

3.2 Implementation details

The training setup for nnU-Net closely followed the guidelines set in the original publication [1], with the models undergoing 1000 epochs of training with 250 iterations per epoch. The automatic calculation of patch sizes resulted in dimensions of 64$\times$128$\times$224 for the pelvis task and 96$\times$160$\times$160 for the brain task. In contrast, the pix2pixHD model was trained using a similar procedure as defined in the original paper [5], with a training duration of 200 epochs with

1000 iterations per epoch, and a uniform patch size of 128×128×128.
All models were trained using PyTorch 2.3.0 and automatic mixed precision to enhance computational efficiency and accelerate the training process. Trainings were configured with a batch size of 2 and completed within 24 to 48 hours, employing an NVIDIA RTX 6000 with 24GB of VRAM. An 80/20 training/validation split was maintained for robust evaluation in both tasks.

3.3 Evaluation

We employ the same metrics used in each challenge to facilitate direct comparison with other methods. For the translation task, we report Mean Absolute Error (MAE), Structural Similarity Index (SSIM), and Peak Signal to Noise Ratio (PSNR). For the inpainting task, we report Mean Squared Error (MSE), SSIM, and PSNR. These metrics are intensity-based and, while they do not fully capture the visual appearance and medical conformity of the images, they are effective for comparing our results with state-of-the-art methods.

4 Results

4.1 MR to CT translation

Table 1 presents a quantitative evaluation of the adapted nnU-Net's performance based on MAE, SSIM, and PSNR metrics between synthesized and ground truth CT images. All metrics for our models are calculated using the SynthRAD2023 challenge validation server on 30 patients from the validation set. This table shows that using a standard 3D pix2pixHD model yields inferior performance compared to the nnU-Net for cross-modality synthesis. Results from the challenge report [9] are also listed for reference, although they are based on the test set and not on the validation set. The trends between nnU-Net with MAE and nnU-Net with MSE are strongly correlated. The challenge ranking is based not only on reported intensity metrics but also on dose evaluation metrics, explaining the observed ranking.
In Fig. 1, a comparison of MR to CT synthesis on the validation set is shown. We observe that nnU-Net with MAE or MSE both perform exceptionally well in reconstructing CT scans, demonstrating greater precision than pix2pixHD, especially for bone reconstruction and body contouring. nnU-Net with MedicalNet also shows visually high-quality results, challenging the accuracy of the intensity metrics reported in Table 1, which indicated poor performance for this model. This discrepancy suggests that visual assessment may provide additional insights beyond the standard intensity-based metrics.

4.2 Local Inpainting

The quantitative validation results for the local inpainting task are presented in Table 2. Additionally, results from participant reports on the BraTS local

Table 1. Comparison of 3D networks performances using MAE, SSIM, and PSNR metrics between real CT and synthesized CT for the pelvis MR to CT translation task. Evaluation is performed on 30 patients.

Model	MAE	SSIM	PSNR
3D nnU-Net MAE	$\mathbf{65.74 \pm 12.72}$	$\mathbf{0.851 \pm 0.030}$	$\mathbf{27.94 \pm 1.48}$
3D nnU-Net MedicalNet	252.36 ± 53.71	0.731 ± 0.050	23.09 ± 1.69
3D pix2pixHD	88.57 ± 17.53	0.815 ± 0.029	26.64 ± 1.33
Results from Synthrad2023 challenge report			
1st - SMU-MedVision			
3D Hybrid Transformer U-Net	$\mathbf{58.83 \pm 13.41}$	$\mathbf{0.885 \pm 0.029}$	$\mathbf{29.61 \pm 1.79}$
2nd - Jetta_Pang			
3D nnU-Net MSE	65.73 ± 13.75	0.869 ± 0.032	28.38 ± 1.68
3rd - FAYIU			
3D Swin U-NetR	61.72 ± 13.32	0.876 ± 0.030	28.83 ± 1.61

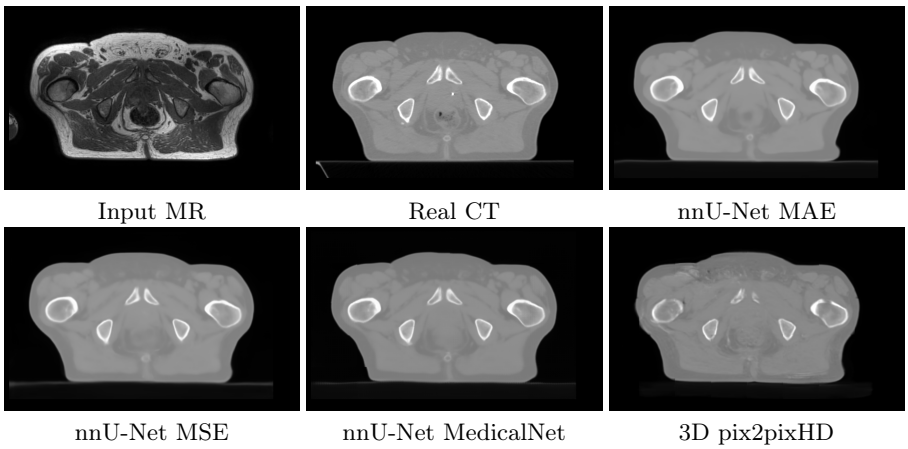
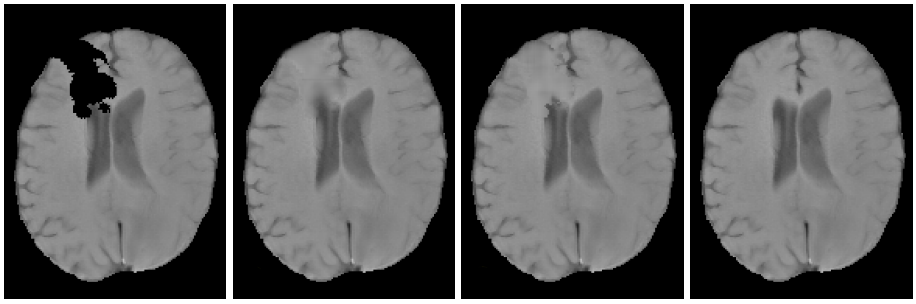


Fig. 1. Comparison of axial slices between input MR image, ground truth CT, and synthesized CT from several nnU-Net implementations and 3D pix2pixHD model

inpainting test dataset are included, with the key difference being that the evaluations were conducted on the test set, whereas our approach utilized a subpart of the training set due to the closure of the BraTS challenge validation server. This table shows that our adapted nnU-Net with a standard MSE loss outpeforms both 3D pix2pixHD and 2D DDPM, and closely matches the 3D pix2pix. Fig. 2 presents an example of a T1-weighted input slice alongside its corresponding ground truth, and reconstruction from best nnU-Net and 3D pix2pix models. Both models attempt to reconstruct the voided areas but differ in their approaches. The nnU-Net, which optimizes the MSE loss known to smooth output intensities, tends to produce blurred regions while preserving general image structure. Conversely, the 3D pix2pixHD model, which also optimizes the GAN loss, attempts to recreate these structures but results in outputs that appear unnatural and show distinct boundaries with the surrounding image areas.

Table 2. Comparison of 3D networks performances using MSE, SSIM, and PSNR metrics between inpainted and original brain images. Evaluation is performed on 238 patients.

Model	MSE	SSIM	PSNR
3D nnU-Net MAE	0.0129 ± 0.0063	$\mathbf{0.797 \pm 0.113}$	19.45 ± 2.63
3D nnU-Net MSE	$\mathbf{0.0116 \pm 0.0056}$	0.797 ± 0.115	$\mathbf{19.87 \pm 2.55}$
3D nnU-Net MedicalNet	0.0142 ± 0.0064	0.749 ± 0.143	18.86 ± 2.11
3D pix2pixHD	0.0141 ± 0.0078	0.787 ± 0.129	19.28 ± 3.11
Results from available participants reports			
N/A - Juexin Zhang, et al. [16]			
3D pix2pix	$\mathbf{0.0111 \pm 0.0068}$	$\mathbf{0.807 \pm 0.129}$	$\mathbf{19.30 \pm 2.49}$
2nd - Alicia Durer et al. [15]			
DDPM 2D slice-wise	0.0160 ± 0.0118	0.785 ± 0.155	18.71 ± 3.08



T1w voided | nnU-Net MSE | 3D pix2pixHD | Ground-Truth

Fig. 2. Comparison of axial slices between the T1w voided image, reconstructions from the best nnU-Net and 3D pix2pixHD models, and the original T1w image

4.3 Computational resources

Table 3. Comparison of networks computational resources in terms of memory cost on GPU, number of parameters and training time per iteration. Here, (G) denotes the generator and (D) the discriminator.

Model	Memory Cost	Parameters	Time per iter.
nnU-Net MAE	8.3 GB	30.8 M	0.150 s
3D Pïx2PixHD	15.9 GB	547.3M (G) + 22.1M (D)	0.630 s

Table 3 provides a comparison of the computational resources required by different network models in terms of GPU memory usage, number of parameters, and training time per iteration. The resources are evaluated for both the generator (G) and the discriminator (D) components where applicable. This table highlights the differences in computational efficiency and resource demands among the models used for MR to CT translation and local inpainting tasks.

5 Discussion and Conclusion

This study presents a robust baseline for medical image synthesis based on an adapted nnU-Net framework. Our findings reveal that this method effectively reconstructs medical images and consistently outperforms GAN-based approaches, challenging the prevalent reliance on such methods within the field. Specifically, for the two evaluated tasks—Pelvis MR to CT translation and brain local inpainting with healthy reconstruction—the adapted nnU-Net, configured with a simple Mean Absolute Error (MAE) loss, produces solid results without the need for a discriminator.

This success can be attributed to two key aspects of the framework. Firstly, nnU-Net's inherent adaptability and optimization capabilities ensure that it is well-suited to a variety of medical imaging tasks. It is designed to automatically configure itself optimally for the task at hand, making it highly versatile and effective. Secondly, the use of a U-Net architecture allows the model to focus directly on the defined cost function, in contrast to traditional GANs, which rely on discriminators that prioritize generating images that visually resemble the training dataset. Furthermore, the use of a simple U-Net architecture results in a lighter, more efficient model. As detailed in Table 3, nnU-Net operates with fewer parameters and less computational costs, while also reducing training times compared to the more complex pix2pixHD. This efficiency not only facilitates faster training cycles and inference but also reduces the likelihood of overfitting or convergence issues commonly associated with GANs.

Although we did not enable data augmentation in our primary experiments to maintain consistency in comparisons, our adapted nnU-Net framework does include capabilities for data augmentation. In trials, data augmentation did not significantly enhance performance for these synthesis tasks, likely because the databases employed were already diverse and of high quality.

Another crucial aspect of this adapted nnU-Net is its contribution to the reproducibility of results. Reproducibility is often a challenge in medical image synthesis, where slight variations in model implementation can lead to significantly different outcomes. The nnU-Net framework helps standardize preprocessing and model configuration, which, as observed in the various implementations of models like pix2pix during the SynthRad2023 challenge [9], can lead to disparate results. By automating many aspects of its setup, nnU-Net makes it easier to replicate results, ensuring that outcomes are consistent and reliable across different studies.

For the inpainting task, while nnU-Net delivers robust performance in terms of intensity-based metrics, it tends to produce images with blurry tissues due to the use of MAE/MSE loss. Although effective in minimizing intensity errors, these losses fail to ensure detailed anatomical structure preservation, highlighting the need for more suitable loss functions. Additionally, recent works on pseudo 3D diffusion models [15] indicate they surpass other methods in the inpainting task, suggesting that diffusion models are particularly well-suited for these challenges. These models often incorporate a U-Net for the denoising process, suggesting that our adapted nnU-Net could serve effectively as this backbone.

Looking ahead, we plan to extend the adapted nnU-Net framework by incorporating an adaptive discriminator component and developing loss functions based on anatomical structures rather than global intensities. This aims to generate images that are both anatomically and medically realistic, further enhancing the utility and accuracy of the model in clinical applications. By sharing our framework on GitHub, we invite the community to collaborate and build upon this work, potentially setting new standards in medical image synthesis.

Acknowledgments

This work was granted access to the HPC resources of IDRIS under the allocations 2022-AD011013848R1 and 2022-AD011013926 made by GENCI.
This work benefited from the support of the project HoliBrain of the French National Research Agency (ANR-23-CE45-0020-01).
This project is supported by the Precision and global vascular brain health institute funded by the France 2030 investment plan as part of the IHU3 initiative (ANR-23-IAHU-0001).
This study received financial support from the French government in the framework of the University of Bordeaux's France 2030 program / RRI "IMPACT and the PEPR StratifyAging.

References

1. F. Isensee, P.F. Jaeger, S. A. Kohl et al. : nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2), 203-211. (2021)
2. F. Isensee, T. Wald, C. Ulrich et al. : nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. arXiv preprint arXiv:2404.09556. (2024)
3. I. Goodfellow, J. Pouget-Abadie, Mirza, B. Xu, et al. : Generative adversarial nets. In: Advances in neural information processing systems. pp 2672–2680 (2014)
4. P. Isola, J. Zhu, T. Zhou, A. Efros : Image-to-Image Translation with Conditional Adversarial Networks. CVPR, pp 5967-5976. (2017)
5. T. Wang, M. Liu, J. Zhu, et al : High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. CVPR, pp 8798-8807. (2018)
6. J. Johnson, A. Alahi, L. Fei-Fei : Perceptual Losses for Real-Time Style Transfer and Super-Resolution. ECCV 2016 (2016)
7. S. Chen, K. Ma and Y. Zheng : Med3D: Transfer Learning for 3D Medical Image Analysis. arXiv preprint arXiv:1904.00625 (2019)
8. A. Thummerer, E. van der Bijl, A. Galapon Jr et al. : SynthRAD2023 Grand Challenge dataset: Generating synthetic CT for radiotherapy. Medical physics, 50(7), 4664-4674. (2023)
9. E. Huijben, M.L. Terpstra, A. Galapon Jr et al. : Generating Synthetic Computed Tomography for Radiotherapy: SynthRAD2023 Challenge Report. arXiv preprint arXiv:2403.08447. (2024)
10. R. Rombach, A. Blattmann, D. Lorenz et al. : High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10684-10695 (2022)

11. J. Wolterink, A. Dinkla, M. Savenije et al. : Deep MR to CT Synthesis using Unpaired Data. (2017)
12. Y. Lei, J. Harms, T. Wang et al. : MRI-Only Based Synthetic CT Generation Using Dense Cycle Consistent Generative Adversarial Networks. Medical Physics. (2019)
13. A. Longuefosse, J. Raoul, I. Benlala et al. : Generating high-resolution synthetic CT from lung MRI with ultrashort echo times: initial evaluation in cystic fibrosis. Radiology, 308(1), e230052. (2023)
14. F. Kofler, F. Meissen, F. Steinbauer et al. : The Brain Tumor Segmentation (BraTS) Challenge 2023: Local Synthesis of Healthy Brain Tissue via Inpainting. arXiv preprint arXiv:2305.08992. (2023)
15. A. Durrer, J. Wolleb, F. Bieder et al. : Denoising Diffusion Models for 3D Healthy Brain Tissue Inpainting. arXiv preprint arXiv:2403.14499. (2024)
16. R. Zhu, X. Zhang, H. Pang et al. : Advancing Brain Tumor Inpainting with Generative Models. arXiv preprint arXiv:2402.01509. (2024)
17. A. Lugmayr, M. Danelljan, A. Romero et al. : Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11461-11471). (2022)