Medical Flow Matching CT Translation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Medical imaging still lacks reliable methods to translate contrast-enhanced arterial-phase CT scans into their non-contrast (native) form. We introduce **Medical Flow Matching (MFM)**, which combines an efficient image-translation paradigm with a bottleneck attention mechanism designed for medical images. On the held-out test set, MFM achieves **MAE** $= 6.229$ HU, **SSIM** $= 0.992$, and **PSNR** $= 38.933$, markedly outperforming traditional methods while cutting generation time by 400 times compared with DDPM. With two-way conversion using MFM, we can double the size of available public pathology-image CT datasets. Training nnUNetv2 on MFM-generated (from contrast) images achieves a **Dice score** of 0.926 versus 0.966 when trained on real (non-contrast) images – 95 % of baseline performance. By removing the need for an additional native scan, MFM can reduce the radiation dose per examination by about 50 %, lowering patients' cumulative exposure and thereby decreasing the risk of radiation-associated diseases.

1 Introduction

In recent years, medical imaging has become central to modern healthcare, providing essential information for patient diagnostics, treatment planning, and monitoring disease progression. Machine learning methods have significantly revolutionized this domain by automating routine radiological tasks [1, 2], showing strong predictive abilities for disease diagnosis [3, 4], and helping reduce the clinical workload of medical specialists [5, 6].

However, building effective machine-learning models for medical imaging still faces big challenges. First, the lack of annotated medical datasets is a major barrier. Rare diseases are usually underrepresented because of their low prevalence, making it harder to gather enough data to train specialized models. Public datasets often include only one modality, for example, contrast-enhanced series, whereas native (non-contrast) images may be scarce or missing. Second, labeling medical images demands specialized expertise, so annotation costs are much higher than in general image-labelling tasks. High-quality, consistent annotations are vital for accurate predictive models, which makes the problem even worse.

Computed tomography (CT) is a key tool in clinical diagnosis, giving important insights into anatomy through both native (non-contrast) and contrast-enhanced scans. The latter involves administering a contrast agent to enhance vascular visibility and support precise diagnosis. Currently, due to the trend toward reducing patient radiation exposure during CT examinations, non-contrast (native) series are not always performed. However in certain cases, they are still required specific findings. Therefore, building reliable methods to convert contrast-enhanced CT scans (arterial phase) into native images without contrast media remains important but challenging.

Our approach matters not only because it can help researchers and developers expand the modality of existing datasets, but also because it offers a new way to utilize contrast-enhanced images when

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

native scans are unavailable. [7] estimates that CT studies in the United States in 2023 could lead to about 103 000 future skin cancers. Our approach could help reduce this number by enabling the creation of potentially diagnostically useful scan sequences without increasing radiation exposure. Although dual-energy CT scanners are already capable of generating virtual non-contrast images, they remain relatively rare [8]. Our method could help move the field in that direction by making this opportunity more accessible.

In this paper, we present a Flow Matching method and TimeResNet model designed to translate arterial-phase contrast-enhanced CT images into corresponding native CT scans. Our proposed method and model address the challenges of limited data high annotation costs by effectively leveraging existing data, which may reduce the need for additional medical image annotations and lower the risks associated with radiation exposure.

This paper offers three main contributions:

1. We propose a new medical image-to-image translation method, that needs far less memory that typical 3D approaches yet keeps consistency across axial slices.
2. We propose a neural network architecture (TimeResNet) for this method, which achieves strong results, compared with existing architectures in MONAI [9].
3. The method can translate contrasts to natives and natives to contrast, letting users train a single network instead of two specific tasks; this halves the training time.

To simplify further narration, by a Native Image or Native we mean a CT scan that was performed without any contrast enhanced. Under Contrast Image or Contrast, we refer to the Arterial phase of CT examination, usually after 15-35 seconds after administration of iodine-containing contrast.

2 Related work

Given the challenges outlined above, it is therefore unsurprising that the generation of synthetic medical images has become increasingly popular. Initially, classical techniques [10, 11, 12, 13, 14] were employed; these methods proved both effective and applicable across a range of tasks, yet they nonetheless faced limitations in producing truly realistic medical imagery. The advent of deep learning subsequently enabled the synthesis of more complex images. Soon thereafter, deep neural networks began to appear, initially demonstrating marked improvements in image quality and generalization performance relative to classical approaches.

As generative adversarial networks (GANs) matured, they found extensive application within medical imaging—for example, in MRI-to-CT translation [15], in MRI-PET translation [16, 17, 18, 19, 20, 21, 22], image reconstruction [23, 24], and super-resolution [25, 26]. One of the most widespread uses of GANs has been as a data-augmentation strategy to bolster the training of models for both segmentation and classification tasks.

Diffusion models have gained considerable traction in recent years, demonstrating substantial promise in the synthesis of medical images owing to their ability to generate high-quality outputs, their stable training dynamics, and their flexibility during optimization [27]. Studies [28, 29, 30, 31, 32, 33, 34, 35, 36, 37] have confirmed the efficacy of these diffusion-based approaches. In two-dimensional medical image generation, state-of-the-art models capture intricate anatomical details with negligible distortion, to the extent that they are occasionally deemed suitable for clinical deployment. In the domain of three-dimensional imaging, NVIDIA's recent publication [38] on whole-patient generation employed diffusion processes within a latent representation space to achieve remarkably realistic volumetric outputs.

In the broader field of computer vision, there has been a marked trend toward generative techniques based on flow matching [39, 40] as well as conditional generation frameworks [41, 42]. Such methodologies are increasingly being adopted within medical imaging, particularly for tasks involving image translation between different modalities [43, 44].

In this work, we focus on building a robust image-translation pipeline. We introduce targeted tweaks to a proven architecture so it works well across different translation tasks without heavy tuning. For instance, we have an AbdomenAtlas [45] dataset for 8,448 CT volumes with masks. Most of them are studies with contrasts. Our method can almost double this dataset, which will allow us to use

data for specific tasks. This approach not only conserves time and computational resources but also enables researchers to focus on other critical aspects, thereby enhancing overall efficiency in a range of clinical scenarios.

3 Methods

To show how well our approach works, we compare it with several alternatives. Our simplest baseline is a U-Net–style network: the contrast image goes into the encoder, and the decoder tries to reproduce the native image. We train this network with mean absolute error (MAE). To be sure that any performance gap is not caused by the choice of architecture alone, we test multiple network setups and loss functions, following the procedure in [46]. $\mathcal{L}_{\text{MAE}}(\theta) = \|g_\theta(x_C) - x_N\|_1$,where g_θ – neural network that we will train for regression task. In our experiments we set $p = 1$.

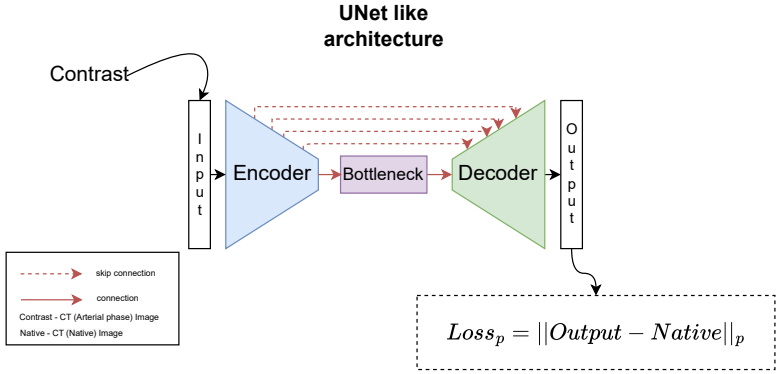

Figure 1: Baseline regression scheme.

3.1 Diffusion

So if we want to use diffusion for conditional image to image translation, we train a UNet denoiser for the following loss. We will apply DDPM [47] to conditional image-to-image generation. π_0 – distribution of target images, π_1 – distribution of conditional images. $\pi_0 \times \pi_1$ – joint distribution of paired target and conditional images.

$$\mathcal{L}_{\text{DDPM}}(\theta) = \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,1)}\mathbb{E}_{t\sim\mathcal{U}\{0,\dots,T\}}\mathbb{E}_{(\mathbf{x}_0,\mathbf{x}_1)\sim\pi_0\times\pi_1} \left\|\varepsilon - \varepsilon_{\theta,t}\left(\alpha(t)\cdot\mathbf{x}_0 + \sigma(t)\cdot\varepsilon, \mathbf{x}_1\right)\right\| \quad (1)$$

We have conditions on $\alpha(t)$ and $\sigma(t)$. $\alpha(t), \sigma(t)$ such that $\forall t \in \{0, ..., T\} \to \alpha(t)^2 + \sigma(t)^2 = 1$. $\mathcal{U}\{0, ..., T\}$ – is a uniform integer distribution.

If we apply this method in our case, we will get that: π_N – distribution of Native images (target images), π_C – distribution of Contrast images (conditional images), $\pi_N \times \pi_C$ – joint distribution of paired Native and Contrast images. So we choose standard linear scheme for β_t [47], using the notation $\alpha_t = 1 - \beta_t$ and $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, so we have $\alpha(t)^2 = \overline{\alpha}_t$, $\sigma(t)^2 = 1 - \overline{\alpha}_t$, $\beta_t = \beta_{start} + (\beta_{end} - \beta start)\cdot\frac{t}{T}$, where $T, \beta_{start}, \beta_{end}$ – hyperparameters that are set before training.

$$\mathcal{L}_{\text{DDPM}}(\theta) = \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,1)}\mathbb{E}_{t\sim\mathcal{U}\{0,\dots,T\}}\mathbb{E}_{(\mathbf{x}_N,\mathbf{x}_C)\sim\pi_N\times\pi_C} \left\|\varepsilon - \varepsilon_{\theta,t}\left(\sqrt{\overline{\alpha}_t}\cdot x_N + \sqrt{1-\overline{\alpha}_t}\cdot\varepsilon, x_C\right)\right\| \quad (2)$$

For sampling we used DDIM-like scheme, to obtain deterministic images from the initial noise.

$$\mathbf{x}_{t-1} = \sqrt{\frac{1}{1-\beta_t}}\cdot x_t + \left[\sqrt{1-\overline{\alpha}_t} - \sqrt{\frac{1-\overline{\alpha}_{t-1}}{1-\beta_t}}\right]\cdot\varepsilon_\theta\big([\mathbf{x}_{\text{img}}, \mathbf{x}_t], t\big)$$

It is also important to understand that we must use the same noise from which the image will be generated for all axial images of a single patient, otherwise this will lead to a very large heterogeneity in the relative and axial projections.

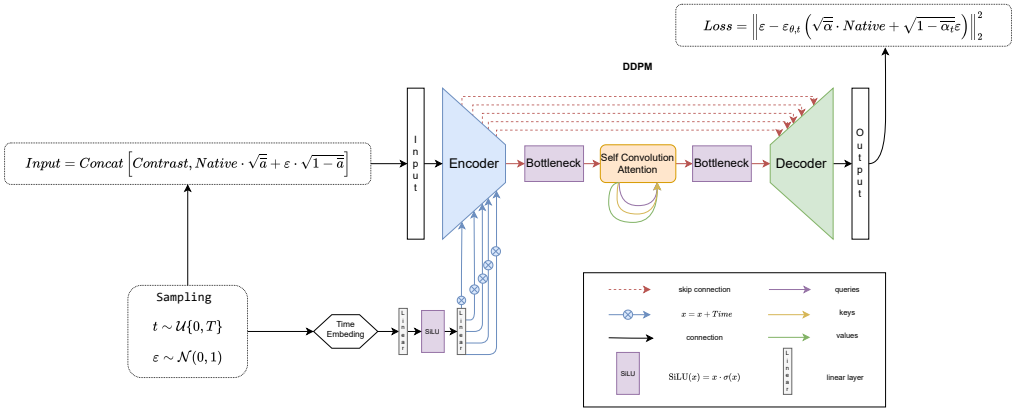

Figure 2: DDPM scheme. Neural network architecture and a visual diagram of training.

3.2 Flow Matching

Let $\{\pi_t\}_{t\in[0,1]}$ be a *probability path* (*probability path is a continuous trajectory of probability distributions that interpolates between an initial distribution and a target distribution over time.*) that continuously interpolates between an easy-to-sample reference π_0 (e. g. Gaussian noise) and the data distribution π_1. Samples $\mathbf{x}_t \sim \pi_t$ evolve according to an ordinary differential equation (ODE)

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = f_\theta(\mathbf{x}_t, t), \qquad \mathbf{x}_0 \sim \pi_0,\ \mathbf{x}_1 \sim \pi_1. \tag{3}$$

In our case we define $\pi_0 \times \pi_1$ - as pairs of images of the same patient but from different series, where π_0 - distribution of native images, π_1 - distribution of contrast images.

Define the convex combination $t \in [0, 1]$, $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$

and the *target velocity*

$$\mathbf{v}_t = \frac{\partial \mathbf{x}_t}{\partial t} = \mathbf{x}_1 - \mathbf{x}_0. \tag{4}$$

Flow Matching (FM) trains a neural vector field f_θ by regressing onto $\mathbf{v}_t$ over random $(\mathbf{x}_0, \mathbf{x}_1)$ pairs and $t \sim \mathcal{U}[0, 1]$:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1]}\mathbb{E}_{(\mathbf{x}_0,\mathbf{x}_1)\sim\pi_0\times\pi_1}\left[\|f_\theta(\mathbf{x}_t, t) - \mathbf{v}_t\|_2^2\right]. \tag{5}$$

Crucially, Eq. (5) is *simulation-free*: no stochastic differential equations need to be solved during training.

Below is a universal diagram of how the model works and learns using the Flow Matching method.

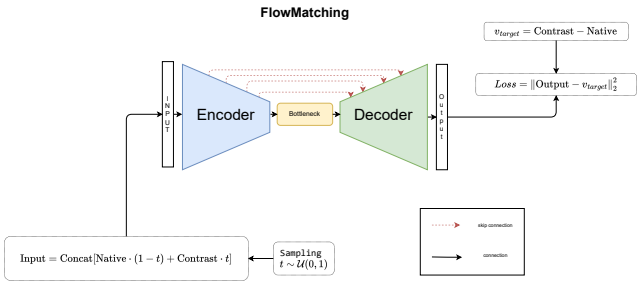

Figure 3: Flow Matching scheme.

We also developed and proposed our own model architecture, TimeResNet.

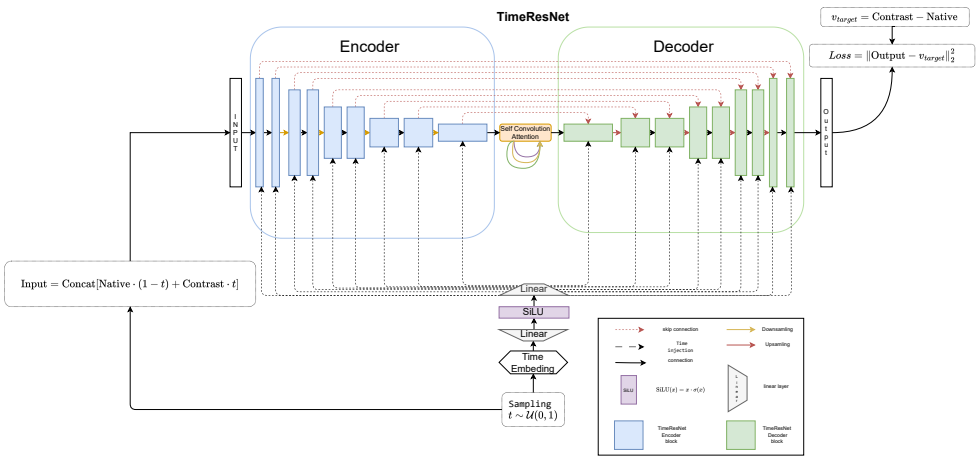

Figure 4: TimeResNet in Flow Matching scheme.

A more detailed description of the Self-Convolution Attention block is given below. We determined that an explicit attention mechanism was necessary to achieve high performance; however, widely adopted architectures such as SwinUNET [48] – which employ a limited, patch-based attention scheme—did not yield substantial gains. Consequently, we elected to insert our attention module between the Bottleneck layers, drawing inspiration from Yang et al. (2019) [49]. To stabilize training, we further incorporated Group Normalization [50] into the block.

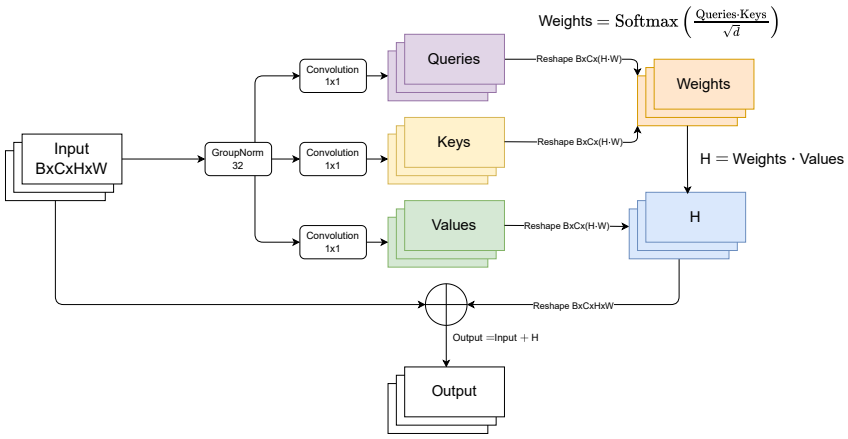

Figure 5: Self convolution attention scheme.

After training, sampling reduces to integrating Eq. (3) from $t = 1$ to $t = 0$ with a numerical ODE solver. We have the opportunity to sample both motifs into contrasts and natives from contrasts.

We make full condition on linear combination (3.2) to improve stability of the work and remove stochasticity. To correctly add time intervals to our model, we do the following. We use a standard scheme with sine and cosine embeddings (link to the article), then we feed it into the MLP head, the results of which are fed into each of the ResNet blocks, which initially also uses a linear layer to translate the size to the current number of channels, and then the intermediate values are used as normalization. $h = \mathbf{A} \cdot h + \mathbf{B}$, where $\mathbf{A}$ and $\mathbf{B}$ are linear and trainable parameters in each block. A more detailed scheme for adding Time embeddings is presented in B

4 Experiments

A total of 120 abdominal CT studies (52561 images) were retrospectively collected, each containing both native and arterial phases (one study per patient), from 3 different CT stations from 3 different hospitals. The dataset was split into 80 (34992 images) training, 20 (8197 images) test, and 20 (9372 images) held-out patients. Seventy imaging studies were obtained as follows: first, the anatomical region of interest was localized in both the native and contrast-enhanced series, next, the contrast-enhanced series were registered to the native series.

All images were taken in the original (512×512) shape and clipped to $[-1000, 1000]$ HU before scaling to $[-1, 1]$. As augmentations, we used axis-aligned flips and affine transformations. No intensity-based perturbations were used in order to preserve Hounsfield integrity.

All models were trained for 30 epochs with Adam optimizer, batch size was equal 2. We conducted all the experiments on a node of four NVIDIA RTX 3090 GPU (24 GB). To implement the experiments, we used the PyTorch [51] and MONAI framework.

4.1 Training Dynamics

Detailed model learning trajectory is presented below. We see that the graphs have reached a plateau, but we assume that further scaling in terms of the amount of data, training time, and model size may yield better results, but we tried to use all the resources available to us to present the most valid results in our opinion.

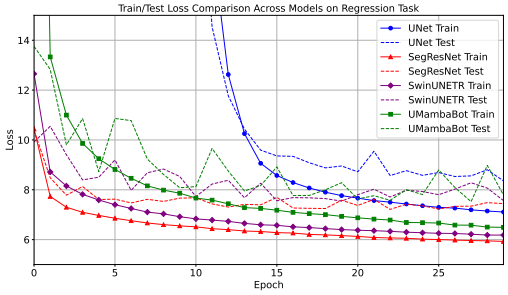

Figure 6: Dynamics of training different models in the regression task.

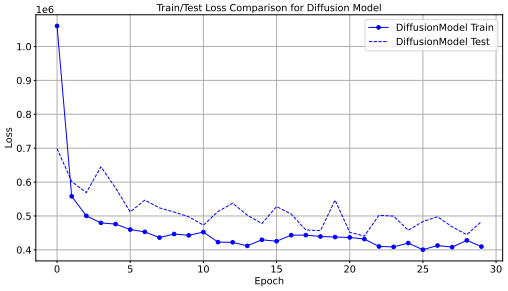

Figure 7: Dynamics of training model in the diffusion task.

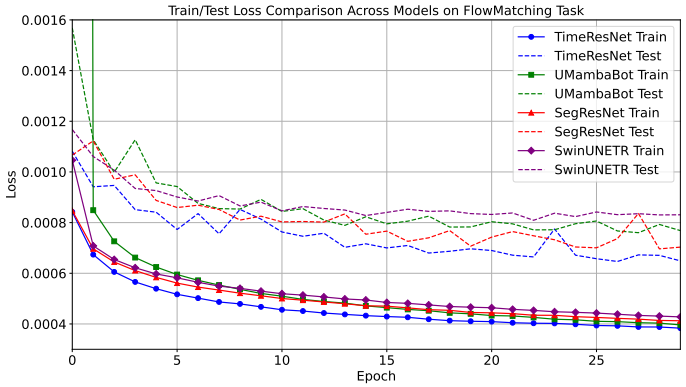

Figure 8: Dynamics of training different models in the flow matching method.

We can note that the TimeResNet model we have proposed shows the best convergence compared to the most popular options among CNN, Transformer, and SSM. Umamba architecture needs a certain number of epochs to start showing good results.

4.2 Visual results

In order to evaluate the generation quality of our model in an additional way, we conducted the following experiment. We generated native images from contrast using the SegResNet model for the regression task, generated images using diffusion, and generated images using the TimeResNet model. Then we transferred 20 pairs consisting of a real native and a generated sample to qualified radiologists to blindly determine where the generated image is and where the original one is. Of the 20 examples generated by regression, all 20 examples were correctly classified, and the diffusion examples had the same result. But using the example of our TimeResNet model, the doctor correctly selected 17 cases out of 20. Next, we decided to conduct an experiment where we provide only one image and ask the reader to label it as a generated or real. We took 20 new cases and generated them in a similar way, so that 10 images were generated and 10 were real. In the same experiment with images generated by diffusion and regression, similar results were obtained and all 20 images were selected correctly. But in the pictures generated using TimeResNet, 14 of the 20 pictures were already correctly selected. It is worth mentioning that our 2 doctors in both experiments used a professional application for viewing medical images - Slicer3D, which allows you to carefully view the image in all 3 projections simultaneously, the doctors performed the screening in the adbomen window. The following conclusions can be drawn from the generated images: in those produced by the regression-based approach, organs such as the kidneys, liver, and aorta are recognizable, but the model tends to blur fine details and smooth out artifacts present in the original CT scans. Diffusion is the easiest to guess due to the fact that the image is unnaturally shifted when viewed in a sagittal projection. But there are difficulties in the pictures obtained using Flow Matching, so with paired studies you can find some details in the liver and kidneys that simply cannot be seen in native images due to the specifics of CT scans. Further improvement of this method can be achieved by improving the quality of the training dataset. You can see from the example that we presented that the difference between a native and contrast image is not only in terms of organs that differ in intensity, but also along the border of the human body. This is due to the fact that the person still moves a little during the CT scan. An example of generation can be found in 10.

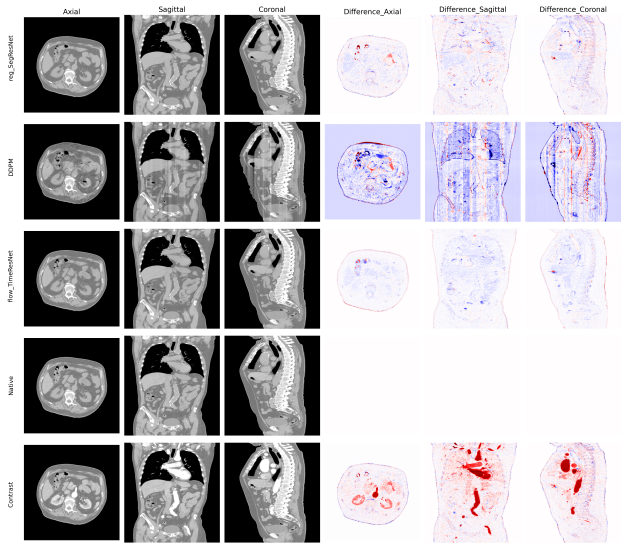

Figure 9: Compare different methods and models for translation from Contrast to Native image. The original Contrast and Native are shown below

4.3 Evaluation Metrics

To evaluate image-to-image translation task we used MAE, although it may not reflect the visual component, but provided that the picture looks visually correct, it is an excellent metric for comparing two models.

$$MAE(x, y) = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} |x_{ij} - y_{ij}|. \tag{6}$$

To balance the MAE we also used the SSIM [52]. The SSIM is a commonly used metric for measuring the structural similarity between two images, it is similar to human perception of image quality. It is defined as:

$$SSIM(x, y) = \frac{\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2 \cdot \sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{7}$$

where μ_x, μ_y and σ_x, σ_y are the means and standard deviations of images x and y respectively, σ_{xy} means the covariance of x and y. C_1 is $(k_1 L)^2$ and C_2 is $(k_2 L)$, where $k_1 = 0.01$, $k_2 = 0.01$ and L is the largest pixel of the image x. Also we used the peak signal-to-noise ratio (PSNR) [53] to evaluate the similarity between the original image and the generated image. PSNR measures the ratio between the maximum possible value of the original image and the power of distorting noise:

$$\text{PSNR} = 20 \cdot \log_{10} \left(\frac{\text{MAX}(image)}{\sqrt{\text{MSE}}} \right), \tag{8}$$

where $\text{MAX}(x)$ denotes the maximum possible pixel value of the image.

4.4 Quantitative Results

All values of the MAE metric are specified for images whose values range from -1000 to 1000 HU, which corresponds to the values of real images. The SSIM values are calculated by translating values from 0 to 1, the PSNR metric is obtained taking into account that the values take values from 0 to 2000, and the maximum value is 2000. All models were compared by translation based on a one-step solution using the Euler method.

Name	**Test**			**Hold-out**			**Time**	**Params**
	MAE↓	SSIM↑	PSNR↑	MAE↓	SSIM↑	PSNR↑	Seconds↓	Millions ↓
UNet	8.132	0.976	37.629	7.051	0.979	39.220	**0.015**	104.3
SegResNet	**7.177**	**0.979**	**38.103**	**6.146**	**0.983**	**40.108**	**0.056**	214.1
SwinUNETR	7.465	0.978	38.046	6.485	<u>0.982</u>	39.865	0.087	120.1
UMambaBot	<u>7.220</u>	**0.979**	**38.103**	<u>6.201</u>	<u>0.982</u>	<u>39.955</u>	<u>0.077</u>	141.6

Table 1: Comparison of neural network regression performance on test and hold-out datasets using MAE, SSIM, PSNR, and inference time in seconds. Best values in **bold**, second-best values are <u>underlined</u>.

The regression-based method produces good metrics for MAE, but you can see that SSIM is significantly inferior to the FlowMatching method. This loss can also be observed visually, the image is blurred.

Split	MAE↓	SSIM ↑	PSNR ↑	Time (seconds) ↓
Test	17.822	0.935	30.192	44.2
Hold-out	18.203	0.934	30.051	44.2

Table 2: Evaluate DDPM performance using MAE, SSIM, and PSNR on test and hold-out datasets.

The diffusion generation method produces good metrics, but it lags far behind other approaches.

For the task of translating Contrast to Native, the best metrics are provided by the SwinUNETR architecture. To compare different architectures, we used the simplest method of solving the equation - the one-step Euler method. We conducted a more thorough analysis in terms of selecting a solver and choosing the number of steps, using the TimeResNet example, you can find the results in B. Careful selection of the solver leads to a significant improvement in the metrics of our proposed model compared to the values of the architecture metrics in the table.

Name	Test			Hold-out			Time	Params
	MAE↓	SSIM↑	PSNR↑	MAE↓	SSIM↑	PSNR↑	Seconds↓	Millions ↓
UMambaBot	<u>6.207</u>	**0.992**	38.623	**6.171**	**0.992**	38.545	<u>0.077</u>	141.6
SegResNet	6.326	0.991	<u>38.783</u>	6.494	0.991	<u>38.572</u>	**0.056**	214.1
SwinUNETR	**6.235**	**0.992**	**38.942**	<u>6.229</u>	**0.992**	**38.933**	0.087	**120.1**
TimeResNet (Our)	6.513	0.990	38.017	6.363	0.991	38.182	0.105	<u>124.7</u>

Table 3: Comparison of neural network Flow Matching performance (from Contrast to Native) using MAE, SSIM, PSNR, and inference time in seconds on test and hold-out datasets. Best values in **bold**, second-best values are <u>underlined</u>.

4.5 Results on subtask segmentation

In order to test the possibility of training real models on our generated images, we conducted the following experiment. We took a hold-out dataset that was not used in training the model, marked the aorta on Native images, and marked the aorta on contrasting images. Next, the contrasting images were converted to native images using the proposed model in the article. We used the nnUNetV2 framework to train the segmentation model. We trained on all 20 images, in order to validate our models, we marked the aorta on an additional 20 native images, on which our models failed. Dice of the model that was trained on real Native images $= 0.966$, Dice of the model that was trained on translated Native images $= 0.926$. These results confirm the huge opportunity for training models on the generated ones, they show metrics comparable with real images.

4.6 Discussion and Limitations

Our proposed method and models yield substantial improvements in CT image translation—thereby expanding the available training data and markedly enhancing performance across diverse domains—they also serve effectively as a data-augmentation tool for training robust models capable of operating in dual-domain settings. It is imperative to define the limits of applicability and to caution that radiologists should not draw clinical conclusions based solely on a single series of images synthesized by these models, whether generating contrast-enhanced scans from native data or vice versa. Our future work will focus on addressing challenges associated with extending these techniques to volumetric imaging and on reducing the requisite GPU memory footprint.

5 Conclusion

In this paper, we propose a novel method in medical imaging that allows us to solve two tasks at once, translation from Contrast to Native and translation from Native to Contrast, which reduces the necessary computational resources to train a model for these tasks, and also demonstrates an unprecedented generation rate compared to diffusion models. We offer an architecture that reliably solves these problems and generate stable images. These developments are expected to enhance the potential accessibility of data across diverse modalities for downstream tasks, as well as to support additional augmentation of medical datasets for foundation models.

6 Data availability

Due to confidentiality, data collected for the study are not publicly available for download, however the corresponding authors can be contacted for academic purposes. Tools for deep learning are indicated in the methods section.

References

[1] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[2] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961, 2019.

[3] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[4] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410, 2016.

[5] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[6] Karin Dembrower, Alessio Crippa, Eugenia Colón, Martin Eklund, and Fredrik Strand. Artificial intelligence for breast cancer detection in screening mammography in sweden: a prospective, population-based, paired-reader, non-inferiority study. *The Lancet Digital Health*, 5(10):e703–e711, 2023.

[7] Rebecca Smith-Bindman, Philip W. Chu, Hana Azman Firdaus, Carly Stewart, Matthew Malekhedayat, Susan Alber, Wesley E. Bolch, Malini Mahendra, Amy Berrington de González, and Diana L. Miglioretti. Projected lifetime cancer risks from current computed tomography imaging. *JAMA Internal Medicine*, 04 2025. ISSN 2168-6106. doi: 10.1001/jamainternmed.2025.0505. URL `https://doi.org/10.1001/jamainternmed.2025.0505`.

[8] Mayur K Virarkar, Sai Swarupa R Vulasala, Anjali Verma Gupta, DheerajReddy Gopireddy, Sindhu Kumar, Mauricio Hernandez, Chandana Lall, and Priya Bhosale. Virtual non-contrast imaging in the abdomen and the pelvis: An overview. *Seminars in Ultrasound, CT and MRI*, 43(4):293–310, 2022. ISSN 0887-2171. doi: https://doi.org/10.1053/j.sult.2022.03.004. URL `https://www.sciencedirect.com/science/article/pii/S0887217122000312`. Dual source CT: Applications, Technology.

[9] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[10] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001. doi: 10.1109/34.927467.

[11] Berengère Aubert-Broche, Mark Griffin, G Bruce Pike, Alan C Evans, and D Louis Collins. Twenty new digital brain phantoms for creation of validation image data bases. *IEEE transactions on medical imaging*, 25(11):1410–1416, 2006.

[12] Marcel Prastawa, Elizabeth Bullitt, and Guido Gerig. Simulation of brain tumors in mr images for evaluation of segmentation efficacy. *Medical Image Analysis*, 13(2):297–311, 2009. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2008.11.002. URL `https://www.sciencedirect.com/science/article/pii/S1361841508001357`. Includes Special Section on Functional Imaging and Modelling of the Heart.

[13] W Paul Segars, G Sturgeon, S Mendonca, Jason Grimes, and Benjamin MW Tsui. 4d xcat phantom for multimodality imaging research. *Medical physics*, 37(9):4902–4915, 2010.

[14] Sébastien Jan, Giovanni Santin, Daniel Strul, Steven Staelens, K Assié, Damien Autret, Stéphane Avner, Remi Barbier, Manuel Bardiès, PM Bloomfield, et al. Gate: a simulation toolkit for pet and spect. *Physics in Medicine & Biology*, 49(19):4543, 2004.

[15] Shengye Hu, Baiying Lei, Shuqiang Wang, Yong Wang, Zhiguang Feng, and Yanyan Shen. Bidirectional mapping generative adversarial networks for brain mr to pet synthesis. *IEEE Transactions on Medical Imaging*, 41(1):145–157, 2021.

[16] Wen Wei, Emilie Poirion, Benedetta Bodini, Stanley Durrleman, Nicholas Ayache, Bruno Stankoff, and Olivier Colliot. Predicting pet-derived demyelination from multimodal mri using sketcher-refiner adversarial training for multiple sclerosis. *Medical Image Analysis*, 58: 101546, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101546. URL https://www.sciencedirect.com/science/article/pii/S1361841519300817.

[17] Haoyu Lan, the Alzheimer Disease Neuroimaging Initiative, Arthur W. Toga, and Farshid Sepehrband. Three-dimensional self-attention conditional gan with spectral normalization for multimodal neuroimaging synthesis. *Magnetic Resonance in Medicine*, 86(3):1718–1733, 2021. doi: https://doi.org/10.1002/mrm.28819. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.28819.

[18] Wanyun Lin, Weiming Lin, Gang Chen, Hejun Zhang, Qinquan Gao, Yechong Huang, Tong Tong, and Min Du. Bidirectional mapping of brain mri and pet with 3d reversible gan for the diagnosis of alzheimer's disease. *Frontiers in Neuroscience*, 15, 2021. URL https://api.semanticscholar.org/CorpusID:233186986.

[19] Apoorva Sikka, Jitender Singh Virk, Deepti R Bathula, et al. Mri to pet cross-modality translation using globally and locally aware gan (gla-gan) for multi-modal diagnosis of alzheimer's disease. *arXiv preprint arXiv:2108.02160*, 2021.

[20] Jin Zhang, Xiaohai He, Linbo Qing, Feng Gao, and Bin Wang. Bpgan: Brain pet synthesis from mri using generative adversarial network for multi-modal alzheimer's disease diagnosis. *Computer Methods and Programs in Biomedicine*, 217:106676, 2022. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2022.106676. URL https://www.sciencedirect.com/science/article/pii/S016926072200061X.

[21] Farideh Bazangani, Frédéric J. P. Richard, Badih Ghattas, and Eric Guedj. Fdg-pet to t1 weighted mri translation with 3d elicit generative adversarial network (e-gan). *Sensors*, 22(12), 2022. ISSN 1424-8220. doi: 10.3390/s22124640. URL https://www.mdpi.com/1424-8220/22/12/4640.

[22] Yan Wang, Yanmei Luo, Chen Zu, Bo Zhan, Zhengyang Jiao, Xi Wu, Jiliu Zhou, Dinggang Shen, and Luping Zhou. 3d multi-modality transformer-gan for high-quality pet reconstruction. *Medical Image Analysis*, 91:102983, 2024. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102983. URL https://www.sciencedirect.com/science/article/pii/S1361841523002438.

[23] Cheng Peng, Pengfei Guo, S Kevin Zhou, Vishal M Patel, and Rama Chellappa. Towards performant and reliable undersampled mr reconstruction via diffusion model sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 623–633. Springer, 2022.

[24] Yutong Xie and Quanzheng Li. Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 655–664. Springer, 2022.

[25] Chi-Hieu Pham, Carlos Tor-Díez, Hélène Meunier, Nathalie Bednarek, Ronan Fablet, Nicolas Passat, and François Rousseau. Multiscale brain mri super-resolution using deep 3d convolutional networks. *Computerized Medical Imaging and Graphics*, 77:101647, 2019.

[26] Cycle Learning Ensemble. Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[28] Alicia Durrer, Julia Wolleb, Florentin Bieder, Tim Sinnecker, Matthias Weigel, Robin Sandkühler, Cristina Granziera, Özgür Yaldizli, and Philippe C Cattin. Diffusion models for contrast harmonization of magnetic resonance images. *arXiv preprint arXiv:2303.08189*, 2023.

[29] Robert Graf, Joachim Schmitt, Sarah Schlaeger, Hendrik Kristian Möller, Vasiliki Sideri-Lampretsa, Anjany Sekuboyina, Sandro Manuel Krieg, Benedikt Wiestler, Bjoern Menze, Daniel Rueckert, et al. Denoising diffusion-based mri to ct image translation enables automated spinal segmentation. *European Radiology Experimental*, 7(1):70, 2023.

[30] Shaoyan Pan, Zach Eidex, Mojtaba Safari, Richard Qiu, and Xiaofeng Yang. Cycle-guided denoising diffusion probability model for 3d cross-modality mri synthesis. In *Medical Imaging 2025: Clinical and Biomedical Imaging*, volume 13410, pages 515–522. SPIE, 2025.

[31] Lingting Zhu, Zeyue Xue, Zhenchao Jin, Xian Liu, Jingzhen He, Ziwei Liu, and Lequan Yu. Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 592–601. Springer, 2023.

[32] Reza Kalantar, Sumeet Hindocha, Benjamin Hunter, Bhupinder Sharma, Nasir Khan, Dow-Mu Koh, Merina Ahmed, Eric Aboagye, Richard Lee, and Matthew Blackledge. Non-contrast ct synthesis using patch-based cycle-consistent generative adversarial network (cycle-gan) for radiomics and deep learning in the era of covid-19. *Scientific Reports*, 13, 06 2023. doi: 10.1038/s41598-023-36712-1.

[33] Jonghun Kim and Hyunjin Park. Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7604–7613, 2024.

[34] Lingting Zhu, Noel Codella, Dongdong Chen, Zhenchao Jin, Lu Yuan, and Lequan Yu. Generative enhancement for 3d medical images. *arXiv preprint arXiv:2403.12852*, 2024.

[35] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[36] Shaoyan Pan, Elham Abouei, Jacob Wynne, Chih-Wei Chang, Tonghe Wang, Richard LJ Qiu, Yuheng Li, Junbo Peng, Justin Roper, Pretesh Patel, et al. Synthetic ct generation from mri using 3d transformer-based denoising diffusion model. *Medical Physics*, 51(4):2538–2548, 2024.

[37] Yitong Li, Igor Yakushev, Dennis M Hedderich, and Christian Wachinger. Pasta: P athology-a ware mri to pet cro s s-modal t r a nslation with diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 529–540. Springer, 2024.

[38] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical ai for synthetic imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4430–4441. IEEE, 2025.

[39] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

[40] Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural odes. *Advances in neural information processing systems*, 34:21325–21337, 2021.

[41] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[42] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

[43] Milad Yazdani, Yasamin Medghalchi, Pooria Ashrafian, Ilker Hacihaliloglu, and Dena Shahriari. Flow matching for medical image synthesis: Bridging the gap between speed and quality. *arXiv preprint arXiv:2503.00266*, 2025.

[44] Hadrien Reynaud, Alberto Gomez, Paul Leeson, Qingjie Meng, and Bernhard Kainz. Echoflow: A foundation model for cardiac ultrasound image and video generation. *arXiv preprint arXiv:2503.22357*, 2025.

[45] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, 97:103285, 2024.

[46] Jungye Kim, Jimin Lee, Bitbyeol Kim, Sangwook Kim, Hyeongmin Jin, and Seongmoon Jung. Generation of deep learning based virtual non-contrast ct using dual-layer dual-energy ct and its application to planning ct for radiotherapy. *PLOS ONE*, 19, 12 2024. doi: 10.1371/journal.pone.0316099.

[47] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[48] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[49] Baosong Yang, Longyue Wang, Derek F. Wong, Lidia S. Chao, and Zhaopeng Tu. Convolutional self-attention networks. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4040–4045, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1407. URL https://aclanthology.org/N19-1407/.

[50] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[51] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[52] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

[53] Jari Korhonen and Junyong You. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth international workshop on quality of multimedia experience*, pages 37–38. IEEE, 2012.

[54] L. Euler. *Institutionum calculi integralis*. Impensis Academiae Imperialis Scientiarum, 1769. URL https://books.google.ru/books?id=XmsCrgEACAAJ.

[55] Mark Lotkin. A note on the midpoint method of integration. *J. ACM*, 3(3):208–211, July 1956. ISSN 0004-5411. doi: 10.1145/320831.320840. URL https://doi.org/10.1145/320831.320840.

[56] C. Runge. Ueber die numerische auflösung von differentialgleichungen. *Mathematische Annalen*, 46:167–178, 1895. URL http://eudml.org/doc/157756.

[57] W. Kutta. *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen*. Teubner, 1901. URL https://books.google.ru/books?id=K5e6kQEACAAJ.

[58] Francis Bashforth and John Couch Adams. *An Attempt to Test the Theories of Capillary Action by Comparing the Theoretical and Measured Forms of Drops of Fluid*. Cambridge University Press, Cambridge, 1883. URL https://archive.org/details/attempttest00bashrich.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

The additional material is organized as follows: section A will contain details about metrics and methods of model inference on a test sample, section B will provide details about the selection of hyperparameters in training, section C will provide visual diagrams of training and operation of various parts of the neural network, and section D will provide additional examples of image generation.

A Additional quantative results

A.1 Details

We measured the ability to convert images with contrast to native images using the models we trained to test their ability to do this. They showed good results. This gives us the opportunity to use the same neural network, but for two types of tasks, it allows us to reduce the time for training two neural networks in the case of UNet Regression, DDPM.

Name	**Test**			**Hold-out**			**Time (s)**	**Params**
	MAE↓	SSIM↑	PSNR↑	MAE↓	SSIM↑	PSNR↑	Seconds↓	Millions ↓
UMambaBot	<u>6.996</u>	**0.989**	<u>39.095</u>	<u>6.882</u>	<u>0.989</u>	<u>38.933</u>	<u>0.077</u>	141.1
SegResNet	**6.431**	**0.989**	**39.544**	**6.426**	**0.990**	**39.369**	**0.056**	214.1
SwinUNETR	7.593	0.987	38.692	7.604	0.987	38.524	0.087	**120.1**
TimeResNet (Our)	7.329	0.988	38.947	7.253	0.988	38.777	0.105	<u>124.7</u>

Table 4: Comparison of neural network Flow Matching performance (from Native to Contrast) using MAE, SSIM, PSNR, and inference time in seconds on test and hold-out datasets. Best values in **bold**, second-best values are <u>underlined</u>.

In the task of translating from Native to Contrast, the SegResNet architecture wins by a large margin.

A.2 ODE-Solver Ablation

Table A.2 benchmarks five numerical solvers (Euler [54], Midpoint [55], RK2 [56], RK3, RK4 [57], Adams [58]) Exponential Integrator) for the best FM model over step budgets $\{1, 3, 5, 10\}$.

We can numerically solve the neural ordinary differential equation using various methods to figure out which method works best and what is the optimal number of steps to choose. With the proper selection of a solver to solve the problem, our model wins by a significant margin over other architectures.

Solver	**Test**			**Hold-out**			**Time**	**Complexity**
	MAE↓	SSIM↑	PSNR↑	MAE↓	SSIM↑	PSNR↑	Seconds ↓	↓
Euler (1)	6.513	0.990	38.017	6.362	0.991	38.182	**0.105**	1
Euler (2)	5.591	0.995	39.039	5.525	0.995	39.017	<u>0.210</u>	2
Euler (3)	<u>5.529</u>	**0.996**	38.941	<u>5.472</u>	0.995	38.994	0.313	3
Euler (5)	5.565	0.995	38.899	5.526	0.995	38.951	0.521	5
RK2 (1)	**5.399**	**0.996**	**40.017**	**5.436**	**0.996**	**39.776**	0.209	2
RK3 (1)	5.553	0.995	39.193	5.553	0.995	39.122	0.313	3
RK4 (1)	5.674	0.995	<u>39.408</u>	5.716	0.995	<u>39.257</u>	0.419	4
Mid Point (1)	5.601	0.995	39.106	5.611	0.995	39.047	<u>0.210</u>	2
Mid Point (3)	5.744	0.995	38.882	5.759	0.995	38.923	0.628	6

Table 5: Comparison of different solvers (from Contrast to Native) for TimeResNet (our) using MAE, SSIM, PSNR, and inference time in seconds on test and hold-out datasets. Best values in **bold**, second-best values are <u>underlined</u>.

By complexity here, we mean the number of calls to the neural network to calculate the gradient when generating a single image.

B Additional implementation details

B.1 2D Processing Justification

Due to the limited computing resources available, we decided to solve this problem in 2D, since 3D requires much more computing resources and a lot of data, but a full-fledged volume does not fit on one GPU, so it would require a latent-space solution, which still demands significant resources. Moreover, clinicians typically review studies in the axial plane (though not always), making 2D axial slices a natural choice. We therefore restrict our method to 2D axial processing.

B.2 Image Preprocessing and Resizing

When down-sampling inputs, one can either resample to a uniform spacing or to a fixed pixel size. We chose a fixed image size of 512×512, since nearly all DICOM slices in our dataset already conform to this dimension. Avoiding on-the-fly resizing during inference eliminates an extra source of error. Additionally, variations in original spacing act as implicit data augmentation, improving model robustness. We acknowledge that 2D methods may lack inter-slice context and can yield inconsistent predictions across different planes (e.g., axial vs. sagittal), as seen in diffusion experiments; however, our approach demonstrates stable and consistent contrast-to-native predictions slice-by-slice within the same study.

B.3 Data Normalization and Clipping

To harmonize input intensities across scanners, we clip CT Hounsfield Units to $[-1000, 1000]$, then linearly rescale to $[-1, 1]$. This range covers the vast majority of soft-tissue contrasts while ensuring numerical stability during training.

B.4 Data Augmentation and Optimization

We apply geometric augmentations via the MONAI framework:

- Random flips along x and y axes (50% each),
- Random 90° rotations (up to 3 turns, 50% probability),
- Random affine transforms (70% probability) with: rotation up to $\pm 45°$, translation up to ± 102.4 pixels, shear up to $\pm 10°$.

We train using Adam with hyperparameters: learning rate $= 2 \cdot 10^{-5}$, weight decay $= 0$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. Batch size was 2.

B.5 Training Data Selection and Dual-Energy Exclusion

To minimize misregistration due to patient motion between acquisitions, we remove series where we have different origins and spacing between two series. Although dual-energy CT could in principle provide "native" images directly, such data are scarce, and the vendor-reconstructed "native" series result from proprietary algorithms that may not faithfully represent true tissue contrast. Therefore, we exclude dual-energy studies to ensure the network learns genuine native-to-contrast mappings.

B.6 Image Registration

In order to improve the quality and purity of the data, we registered images with contrast to native images. We employ the ANTs library with the "deformable SyN only" transform, providing a body-mask to focus optimization on the anatomy. This configuration yielded the best mean absolute error (MAE) and mutual information (MI) improvements, and was faster than alternative ANTs transforms.

C Additional visual results and discussion

The generation examples are presented in 3 projections, and each of the projections shows the difference between the original native image. The picture was clipped into the abdominal window for

a clearer understanding of the difference, since in other windows or without an image clip at all, the difference is much less noticeable, and we are primarily interested in the abdominal window.

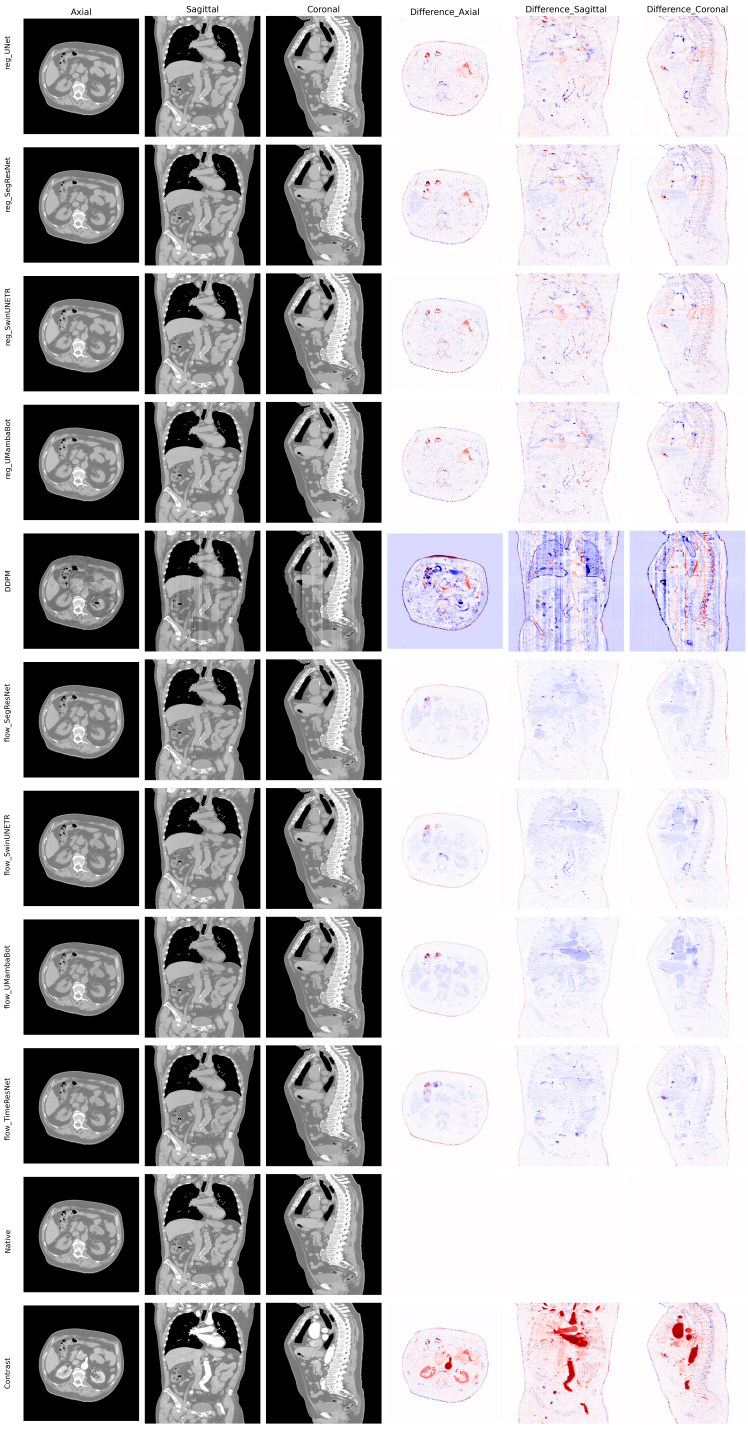

Figure 10: Compare different methods and models for translation from Contrast to Native image. The original Contrast and Native are shown below

D Architecture details

D.1 Time embedding mechanism

One nontrivial challenge is how to encode temporal information so that the model fully captures context and can accurately predict the vector field between two images at a given time point. We evaluated several strategies and found the following approach to be most effective, yielding substantial performance gains. Because the model tends to "forget" the temporal signal, we inject temporal embeddings into every block of the encoder. In addition, we introduce a lightweight linear projection layer in each block, enabling it to learn the specific dependencies it requires without unduly increasing the overall parameter count. This linear layer operates analogously to an instance-normalization layer, allowing independent modulation of each channel's activations. The temporal embeddings themselves are constructed using the sinusoidal scheme described in [59]. To capture nonlinear temporal interactions, we further process these sinusoidal embeddings through a small multilayer perceptron, and the resulting feature vector is added to the input of each ResNet block. In the schematic below A, B denote trainable vectors of length C, which are applied element-wise to the output of the SiLU activation. For convenience, one linear layer of length $2 \cdot C$ is trained, which is then divided into two vectors – A, B. If the output of the nonlinear layer, h, has dimensions (B x C x H x W), then the resulting tensor likewise has dimensions (B x C x H x W), with the multiplication broadcast along the channel dimension.

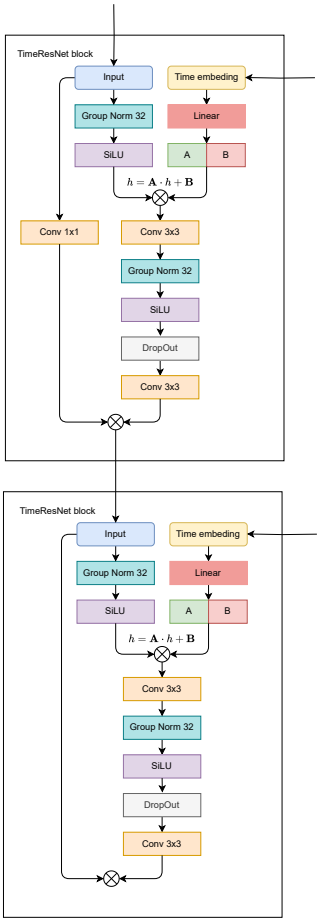

Figure 11: Time embedding mechanism.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of the work in Section 4.6

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All details of the proposed method are included in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All the code and model weights will be released upon paper acceptance, data will be available on request.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All details included in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: NA

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All the experiments are conducted on 4 RTX 3090 24G GPUs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We contribute translation model and method, which can benefit numerous clinical study and applications. All limitations discussed in the Section 4.6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: No such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The trained model will be released after reviewing.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.