

Unsupervised Skull Segmentation via Contrastive MR-to-CT Modality Translation

Kamil Kwarciak¹, Mateusz Daniol¹, Daria Hemmerling¹, and Marek Wodzinski^{1,2}

¹ Department of Measurement and Electronics, AGH University of Krakow, Kraków, Poland

² Institute of Informatics, University of Applied Sciences Western Switzerland (HES-SO Valais), Sierre, Switzerland

Abstract. The skull segmentation from CT scans can be seen as an already solved problem. However, in MR this task has a significantly greater complexity due to the presence of soft tissues rather than bones. Capturing the bone structures from MR images of the head, where the main visualization objective is the brain, is very demanding. The attempts that make use of skull stripping seem to not be well suited for this task and fail to work in many cases. On the other hand, supervised approaches require costly and time-consuming skull annotations. To overcome the difficulties we propose a fully unsupervised approach, where we do not perform the segmentation directly on MR images, but we rather perform a synthetic CT data generation via MR-to-CT translation and perform the segmentation there. We claim that translating the process to the CT modality is essential, as it significantly simplifies the overall procedure by transforming the complex segmentation in MR into a more straightforward segmentation in CT. We address many issues associated with unsupervised skull segmentation including the unpaired nature of MR and CT datasets (contrastive learning), low resolution and poor quality (super-resolution), and generalization capabilities. We demonstrate the effectiveness of our methodology through a quantitative analysis using Dice and Surface Dice metrics on the validation dataset, as well as on the test set to highlight its adaptability to new datasets. The research has a significant value for downstream tasks requiring skull segmentation from MR volumes such as craniectomy or surgery planning and can be seen as an important step towards the utilization of synthetic data in medical imaging.

Keywords: Deep Learning · Generative AI · Modality Translation · Contrastive Learning · Super-Resolution · Synthetic Data · CT Synthesis

1 Introduction

Skull segmentation is a common objective of many medical imaging studies, as it is usually the first step in processing pipelines and plays a crucial role in various diagnostic solutions. Among them, it is safe to mention: dental planning,

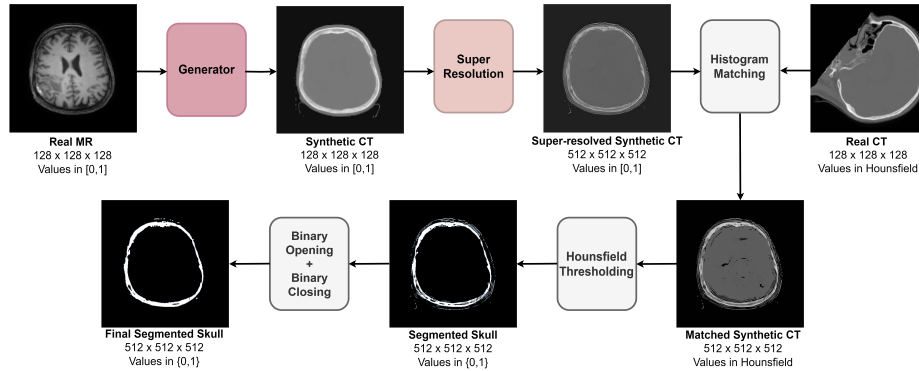


Fig. 1: Inference with the use of the proposed solution. Firstly, the generator from the CUT framework is applied. This is followed by a super-resolution module. Next, we perform histogram matching using a different CT sample from the dataset, which does not need to be correlated with the input MR image. After this, Hounsfield thresholding is applied, followed by binary operations, to achieve the final skull segmentation results. The entire process leverages synthetic CT for skull segmentation via MR-to-CT modality translation.

patient positioning, or cranial implants design [39]. On one hand, this process in computed tomography (CT) may seem straightforward, as it commonly utilizes the thresholding in the Hounsfield scale [33], while on the other hand, when it comes to magnetic resonance (MR) modality, this procedure gets more complicated. The reason behind this is that the main strength of MR lies in providing detailed images of soft tissues rather than bone structures. Hence, the problem becomes highly complex, but also indispensable, as many challenges require working on the skull (or in general bone) data, while only MR images are available or the acquisition of CT is undesired. As an example, one can mention the case of pediatric imaging where MR is favored over CT to avoid of ionizing radiation [2].

Among the techniques used for the separation of skull and brain structures, skull stripping [7, 12, 35] seem to be standard solutions. Skull stripping is a technique that involves removing non-brain tissues, such as the skull and extracranial structures. This is typically achieved through a wide range of different methods such as classical morphological operations, or novel deep neural network architectures. Among the most widely used skull stripping methods, we find highlight Bet [35] which is the most classic approach and relies on deformable surface models, DeepBet [7] which can be seen as a deep learning-based extension of Bet and benefits from modifying the U-Net [32] and applying bias correction and normalization, and finally SynthStrip [12] which uses a 3-D U-Net and is trained on synthetic data. Then, to obtain a skull segmentation mask, one, can perform a subtraction between the input MR image and output of skull stripping and apply binary masking on top of it. However, it should be kept in mind that skull strip-

ping methods are not directly dedicated to skull segmentation, and they may produce some artifacts. Various supervised segmentation models have been proposed in the literature to address different medical tasks. One notable solution is the SwinUNETR [10], suited for tasks such as brain tumor segmentation from MR images and lung tumor segmentation from CT images. However, the application of SwinUNETR for skull segmentation has not been explored, primarily due to the lack of annotated datasets that include MR images of the head along with corresponding skull segmentation masks. Additionally, supervised segmentation methods trained on specific tasks often exhibit poor generalization capabilities, struggling not only with different anatomical structures but also with the same anatomical structures represented in different datasets. Finally, an approach that can be seen as a potential go-to solution is the utilization of foundation models, specifically Segment Anything Model [18] and to be precise, its medical variant, MedSAM [21]. Due to the robust MedSAM training routine which also involves training on a large-scale dataset that can handle diverse segmentation tasks, segmenting the skull from an MR image should also be possible. Finally, the skull segmentation models could be trained in a supervised manner, however, such an approach would require a significant number of annotated volumes from various medical centers and acquisition protocols.

The concept of modality translation has already emerged in the field of generative AI, as a very wide range of models and architectures have been developed. The general concept of image-to-image translation is to find a mapping between images from the source and target domains. The popular techniques include GAN-based approaches [13, 41], contrastive learning [25], or diffusion models [34]. Many of these solutions are applied in the medical image analysis to translate between modalities, where the most important one from the perspective of this work is MR-to-CT [24, 40]. Existing solutions in this field yield satisfying outcomes. Nevertheless, their limitation lies in their inability to generalize effectively due to a fact of training on relatively homogeneous datasets. This causes a challenge for skull segmentation, especially in the context of clinical treatments, where data sources vary considerably. To take the matter further we find two major bottlenecks of contemporary strategies. Firstly, many approaches use solely 2-D data [20, 40] rather than fully adopting a 3-D perspective, which is computationally cheaper, however limits inference performance in real-life scenarios with volumetric data, particularly when considering information between volumes. Secondly, various models are trained in the paired manner [24] which means that there is a direct label and alignment between images from both domains in the dataset, and in most of the clinical cases, the data is available only in one modality. One more concept that should also be mentioned in terms of working on MR-to-CT translation is the resolution of these imaging techniques. Most commonly, MR data has a lower resolution than CT data, and high-resolution CT is desired in surgical procedures, such as craniectomy. Hence, in translation works it is important to further target obtained synthetic CT images and upsample them into higher-resolution, with the use of classic interpolation techniques or more robust, super-resolution networks.

Contribution: In this study, we aim to explore the emerging field of synthetic data generation for medical imaging, specifically focusing on the application of modality translation in cranioplasty to achieve unsupervised skull segmentation in MR images. We solve the previously mentioned dilemmas by designing a pipeline that enables the training on highly diverse datasets of MR and CT images in a fully volumetric and unpaired manner with the use of contrastive learning. We additionally provide a super-resolution module to further address opportunities in clinical use cases, and finally, we experiment with downstream tasks, such as the generation of synthetic CT skulls of children, and modality translation on defective skulls. We compare the proposed approach to state-of-the-art skull stripping methods with postprocessing and medical segmentation foundation model, MedSAM. Importantly, our model does not require any annotations during training. We would also like to emphasize the difference between skull segmentation and skull stripping. The primary goal of skull stripping is to separate the brain from non-brain elements, including the skull. If the final output needs to include the skull, additional postprocessing steps will be required. In contrast, the goal of skull segmentation is to specifically extract the cranial bone, making non-bone structures unimportant. The tasks are entirely different and face distinct challenges.

2 Methods

2.1 Pipeline Overview

The two main components of our solution are the Contrastive Unpaired Translation (CUT) [25] module and the Laplacian Pyramid Super-Resolution Network (LapSRN) [19] module, both utilized for 3-D volumetric data processing. They work separately and use different techniques for sampling the data from the whole space of gathered datasets. CUT is responsible for the MR-to-CT translation in an unpaired manner, it takes unpaired samples and tries to generate a synthetic CT image for the corresponding MR image without a vision of a direct mapping. LapSRN is trained directly and only on high-resolution CT images to create a pyramidal structure of increasing resolutions, with the usage of extreme data augmentation (which we discuss further) to enhance the generalization capabilities of the model. The full architecture is presented in Figure 2 and the inference process is presented in Figure 1.

2.2 Preprocessing

For MR data we simply perform min-max normalization for every instance separately to obtain values in the range of $[0,1]$ with preservation of the local dynamics. For CT data, we perform Hounsfield rescaling to remove the structures that have values below $-500 HU$ (air/background), as it enhances the training process, and has been found to be the most efficient for the whole dataset. After that, we also apply min-max scaling to values in the $[0,1]$ range.

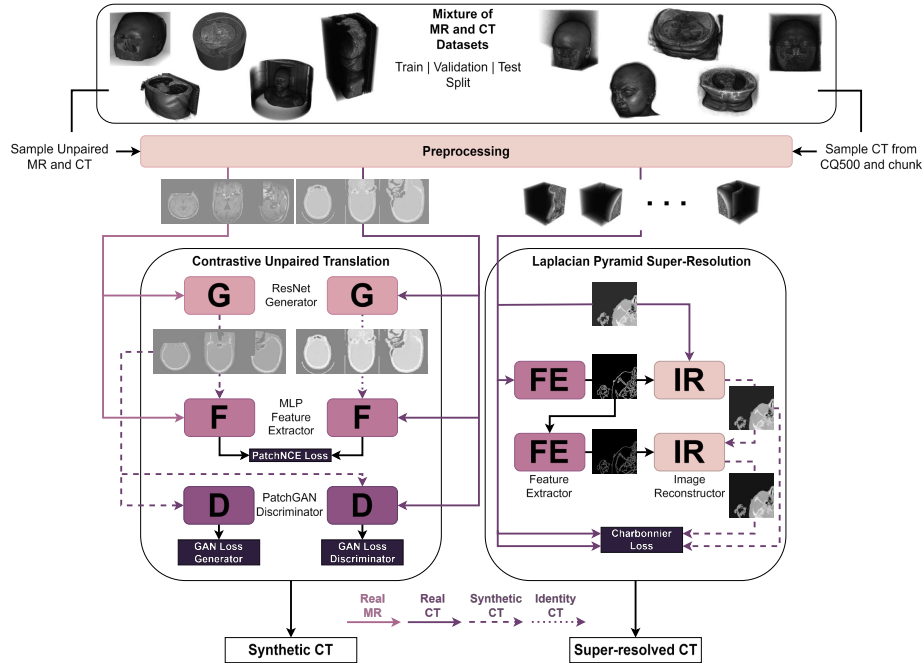


Fig. 2: The overview of the pipeline for the Contrastive Unpaired Translation and Laplacian Pyramid Super-Resolution Network. Note that we work on 3-D tensors, the 2-D representations are used only for visualization simplicity.

2.3 Contrastive Unpaired Translation

The CUT component of the pipeline is heavily inspired by the original work [25], however, we add several updates to address the domain specificity. The CUT framework consists of three networks, generator G , discriminator D , and feature extractor F , where each of them is implemented to operate on 3-D data representations. G in the context of this work is a ResNet-like encoder-decoder network, D is a PatchGAN discriminator [13], and F is a shallow multilayer perceptron. In general, G and D follow a classic GAN minimax game, where G aims to produce realistic CT data to fool a D network, while the D aims to distinguish between real and generated CT data:

$$\begin{aligned} \mathcal{L}_{\text{GAN}}(D, G, \text{MR}, \text{CT}) = & \mathbb{E}_{\mathbf{y}_{\text{real}}^{\text{CT}} \sim \text{CT}} [\log D(\mathbf{y}_{\text{real}}^{\text{CT}})] \\ & + \mathbb{E}_{\mathbf{x}_{\text{real}}^{\text{MR}} \sim \text{MR}} [\log (1 - D(G(\mathbf{x}_{\text{real}}^{\text{MR}})))] , \end{aligned} \quad (1)$$

where G as input takes a real MR image $\mathbf{x}_{\text{real}}^{\text{MR}}$ and produces synthetic CT image $\mathbf{y}_{\text{syn}}^{\text{CT}}$ by learning a mapping $G : \mathbf{x}_{\text{real}}^{\text{MR}} \rightarrow \mathbf{y}_{\text{syn}}^{\text{CT}}$. The second component of the CUT training objective is the use of noise contrastive estimation (NCE) [22].

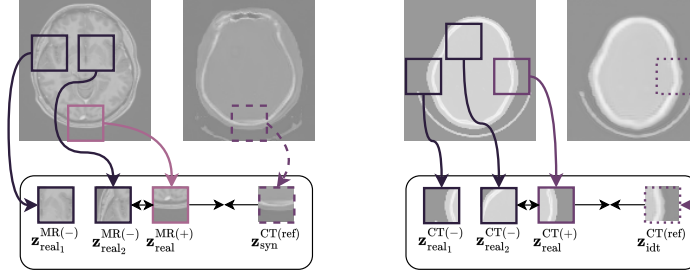


Fig. 3: Sampling procedure for patchwise contrastive estimation of real MR \leftrightarrow synthetic CT, and real CT \leftrightarrow identity CT. We show a 2-D view for better visualization.

Specifically, it employs the InfoNCE loss to maximize mutual information by distinguishing the positive sample from a set of unrelated noise samples. However, an important aspect of the CUT framework, in comparison to the original InfoNCE work, is that the comparison is not performed between a positive sample and noise samples. Instead, it is done by comparing patches sampled from different spatial locations. The source MR image $\mathbf{x}_{\text{real}}^{\text{MR}}$, together with its corresponding synthetic CT image $\mathbf{y}_{\text{syn}}^{\text{CT}}$, are passed through the encoding component of the generator G_{enc} , followed by the feature extractor F , to generate latent feature vector representations in a shared embedding space. Next, a reference patch $\mathbf{z}_{\text{syn}}^{\text{CT}(\text{ref})}$ is extracted from the intermediate representation of $\mathbf{y}_{\text{syn}}^{\text{CT}}$ and compared to a positive patch $\mathbf{z}_{\text{real}}^{\text{MR}(+)}$ extracted from the intermediate representation of $\mathbf{x}_{\text{real}}^{\text{MR}}$ at the same spatial location, as well as to $N - 1$ negative samples $\mathbf{z}_{\text{real}}^{\text{MR}(-)}$ also taken from the intermediate representation of $\mathbf{x}_{\text{real}}^{\text{MR}}$, but taken from different spatial locations. This setup yields the following formulation of InfoNCE loss, with the dot product used as a similarity measure:

$$\begin{aligned} \ell_{\text{MR}} &= \mathcal{L}_{\text{InfoNCE}|\text{MR}_{\text{real}}, \text{CT}_{\text{syn}}}(\mathbf{z}_{\text{syn}}^{\text{CT}(\text{ref})}, \mathbf{z}_{\text{real}}^{\text{MR}(+)}, \mathbf{z}_{\text{real}}^{\text{MR}(-)}) \\ &= -\log \frac{\exp(\mathbf{z}_{\text{syn}}^{\text{CT}(\text{ref})} \cdot \mathbf{z}_{\text{real}}^{\text{MR}(+)})}{\exp(\mathbf{z}_{\text{syn}}^{\text{CT}(\text{ref})} \cdot \mathbf{z}_{\text{real}}^{\text{MR}(+)}) + \sum_{j=1}^{N-1} \exp(\mathbf{z}_{\text{syn}}^{\text{CT}(\text{ref})} \cdot \mathbf{z}_{\text{real}}^{\text{MR}(-)})}. \end{aligned} \quad (2)$$

The CUT methodology also incorporates the concept of identity mapping and utilizes InfoNCE in this setup to stabilize training, functioning similarly to an identity loss. These identity CT mappings are obtained via $G : \mathbf{y}_{\text{real}}^{\text{CT}} \rightarrow \mathbf{y}_{\text{idt}}^{\text{CT}} \approx \mathbf{y}_{\text{real}}^{\text{CT}}$. The latent representation generation process is applied in a similar manner: the data is passed through the encoding part of the generator G_{enc} and then through the feature extractor F . This process results in the following set of patches: $\mathbf{z}_{\text{idt}}^{\text{CT}(\text{ref})}$, which is a reference patch from identity CT, and $\mathbf{z}_{\text{real}}^{\text{CT}(+)}$, $\mathbf{z}_{\text{real}}^{\text{CT}(-)}$ being positive and negative samples from real CT, this provides a second

InfoNCE formulation as:

$$\begin{aligned} \ell_{\text{CT}} &= \mathcal{L}_{\text{InfoNCE}|\text{CT}_{\text{real}}, \text{CT}_{\text{idt}}}(\mathbf{z}_{\text{idt}}^{\text{CT(ref)}}, \mathbf{z}_{\text{real}}^{\text{CT(+)}} , \mathbf{z}_{\text{real}}^{\text{CT(-)}}) \\ &= -\log \frac{\exp(\mathbf{z}_{\text{idt}}^{\text{CT(ref)}} \cdot \mathbf{z}_{\text{real}}^{\text{CT(+)}})}{\exp(\mathbf{z}_{\text{idt}}^{\text{CT(ref)}} \cdot \mathbf{z}_{\text{real}}^{\text{CT(+)}}) + \sum_{j=1}^{N-1} \exp(\mathbf{z}_{\text{idt}}^{\text{CT(ref)}} \cdot \mathbf{z}_{\text{real}}^{\text{CT(-)}})}. \end{aligned} \quad (3)$$

The idea of comparison behind Equation 2 and 3 is presented in Figure 3. Moreover, the CUT methodology broadens the mentioned InfoNCE loss components to the intermediate outputs of G_{enc} layers. This ensures that the features learned at multiple levels are aligned and collectively enhance the quality and consistency of the final generated output. If we denote the layers of the encoding part of G (G_{enc}) as L , then at the l -th layer, we obtain a set of features for the real MR and real CT images as follows: $\{\mathbf{z}_{\text{real}(l)}^{\text{MR}}\}_L = \{F_l(G_{\text{enc}}^l(\mathbf{x}_{\text{real}}^{\text{MR}}))\}_L$ and $\{\mathbf{z}_{\text{real}(l)}^{\text{CT}}\}_L = \{F_l(G_{\text{enc}}^l(\mathbf{y}_{\text{real}}^{\text{CT}}))\}_L$. If we also address the synthetic and identity representations we obtain the following sets: $\{\mathbf{z}_{\text{syn}(l)}^{\text{CT}}\}_L = \{F_l(G_{\text{enc}}^l(G(\mathbf{x}_{\text{real}}^{\text{MR}})))\}_L$ and $\{\mathbf{z}_{\text{idt}(l)}^{\text{CT}}\}_L = \{F_l(G_{\text{enc}}^l(G(\mathbf{y}_{\text{real}}^{\text{CT}})))\}_L$. Now, as we take S_l patches at layer l we have a set of S/s negative patches (-) and s positive patch (+), and we are able to formulate PatchNCE loss in two setups, the first one as the main training objective of synthesizing CT images from MR inputs:

$$\begin{aligned} \mathcal{L}_{\text{PatchNCE}}(G, F, \text{MR}) \\ = \mathbb{E}_{\mathbf{x}_{\text{real}}^{\text{MR}} \sim \text{MR}} \sum_{l=1}^L \sum_{s=1}^{S_l} \ell_{\text{MR}}(\mathbf{z}_{\text{syn}(l)}^{\text{CT(ref)}(s)}, \mathbf{z}_{\text{real}(l)}^{\text{MR}(+) (s)}, \mathbf{z}_{\text{real}(l)}^{\text{MR}(-) (S/s)}), \end{aligned} \quad (4)$$

and the second being a stabilization term between real CT and its identity CT (real CT passed through the generator):

$$\begin{aligned} \mathcal{L}_{\text{PatchNCE}}(G, F, \text{CT}) \\ = \mathbb{E}_{\mathbf{y}_{\text{real}}^{\text{CT}} \sim \text{CT}} \sum_{l=1}^L \sum_{s=1}^{S_l} \ell_{\text{CT}}(\mathbf{z}_{\text{idt}(l)}^{\text{CT(ref)}(s)}, \mathbf{z}_{\text{real}(l)}^{\text{CT}(+) (s)}, \mathbf{z}_{\text{real}(l)}^{\text{CT}(-) (S/s)}). \end{aligned} \quad (5)$$

To conclude, for CUT in terms of MR-to-CT translation, PatchNCE loss is computed in two manners, firstly by comparing $\mathbf{x}_{\text{real}}^{\text{MR}}$ with $\mathbf{y}_{\text{syn}}^{\text{CT}}$ as a main training objective, and secondly by comparing $\mathbf{y}_{\text{real}}^{\text{CT}}$ with $\mathbf{y}_{\text{idt}}^{\text{CT}}$ for training stabilization and regularization. Finally, we can formulate the final objective of CUT training as:

$$\begin{aligned} \mathcal{L}_{\text{CUT}} &= \lambda_{\text{GAN}} \mathcal{L}_{\text{GAN}}(G, D, \text{MR}, \text{CT}) \\ &\quad + \lambda_{\text{syn}} \mathcal{L}_{\text{PatchNCE}}(G, F, \text{MR}) + \lambda_{\text{idt}} \mathcal{L}_{\text{PatchNCE}}(G, F, \text{CT}) \end{aligned} \quad (6)$$

2.4 Laplacian Pyramid Super-Resolution

The second component of the designed pipeline is the modification of the original LapSRN [19] used for super-resolution of 3-D CT volumes. It uses two sub-networks: feature extractor and image reconstructor, which progressively generate images at higher resolutions from the low-resolution ones. The feature extractor is responsible for extracting the features at a coarse (lower) level and generating feature maps at a finer (higher) level. It enhances the representation of the input by capturing important details. The image reconstructor upsamples the lower-resolution input and then via element-wise summation with residuals obtained from the feature extractor creates a higher-resolution output with improved visual quality. By its nature, progressive reconstruction provides task-dependent flexibility and adjustability, as by bypassing the pyramid at certain levels we can obtain representations at the resolution required for a given task. Furthermore, to address the requirement of high generalization capabilities, our LapSRN uses not only a set of diverse data augmentations (flipping, affine transformations, motion artifacts, blurriness, contrast), but is also trained on overlapping chunks of input tensors (see Figure 2). This configuration enhances the model’s generalization capabilities, enabling it to effectively super-resolve single chunks and capture inconsistencies in various skulls, such as those with defects. Additionally, it significantly reduces memory requirements during training, as the chunks can be processed individually or combined into small micro-batches within the mini-batch. LapSRN uses Charbonnier loss, where for M chunks of the input tensor, and pyramid of L levels, the Charbonnier loss is defined as:

$$\mathcal{L}_{\text{Charbonnier}}(\mathbf{y}, \mathbf{x}, \mathbf{r}) = \frac{1}{M} \sum_{m=1}^M \sum_{s=1}^L \sqrt{(\mathbf{y}_s - \mathbf{x}_s - \mathbf{r}_s)^2 + \epsilon^2}, \quad (7)$$

where for layer s , chunk residual is denoted as \mathbf{r}_s , upsampled lower-resolution chunk as \mathbf{x}_s and ground truth higher-resolution chunk \mathbf{y}_s and ϵ (set to small value like 10^{-3}) controls similarity to L_1 loss while staying differentiable.

2.5 Postprocessing

Synthetic CT and high-resolution CT obtained from pipeline components are in the $[0, 1]$ range and it is desired to have them in the Hounsfield scale, to perform threshold-based skull segmentation. This can be easily achieved by performing a histogram matching with one of the real CTs from the train set for CUT, and low-resolution CT for LapSRN. Furthermore, it was experimentally found that following Hounsfield-based thresholding with binary opening and binary closing is beneficial for artifacts removal. It is also important to note, that training both modules, CUT and LapSRN on the original Hounsfield scale can yield training instabilities due to the wide range of values and potential outliers, which can lead to gradient issues and slow convergence.

3 Experiments

3.1 Datasets

To address the generalization possibilities we create a huge dataset of 1,521 MR and 879 CT 3-D images from publicly available datasets [1, 3–6, 8, 9, 14–16, 23, 27–31, 36, 37] (See Supplementary materials for more information). Importantly, we extract a portion of the dataset from the SynthRAD 2023 Challenge [37] for validation purposes (25 images), as it is the only dataset that includes paired MR and CT samples, hence it can be used for evaluation with metrics like Dice coefficient. We also utilize the Han-Seg Dataset [28] for testing purposes (42 images), as it includes images of heads in both MR and CT formats. However, it is important to note that the images in this dataset are not registered. Therefore, we perform image registration before using this dataset for testing. The Han-Seg dataset demonstrates the generalization capabilities of the proposed method. To fully address the requirements of generalization, we construct the dataset of skull segmentation by including also other anatomical structures, such as the pelvis, aorta, or kidneys. This setup enables a more robust capturing of style rather than context, as commonly different institutions use various imaging devices, making single-institution models hard to generalize. To address this, we combined datasets with images of different anatomical structures to enhance cross-institution generalizability via different anatomical structures and different acquisition settings. We train CUT with the use of all datasets, and for LapSRN we only use a dataset of high-resolution CT skulls, namely CQ500 [4] (300 images for training and 50 images for testing).

3.2 Networks

We implement the proposed networks with the use of PyTorch library [26]. CUT and LapSRN subnetworks use 3-D convolutional layers. What’s important, for the CUT model we use instance normalization [38] as it prevents instance-specific mean and covariance shifts, hence it is highly beneficial for modality translation tasks. CUT’s generator is a 9-layer ResNet-based network, with residual connections between downsampling and upsampling blocks, it uses instance normalization and ReLU nonlinearity besides the decision layer for which we use hyperbolic tangent. The discriminator is a 3-layer convolutional network with Leaky ReLU and instance normalization (like PatchGAN discriminator [13]). CUT’s feature extractor is a simple 2-layer multilayer perceptron with ReLU which operates on 64 patches of features extracted from the generator’s flow. Hence, regarding Equation 4 and Equation 5, we operate on a fixed amount of patches equal to 64, and a fixed amount of G_{enc} layers equal to 9. LapSRN’s image reconstructor is a 2-layer convolution/deconvolution network and the feature extractor is an 8-layer convolution/deconvolution network with Leaky ReLU nonlinearities. LapSRN’s convolutional layers use $3 \times 3 \times 3$ kernels with 64 filters with He initialization [11], deconvolutions use the kernel of $4 \times 4 \times 4$ (upsampling by a factor of 2) and weights are initialized from a trilinear filter, as suggested in the original implementation [19].

Table 1: Comparison of methods with DSC and SDSC, evaluated on the subset of SynthRAD 2023 training dataset [37], extracted for a test set in this work.

Method	DSC \uparrow	SDSC \uparrow
SynthStrip [12] + subtraction	0.281	0.580
Bet [35] + subtraction	0.273	0.563
DeepBet [7] + subtraction	0.298	0.606
MedSAM [21]	–	–
Ours	0.512	0.770

3.3 Setup

We resize the data to $128 \times 128 \times 128$ for training both the CUT and LapSRN models. We train CUT using ADAM optimizer [17] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We set the initial learning rate to $2 \cdot 10^{-4}$ and decrease it linearly every 50 epochs if the loss doesn’t decrease. We set λ_{GAN} , λ_{syn} and λ_{idt} all equal to 1 and train CUT with the real batch size of 1 (batch size of 8 distributed across 8 GPUs). LapSRN is trained using SGD with momentum term set to 0.9 and weight decay of 10^{-4} . The learning rate is initialized to 10^{-5} and decreases linearly every 5 epochs if the loss doesn’t decrease. The chunk size for splitting the inputs is set to 8 with 8 voxels overlapping ($128^3 \rightarrow 8 \times 64(+8)^3$), the batch size is 1 and we accumulate gradients every 16 batches. Models were trained until convergence, with the use of 8 NVIDIA A100 40GB GPUs.

4 Results

We evaluate the proposed method on the subset of samples extracted from the training dataset from SynthRAD 2023 Challenge [37] as it consists of paired MR and CT images. We present the Dice coefficient (DSC) and surface Dice coefficient (SDSC) results, calculated from segmentation masks obtained using Hounsfield scale thresholding in Table 1. Importantly, it should be noted that, as previously mentioned, skull stripping differs from skull segmentation. The results of skull stripping methods were followed by postprocessing, which involved manual thresholding for skull subtraction. Furthermore, we demonstrate that the medical imaging foundation model, MedSAM [21], fails to perform the complex task of skull segmentation from MR images. The failed results are presented in Figure 4a for the bounding box approach and Figure 4b for the experimental point prompt approach. We investigate the quality of generated synthetic CT via a small set of ablation studies which are presented in Figure 5. They showcase the following settings: (i) the performance of LapSRN using a test sample from the CUT module, (ii) generalization capabilities to the downstream task of translation on the MR image of the child’s skull, and (iii) the translation of the MR image of a defected (and reconstructed) skull. Finally, to demonstrate the generalization potential of the proposed solution from a quantitative perspective,

we investigate additionally the MR subset of the HaN-Seg dataset [28] which was not present in the original training dataset. As the availability of paired MR and CT datasets is highly limited, besides SynthRAD 2023 dataset [37], we find HaN-Seg to be the most suitable option for this evaluation. Importantly, additional preprocessing in the form of image registration between CT and MR samples in this dataset was required to enable the quantitative evaluation. For comparison, we decided to train a state-of-the-art segmentation network, SwinUNETR [10], on the SynthRAD 2023 dataset [37], where the input was an MR image and the target was a skull mask derived from Hounsfield thresholding of its paired CT image. We also trained the model as a standard paired MR-to-CT translation and applied the same Hounsfield thresholding on the resulting synthetic CTs for consistency with the pipeline from Figure 1. Table 2 shows that the proposed solution presents superior results in terms of the generalization capabilities in comparison to both SwinUNETR setups.

5 Discussion

Quantitative analysis of the CUT method shows that it can provide better results in terms of DSC and SDSC than the combination of skull stripping with subtraction and masking as shown in Table 1. Importantly, we find out that MedSAM [21] struggles with the skull segmentation in MR images and fails to propagate the segmentation mask in both bounding-box and point-prompt (experimental) approaches. This can be potentially solved via task-specific fine-tuning (as the authors mention this MedSAM capability in their work). Good results of our method suggest that it is a promising direction for designing a fully unsupervised framework of skull segmentation from MR images, that can be used for downstream tasks (such as defect segmentation), and with the use of super-resolution LapSRN module, the requirement of high-resolution CT images for tasks like craniectomy or surgery planning is also met. What’s more, both CUT and LapSRN are relatively simple in terms of architectural design and also their learning objectives, hence they can be adapted to other translation tasks, or fine-tuned for other anatomical structures. With CUT we were able to obtain good generalization capabilities due to the very high diversity of the used dataset. Importantly, we also note several limitations of our methodology and leave it as a potential direction for further work. First of all, produced results sometimes include some blurriness; this can be attributed to using convolutional networks with a relatively small number of filters and a shallow network depth. This issue

Table 2: Evaluation of the generalization capabilities on the HaN-Seg dataset [28] in terms of DSC and SDSC.

Method	DSC \uparrow	SDSC \uparrow
SwinUNETR (MR-Mask)	0.084	0.135
SwinUNETR (MR-CT)	0.113	0.149
Ours	0.310	0.396

stems from the challenges of designing large networks for high-resolution volumetric data. Even the latest GPUs struggle with VRAM capacities when trying to fit large tensors and extensive 3-D networks. Nevertheless, we are able to partially reduce this issue as the whole pipeline (presented in Figure 1) employs postprocessing in the form of morphological operations which diminishes this effect in the task of segmentation. Secondly, while DSC results are better than other methods, they still require improvement to meet the demands of potential real-world medical applications. Finally, the primary motivation of this work was to achieve generalization capabilities for the segmentation task in an unsupervised and unpaired manner. Therefore, we do not compare our results with other methods used in MR-to-CT translation, as these are not primarily focused on segmentation. Nevertheless, we acknowledge that some existing translation methods may be more effective than CUT, and other super-resolution modules could outperform LapSRN. We consider this a potential area for further improvement, research, and investigation.

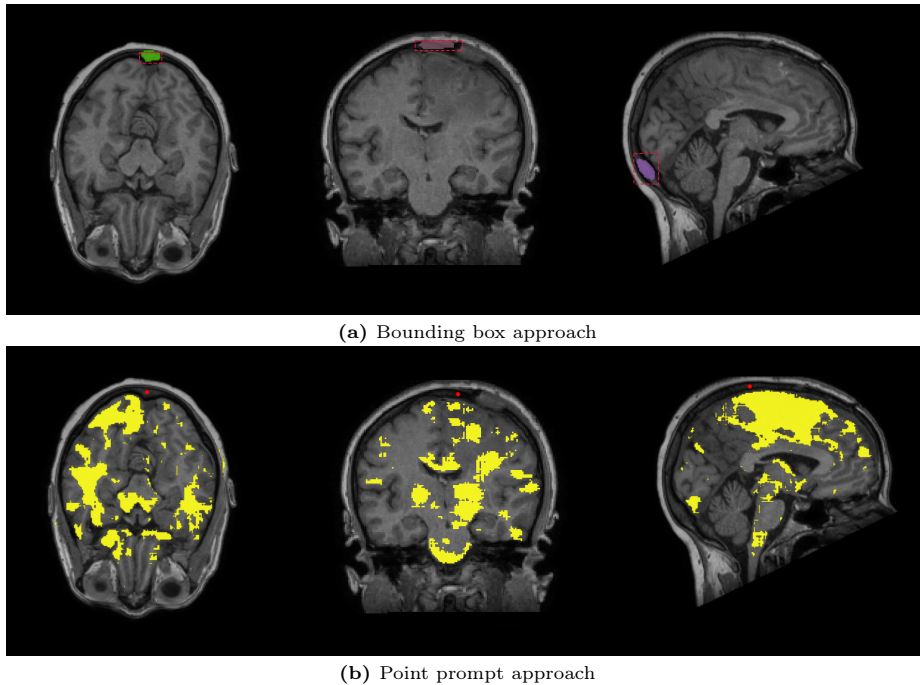


Fig. 4: Results of skull segmentation from MR images with the use of MedSAM: (a) bounding box and (b) point prompt approaches. Bounding box approach failure stems from a fact that segmented skull structure is not propagated through the whole image, and only small parts are captured. For point prompt, the model is unable to identify the skull and propagates segmentation into brain.

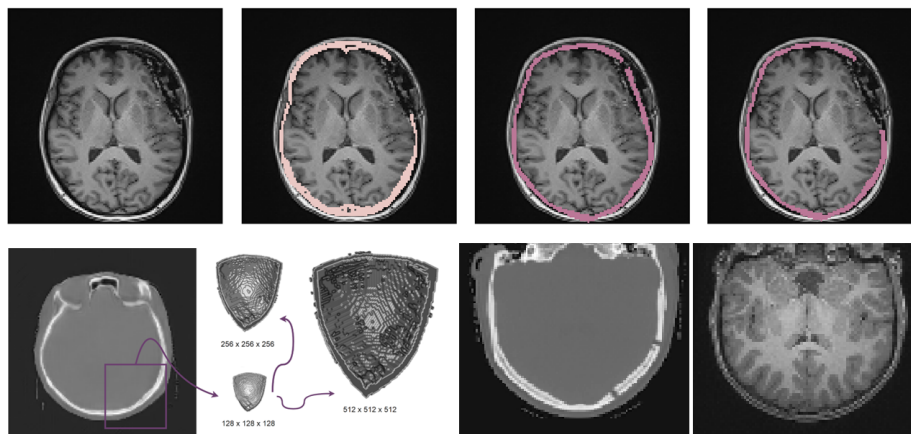


Fig. 5: Top: Results of translation and segmentation on defected skulls (from left to right: input MR, matched CT mask, synthetic CT mask, synthetic CT mask with removed implant area). Bottom left: Super-resolution of synthetic skull. Bottom right: MR-to-CT translation on child’s skull.

6 Conclusion

In conclusion, we presented a novel generative AI-based methodology for synthetic CT generation, specifically designed for skull segmentation from MR images. Our approach operates on 3-D data, is trained in an unsupervised manner, and demonstrates strong generalization capabilities while also super-resolving into higher resolutions. We analyze its performance utilizing quantitative metrics on two separate datasets to present its performance and generalization. We believe this opens up new research opportunities in this field, and we plan to further enhance the proposed solution.

Acknowledgements

The project was funded by The National Centre for Research and Development, Poland under Lider Grant no: LIDER13/0038/2022 (DeepImplant). We gratefully acknowledge Polish HPC infrastructure PLGrid support within computational grants no. PLG/2023/016239 and PLG/2024/017079.

References

1. Banfi, C., et al.: Reading-related functional activity in children with isolated spelling deficits and dyslexia. *Language, Cognition And Neuroscience* **36**(5), 543–561 (2021)

2. Bosch de Basea Gomez, M., et al.: Risk of hematological malignancies from ct radiation exposure in children, adolescents and young adults. *Nature Medicine* **29**(12), 3111–3119 (2023)
3. Burleigh, L., Greening, S.G.: Fear in the mind’s eye: the neural correlates of differential fear acquisition to imagined conditioned stimuli. *Social Cognitive and Affective Neuroscience* **18**(1), nsac063 (2023)
4. Chilamkurthy, S., et al.: Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study. *The Lancet* **392**(10162), 2388–2396 (2018)
5. Etzel, J.A., et al.: The dual mechanisms of cognitive control dataset, a theoretically-guided within-subject task fmri battery. *Scientific data* **9**(1), 114 (2022)
6. Fialkowski, K.P., Bush, K.A.: Identifying the neural correlates of resting state affect processing dynamics. *Frontiers in Neuroimaging* **1**, 825105 (2022)
7. Fisch, L., et al.: Deepbet: Fast brain extraction of t1-weighted mri using convolutional neural networks. *arXiv preprint arXiv:2308.07003* (2023)
8. Gaesser, B., et al.: A role for the medial temporal lobe subsystem in guiding prosociality: the effect of episodic processes on willingness to help others. *Social cognitive and affective neuroscience* **14**(4), 397–410 (2019)
9. Greene, D.J., et al.: Behavioral interventions for reducing head motion during mri scans in children. *Neuroimage* **171**, 234–245 (2018)
10. Hatamizadeh, A., et al.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: *International MICCAI Brainlesion Workshop*. pp. 272–284. Springer (2021)
11. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE international conference on computer vision*. pp. 1026–1034 (2015)
12. Hoopes, A., et al.: Synthstrip: Skull-stripping for any brain image. *NeuroImage* **260**, 119474 (2022)
13. Isola, P., et al.: Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1125–1134 (2017)
14. Jo, H., et al.: A brain network that supports consensus-seeking and conflict-resolving of college couples’ shopping interaction. *Scientific reports* **10**(1), 17601 (2020)
15. Kavur, A.E., et al.: Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis* **69**, 101950 (2021)
16. Keane, B.P., et al.: Brain network mechanisms of visual shape completion. *NeuroImage* **236**, 118069 (2021)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
18. Kirillov, A., et al.: Segment anything. *arXiv preprint arXiv:2304.02643* (2023)
19. Lai, W.S., et al.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 624–632 (2017)
20. Liu, Y., et al.: Ct synthesis from mri using multi-cycle gan for head-and-neck radiation therapy. *Computerized medical imaging and graphics* **91**, 101953 (2021)
21. Ma, J., et al.: Segment anything in medical images. *Nature Communications* **15**(1), 654 (2024)
22. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018)

23. Ottesen, T.D., et al.: Differences in chemo-signaling compound-evoked brain activity in male and female young adults: A pilot study in the role of sexual dimorphism in olfactory chemo-signaling. *bioRxiv* pp. 2020–09 (2020)
24. Pan, S., et al.: Synthetic ct generation from mri using 3d transformer-based denoising diffusion model. *arXiv preprint arXiv:2305.19467* (2023)
25. Park, T., et al.: Contrastive learning for unpaired image-to-image translation. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX* 16. pp. 319–345. Springer (2020)
26. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019)
27. Peelle, J.E., et al.: Increased connectivity among sensory and motor regions during visual and audiovisual speech perception. *Journal of neuroscience* **42**(3), 435–442 (2022)
28. Podobnik, G., et al.: Han-seg: The head and neck organ-at-risk ct and mr segmentation dataset. *Medical physics* **50**(3), 1917–1927 (2023)
29. Racey, C., et al.: An open science mri database of over 100 synaesthetic brains and accompanying deep phenotypic information. *Scientific Data* **10**(1), 766 (2023)
30. Radl, L., et al.: Avt: Multicenter aortic vessel tree cta dataset collection with ground truth segmentation masks. *Data in brief* **40**, 107801 (2022)
31. Rogers, C.S., et al.: Real-time feedback reduces participant motion during task-based fmri. *bioRxiv* (2023)
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18. pp. 234–241. Springer (2015)
33. Rulaningtyas, R., et al.: Ct scan image segmentation based on hounsfield unit values using otsu thresholding method. In: *Journal of Physics: Conference Series*. vol. 1816, p. 012080. IOP Publishing (2021)
34. Sasaki, H., Willcocks, C.G., Breckon, T.P.: Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358* (2021)
35. Smith, S.M.: Fast robust automated brain extraction. *Human brain mapping* **17**(3), 143–155 (2002)
36. Thornton, M.A., et al.: People represent their own mental states more distinctly than those of others. *Nature communications* **10**(1), 2117 (2019)
37. Thummerer, A., et al.: Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy. *Medical Physics* (2023)
38. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016)
39. Wodzinski, M., Daniol, M., Hemmerling, D.: Improving the automatic cranial implant design in cranioplasty by linking different datasets. In: *Towards the Automation of Cranial Implant Design in Cranioplasty II: Second Challenge, AutoImplant 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 2*. pp. 29–44. Springer (2021)
40. Wolterink, J.M., et al.: Deep mr to ct synthesis using unpaired data. In: *Simulation and Synthesis in Medical Imaging: Second International Workshop, SASHIMI 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 10, 2017, Proceedings 2*. pp. 14–23. Springer (2017)
41. Zhu, J.Y., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2223–2232 (2017)