

Unsupervised Tumor Segmentation

Mehdi Astaraki^{1,2}, Francesca De Benetti³, Yousef Yeganeh³, Iuliana Toma-Dasu^{2,4}, Örjan Smedby¹, Chunliang Wang¹, Nassir Navab^{3,5}, Thomas Wendler³

¹KTH Royal Institute of Technology, Department of Biomedical Engineering and Healthy Systems, SE-14157 Huddinge, Sweden

²Karolinska Institutet, Department of Oncology-Pathology, SE-17176 Stockholm, Sweden

³Chair for Computer Aided Medical Procedures and Augmented Reality, Technische Universität München, Boltzmannstr. 3, 85748 Garching bei München, Germany

⁴Stockholm University, Department of Physics, SE-106 91 Stockholm, Sweden

⁵Chair for Computer Aided Medical Procedures Laboratory for Computational Sensing and Robotics, Johns-Hopkins University, Baltimore, MD, USA

Abstract

To be completed!

1. Introduction

Medical image segmentation refers to the process of partitioning the voxels/pixels of tissues, organs, or pathologies from background anatomical structures in medical images such as Computed Tomography (CT), Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI). Target region segmentation in medical images is recognized as one of the most challenging tasks in medical image analysis due to the complexity of human anatomy, lack of intensity/textural contrast between adjacent tissues, presence of noise/artifacts, and boundary missing [1]. This process is often done manually by expert radiologists, which is not only a demanding task but also subjects to inter/intra-observer variabilities [2]. However, the quantifications derived from the segmentation step deliver critical information regarding the characteristics of the segmented regions such as shape, area/volume, and intensity/textural distributions that can be further used for diagnosis, prognosis, and interventional purposes. In the context of oncological images, the aim of image segmentation is to delineate the boundaries of target tumoral regions [3] and/or nearby healthy organs known as organs at risk [4].

In the past three decades, a variety of computerized methods have been developed to speed up the delineation time without compromising the segmentation accuracy. In a broad view, these methods can be categorized as either deep learning or non-deep learning techniques. In the context of non-deep learning techniques, a wide range of rule-based methods have been proposed for different segmentation tasks. Region-growing [5], watershed [6], level-set [7], Markov random fields [8], graph cut [9], atlas-based [10] and statistical shape modelling [11] approaches, are only a few examples of rule-based segmentation methods that were employed to segment different types of tumors such as liver [12], [13], kidney [14], [15], and prostate [16], [17].

Supervised segmentation: Capability and limitations. Thanks to the recent rapid advances in the deep learning fields, a great level of progress have been witnessed in the performance of medical image segmentation tasks. Inspiring by the breakthrough of U-Net model [18], many different techniques have been proposed to tackle a variety of challenging segmentation problems. The novelties introduced by such models are mainly focused on modifications of the network architecture and/or optimization process. In this context, Attention U-Net was proposed by integrating the attention gate [19] into the plain U-Net model to guide the learning process more on the target area that successfully improved the segmentation performance of brain tumors [20] and retinal vessels [21]. By replacing convolutional blocks with inception blocks [22], computationally efficient deeper U-Nets were developed to deal with large variations in size and morphology within the salient regions. The superiority of the segmentation accuracy of such models was reported for the challenging task of lung nodule detection [23]. Similarly, Dense U-Net and Residual U-Net were developed by using Dense blocks [24] and Residual blocks [25], respectively, in the encoder-decoder paths that lead to outstanding segmentation accuracy of the prostate [26] and lung cancer [27]. More powerful segmentation network families such as U-Net⁺⁺ [28] and Adversarial U-Net [29] have been developed and tested on large-scale datasets with remarkable improvement in

segmentation performance in different tasks. Despite the promising potential of such models, which can achieve clinical expert level accuracies, they require a large number of labeled data due to their supervised training fashion. In fact, supervised training of such data greedy models suffers from two types of limitations. First, the number of training medical images is often limited because of the costly slice-by-slice data annotation. Second, even if large-scale training data is available, the generalization power of the learned models over the unseen classes is poor, which necessarily requires the collection of annotated data from the new class followed by retraining of the model [30].

Unsupervised segmentation. Unsupervised deep learning methods tend to be a natural fit for gaining insights into medical image analysis tasks as their optimizations do not entail labeled datasets. In this domain, Unsupervised Anomaly Detection (UAD) is an active field of research that aims to identify the data that does not fit the learned distribution from normal data [31]. The main advantage of UAD approaches is their similarity to the learning procedures of radiologists who are trained to learn the appearance and characteristics of healthy anatomical structures to potentially detect any arbitrary abnormalities without a-priori knowledge of their attributes [32]. This essentially means that the training process of such models requires only unlabeled data acquired from healthy subjects. The underlying hypothesis is to capture the distribution of healthy anatomical organs by training deep representation learning models in order to identify anomalies as outliers with respect to the normative distribution [33]. In the domain of medical image segmentation, the applications of UAD techniques have been extensively investigated for the task of lesion segmentation [34]–[36]. In a series of contributions, Baur et al. investigated the potential of the deep AutoEncoder (AE) models for unsupervised brain lesion segmentation from MR images [32]. Specifically, by integrating the adversarial training into spatial Variational AutoEncoder (VAE), they could map the healthy anatomies into latent manifolds and further reconstruct fairly high-resolution images. With this model, they achieved a segmentation accuracy of 0.605 in terms of Dice score for Multiple Sclerosis (MS) lesion segmentation in a dataset containing 49 subjects [33]. They later developed a SteGANomaly [36] model, which gains from the steganographic abilities of CycleGAN in removing high-frequency patterns that, to some extent, was a beneficial strategy for preventing the learned model from reconstructing the images with pathological regions that achieved the best Dice score, $[Dice]$, of 0.608 on the same MS dataset. The same authors employed the inherent multi-scale nature of the Laplacian pyramid within a family of AE models to compress and reconstruct MR images of different resolutions in a scale-space [37] approach. With this method, they reported a $[Dice]$ value of 0.590 on the same MS dataset. Schlegel et al. [31] built a generative model of healthy training data and used the GAN’s latent space along with an anomaly score to comprise a discriminator feature residual error and image reconstruction error. The proposed f-AnoGAN model was tested on optical coherence tomography images with superiority over the conventional AE-based models. To efficiently learn fine-grained feature representations, Tian et al. [35] developed a Constrained Contrastive Distribution (CCD) model to simultaneously predict the augmented data as well as image contexts. This model was tested on colonoscopy and fundus screening datasets and outperformed a few other UAD models. Sergio et al. [38] lifted the need for an encoder network to capture the latent representation of healthy data by substituting the AE architecture with an auto-decoder along with a modified version of the implicit field learning technique to reconstruct high-resolution anomaly-free images. This model was tested on a brain tumor segmentation task in MR images with an outstanding performance against a family of VAE models. Last but not least, Dey et al. [39] developed an Adversarial-based Selective Cutting neural network (ASC-net) by integrating the adversarial learning into a U-Net-like model with two decoders to decompose the images into two cuts based on a reference learned distribution of healthy images. The focus of this model is to obtain a joint estimation of anomaly and the corresponding normal images rather than to reconstruct a high-fidelity normal-looking image. This model was tested on several different pathology segmentations, including MS and brain tumor in MR images and liver tumor in CT images, and outperformed the segmentation accuracy of AnoGAN families.

Anomaly detection challenges. Despite the promising results achieved by the current UAD models, such models suffer from a number of limitations: 1) The first issue is related to learning the distribution of healthy anatomies in full image resolution. In fact, there are many fine-grained details in healthy anatomies that pose similar attributes with respect to the pathologies. However, the current methods cannot deal with such anatomical details and are unable to discriminate the fine-grained healthy structures from abnormalities. To tackle this issue, the current methods reduce the dimensionality of the original images to eliminate the fine-grained details and train the models with low-resolution data [32]. Such a downsampling procedure, however, abandons important image characteristics and therefore yields in learning the distributions of incomplete anatomies. 2) They often focus on detecting anomalies with different intensity patterns with

respect to nearby normal tissues, such as glioma and MS lesions, in a specific sequence(s) of MR images. However, to the best of our knowledge, detecting pathologies with similar intensity/textural patterns w.r.t adjacent healthy organs has not been investigated. The fact that AE models often reconstruct a blurry version of the down-sampled original image challenges the underlying hypothesis of capturing the distribution of healthy anatomies [40]. In other words, the hyperintensity patterns of the studied pathologies within generated images from the learned low dimensional representation space naturally tend to be suppressed. Hence, the residual images followed by some thresholding would consist of the anomaly regions regardless of the quality of the generated image. 3) Another important limitation of the current UAD techniques is their difficulties in preserving the anatomical constraint within the generated images. In fact, generating a healthy image from the corresponding pathological image does not necessarily guarantee the retaining of the anatomical constraints of other tissues and structures. Therefore, the residual images calculated from the difference between the original images and the unrealistic-looking generated images often consist of quite many false positives.

Image inpainting. Image inpainting is the process of synthesizing alternative contents in the missing parts of an image with semantically meaningful patterns to reconstruct a seamless and realistic-looking image. It can be used for a variety of image editing tasks such as text removal, object removal, and missing part recovery [41], [42]. Although a variety of CNN-based models have been proposed for image inpainting, typical convolutional operators are naturally unsuitable for hole filling as they treat all the valid and invalid pixels as the same. To tackle this issue, Liu et al. [43] proposed a Partial Convolution (PConv) neural network in which the typical convolution operator is masked and renormalized to be conditioned only on the valid pixels. The invalid pixels are replaced by adjacent textures following a rule-based mask updating procedure. The model was trained with randomly generated irregular masks, and its superior performance was verified on large-scale datasets both quantitatively and qualitatively. In order to condition the prediction of missing pixels at each coordinate on the valid pixels from the input image, Yu et al. [44] replaced the PConv layers with Gated Convolution (GConv) layers along with adding a contextual attention layer and Spectral Normalized Markovian Discriminator (SN-PatchGAN). The advantage of this GConv layer is that they are able to learn features from input images progressively for each channel of the network. The network architecture consists of two encoder-decoder networks named as coarse and refinement networks, followed by a fully convolutional SN-PatchGAN. Due to the learnable dynamic mask updating procedure, Gconv model generates images with more color and texture consistency than Pconv model. Last but not least, different studies show that inpainting models trained with irregular-shaped holes distributed randomly over the image plane can generate images with more semantic context than those trained with simple-shape holes such as rectangles [45], [46].

Contribution. In this study, we propose an inpainting-based UAD method for tumor segmentation in single/multimodal medical images. Specifically, we propose a robust inpainting method to reconstruct high-resolution medical images from corrupted ones while preserving fine-grained details. To efficiently train the inpainting model, healthy images were corrupted by carefully generated random irregular holes to simulate the morphological characteristics of heterogeneous tumors. The learned model is then employed for automatic tumor removal in the test phase in an autoinpainting pipeline. In particular, a set of subregions within the main image is defined through a sliding window approach to be inpainted. The autoinpainting procedure is followed by a postprocessing strategy to detect the candidate region for the final tumor removal. Finally, image slices of each subject are aggregated to form a volume from which residual volumes are calculated to segment the tumors. The proposed inpainting model is optimized with a multi-term objective function to fill the invalid holes with plausible imagery characteristics as well as to preserve the anatomical constraints. The developed pipeline was tested for unsupervised segmentation of two challenging types of tumors: Non-Small Cell Lung Cancer (NSCLC) and Head-and-Neck (HN) on single modalities of CT and PET as well as multimodal PET-CT images.

2. Methods

2.1. Dataset

Two datasets were examined to investigate the potential of the proposed method for segmenting different types of tumors.

PET-CT dataset for Non-Small Cell Lung Cancer (NSCLC) tumor segmentation

This internal dataset includes 33 subjects, all diagnosed with NSCLC in stage III except three subjects who were categorized as stage I, II, and IV. All subjects were scanned with a Biograph 40 PET scanner (Siemen Medical Solution) to acquire one FDG-PET-CT scan before the beginning of radiation therapy and another one after a few weeks of treatment. The acquisition parameters of the scans varied in a wide range. While the voxel spacing in the CT images were fixed to $(0.976 \times 0.976 \times 3)mm^3$, this parameter was fixed to $(4.072 \times 4.072 \times 3)mm^3$ for the corresponding PET images. A semi-automatic segmentation tool based on the level-set algorithm was utilized to generate the grand truth segmentation mask [47]. In specific, initial contours were set around the tumors by an experienced user to instantiate the intensity-based contour evolution algorithm. The final contours were then visually examined and manually refined by an expert radiologist.

PET-CT dataset for Head-and-Neck (HN) tumor segmentation

This multi-institutional publicly available dataset originally consisted of 300 HN cancer patients all diagnosed with squamous cell carcinoma [48], [49]. All patients underwent FDG-PET-CT scans with a median of 18 days before starting the treatment. Tumor delineations were done manually by expert radiation oncologists. In specific, for 93 out of 300 patients, the original radiotherapy contours were directly drawn on the CT images of the PET-CT data, which were used for treatment planning. For the other 207 patients, the radiotherapy contours were delineated on another CT image particularly acquired for treatment planning. The drawn contours were then registered to the original FDG-PET-CT scans using an intensity-based free-form deformable registration tool provided by the software MIM (MIM Software Inc., Cleveland, OH) [48].

From the original 300 subjects, a total number of 298 data were accessible. As annotations were done by various experts from different institutions, large variations were observed within the delineated contours. Accordingly, all the contours were visually examined by an experienced user, which led to excluding 70 subjects. Therefore, from this dataset, 228 PET-CT scans were used for this study in which the CT image resolution varied from $(0.683, 0.683, 2.0)mm^3$ to $(2.343 \times 2.343 \times 1.5)mm^3$ and the PET resolution lies in the range of $(3.515 \times 3.515 \times 3.269)mm^3$ to $(5.468 \times 5.468 \times 3.269)mm^3$.

2.2. Image preparation and preprocessing

The following preprocessing was applied to the employed dataset. First, a third-order Spline interpolation method was used for the PET images to resample the voxel spacing of PET data into the corresponding CT volumes. Second, the intensity values of PET images were converted into Standardized Uptake Values (SUV). Third, to enhance the contrast between the tissues within the target organ, intensity values of CT and PET images were clamped. Particularly, the Hounsfield values of CT images were clamped into the range of $[-1000, 500]$ for NSCLC data and $[-200, 200]$ for the HN dataset. The SUV values of PET images were constrained in the range of $[0, 12]$ as well. The axial slices from signed 16bit volumes were extracted and converted into 8bit gray-level images with the size of 512×512 pixels. Finally, the intensity range of images was normalized by maximum values and rescaled into the range of 0 to 1. Figure 1 shows the diversity of shape, size, and location of the tumors among the employed datasets.

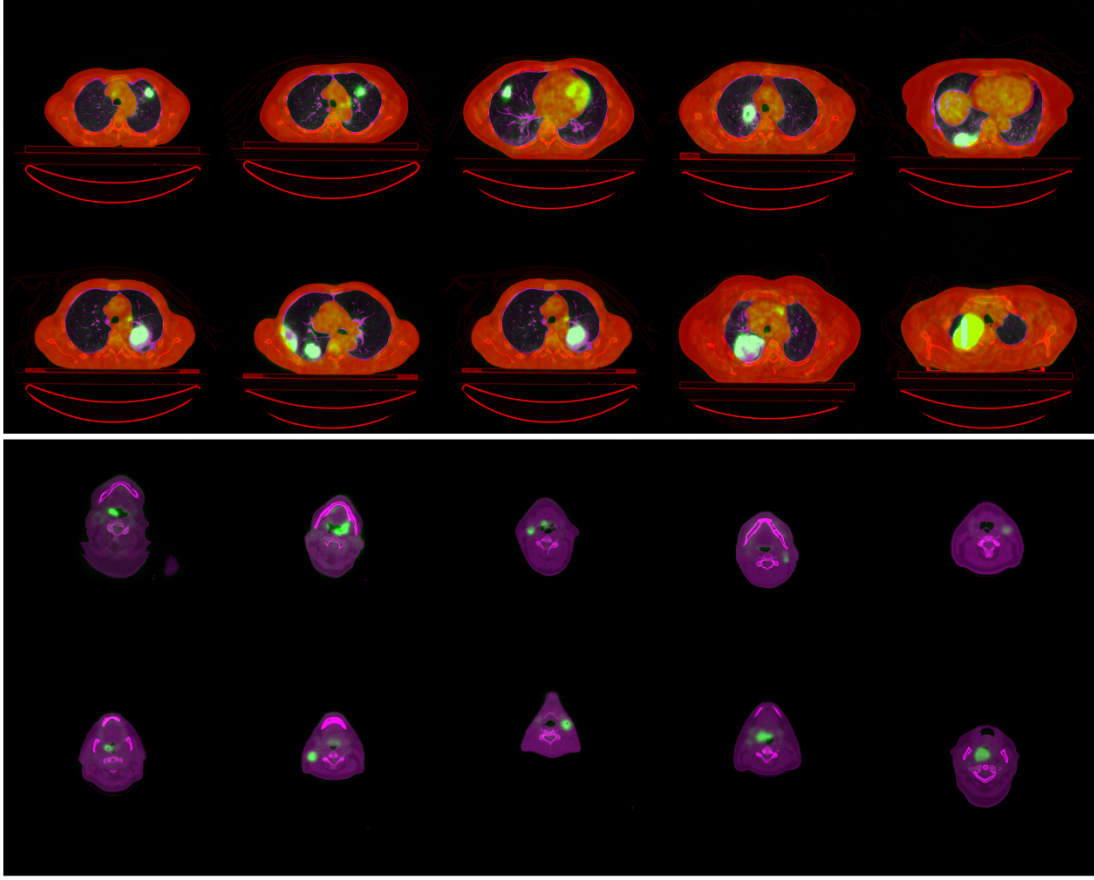


Figure 1. Heterogeneous tumors appear in a diverse range of shapes and sizes at different locations. The first two rows show the diversity of NSCLC tumors, and the second two rows depict different HN tumors.

2.3. Image inpainting model

Assume that $I_{(x,y)}$ stands for a c -channel input image (or input feature map), and W represents a set of filters. The conventional convolutional operator filters the input image and returns a c' -channel output, $O_{(x,y)}$. Mathematically, this function can be represented as:

$$O_{(x,y)} = I_{(x,y)} * W_{(x,y)} = \sum_{i=-k'_m}^{k'_m} \sum_{j=-k'_n}^{k'_n} W_{k_m+i, k_n+j} \cdot I_{x+i, y+j}$$

Where k_m and k_n show the kernel size, $k'_m = \frac{k_m-1}{2}$ and $k'_n = \frac{k_n-1}{2}$. Please note that for the simplicity of the notation, the bias term was skipped. Although this type of convolutional operator works well for several tasks such as image classification, segmentation, and detection, it is not definitely, suitable for the task of image inpainting. In fact, the sliding window scans all the pixels and elements within the image/feature maps and applies the same filters at different spatial coordinates. Thus it simply ignores the presence of holes within a subregion and considers the valid and invalid pixels as the same. As a result, the inpainted holes do not fully match with the nearby textures, and the generated images contain textural/color inconsistencies.

Pconv operator [43] is considered a promising attempt to tackle the mentioned issues faced by the convolutional operators. Let M be a binary mask with the same size as the input image, the partial convolution at every spatial location for the current sliding window can be defined as:

$$O_{(x,y)} = \begin{cases} W_{(x,y)}^T (I_{(x,y)} \odot M_{(x,y)}) \frac{\text{sum}(1)}{\text{sum}(M)} & \text{if } \text{sum}(M) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Where \odot denotes element-wise multiplication, 1 is an all-one matrix with the same size of M . Compared to the ordinary convolution operator, one can understand that the output values of Pconv depend only on the valid areas defined by the binary mask (M). The role of the scaling factor ($\frac{\text{sum}(1)}{\text{sum}(M)}$) is to adjust for varying amounts of valid inputs. After each Pconv operator, the binary mask will be updated by the following rule: if Pconv could condition its output on at least one valid input value, that spatial coordinate will be updated to become valid. However, this kind of rule-based mask updating is problematic because: 1) all feature channels in each convolutional layer share the same mask regardless of their inconsistencies. This limitation will be problematic, especially for multi-channel input images such as multimodal PET-CT slices. 2) The binary mask will be updated progressively as the network goes deeper so that all the invalid pixels will be disappeared no matter how many pixels were covered in the previous layers.

Gconv operator [44] has been proposed to turn the problematic rule-based mask updating of Pconv into a learnable procedure. In specific, gated convolutions learn soft mask updating automatically from the image/feature maps. It will able the convolutional operators to learn the dynamic feature selection mechanism for each channel and each spatial coordinate independently. This process can be formulated as:

$$\begin{aligned} \text{Gating}_{(x,y)} &= \sum \sum W_g \cdot I \\ \text{Feature}_{(x,y)} &= \sum \sum W_f \cdot I \\ O_{(x,y)} &= \varphi(\text{Feature}_{(x,y)}) \odot \sigma(\text{Gating}_{(x,y)}) \end{aligned}$$

Where σ refers to the sigmoid function that scales the output of the gating signal into the range of 0 to 1; φ can be any kind of nonlinear activation function; W_g and W_f are two separate convolutional filters.

Inspired by the concept of the Pconv model and Gconv operator, in this study, we design a U-Net-like architecture, replacing all the ordinary convolutional layers with the Gconv layer and using the nearest neighbor upsampling method in the decoder path. Specifically, the encoder part of the model consists of 8 Gconv blocks, each of which includes a Gconv layer with the stride of 2, followed by an optional Batch Normalization (BN) layer and a Rectified Linear Unit (ReLU) activation function. The decoder stage of the model, similarly, contains 8 Gconv blocks, each of which consists of a nearest neighbor upsampling layer, a Gconv layer, an optional BN layer, followed by a LeakyReLU activation function. Similar to the U-Net model, the skip connections concatenate the feature maps and corresponding binary masks from the encoder to the decoder path, acting as the feature and mask inputs to the next Gconv block. The final output layer of the model is an ordinary convolutional layer with a sigmoid activation function which is fed by a concatenation of the last Gconv block from the decoder path and the original input image with holes along with the original binary mask from the encoder. This strategy ables the model to directly transfer and copy the values of the valid pixels to the output layer. Figure 2 demonstrates a graphical illustration of the network architecture.

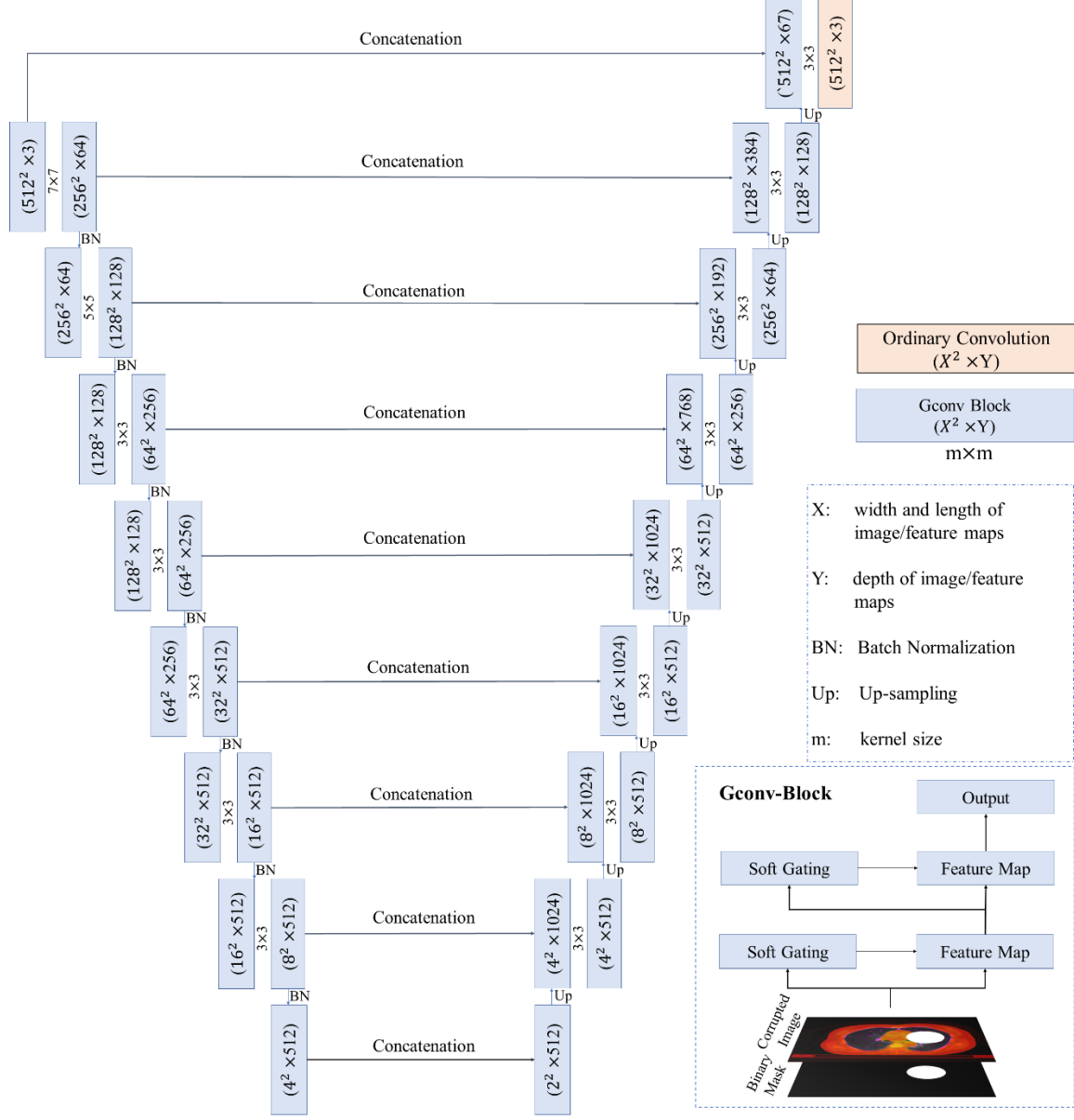


Figure 2. Schematic illustration of the model architecture.

In order to fill the holes with meaningful semantic patterns, the proposed model is optimized with a multi-term objective function [43] that takes into account both pixel-wise reconstruction accuracy and context information. Let the input image with holes be I_{in} ; I_{gt} represents the original image without holes (grand truth), I_{out} indicates the predicted image, and M denotes the binary mask used for corrupting the image; the first two terms in the objective functions are pixel-wise errors that can be calculated separately for the valid and invalid regions as the least absolute errors (L^1 norm). These two terms aim to minimize the intensity differences between the predicted and grand truth images inside and outside the hole regions separately:

$$\mathcal{L}_{valid} = \frac{1}{N_{I_{gt}}} \|M \odot (I_{out} - I_{gt})\|_1$$

$$\mathcal{L}_{hole} = \frac{1}{N_{I_{gt}}} \|(1 - M) \odot (I_{out} - I_{gt})\|_1$$

Where $N_{I_{gt}}$ shows the number of pixels in the I_{gt} .

The third term is the perceptual loss which aims to minimize the discrepancies between the high-level feature representations extracted from the predicted and grand truth images in order to maximize the perceptual similarity between these two images. It calculates the L^1 norm between two sets of high-level features extracted from I_{out} and I_{comp} where I_{comp} is the composite output which is similar to the predicted image but with the intensity of valid pixels replaced by those of the grand truth. 1st, 2nd, and 3rd layers of a VGG16 [50] network pre-trained on ImageNet were used to extract the features:

$$\mathcal{L}_{perceptual} = \sum_{p=0}^{p-1} \frac{\|\Psi_p^{I_{out}} - \Psi_p^{I_{gt}}\|_1}{N_{\Psi_p^{I_{gt}}}} + \sum_{p=0}^{p-1} \frac{\|\Psi_p^{I_{comp}} - \Psi_p^{I_{gt}}\|_1}{N_{\Psi_p^{I_{gt}}}}$$

Here, $\Psi_p^{I_*}$ refers to the outputs of the activation function of the p th layer of the pre-trained network given the input I_* .

In order to minimize the style differences between the synthesized and grand truth images, style loss was computed as well. To reconstruct images with high level of style similarities inside and outside of the holes, the style error was calculated for predicted and composite images separately:

$$\begin{aligned} \mathcal{L}_{style_{out}} &= \sum_{p=0}^{p-1} \frac{1}{C_p^2} \left\| \frac{1}{K_p} ((\Psi_p^{I_{out}})^T (\Psi_p^{I_{out}}) - (\Psi_p^{I_{gt}})^T (\Psi_p^{I_{gt}})) \right\|_1 \\ \mathcal{L}_{style_{comp}} &= \sum_{p=0}^{p-1} \frac{1}{C_p^2} \left\| \frac{1}{K_p} ((\Psi_p^{I_{comp}})^T (\Psi_p^{I_{comp}}) - (\Psi_p^{I_{gt}})^T (\Psi_p^{I_{gt}})) \right\|_1 \end{aligned}$$

As can be seen, the style loss is similar to the perceptual loss, but it first calculates the autocorrelation of extracted features and then computes the L^1 norm. In this notation, C_p indicates the depth of the channels in Ψ_p , and K_p refers to the number of elements in Ψ_p tensor.

The sixth loss term is Total Variation (TV) which is a conventional objective function for noise reduction applications. In fact, it functions as a smoothing term that makes the intensity values of the neighboring pixels in the synthesized image closer to each other:

$$\mathcal{L}_{tv} = \sum_{(i,j) \in R, (i,j+1) \in R} \frac{\|I_{comp}^{i,j+1} - I_{comp}^{i,j}\|_1}{N_{I_{comp}}} + \sum_{(i,j) \in R, (i+1,j) \in R} \frac{\|I_{comp}^{i+1,j} - I_{comp}^{i,j}\|_1}{N_{I_{comp}}}$$

Where $N_{I_{comp}}$ is the number of pixels in the composite image.

Finally, since the early layers of the model focus on capturing edge-based features, the described pixel-wise, perceptual, style, and TV losses alone cannot well preserve the high-frequency patterns and would lead to reconstructing blurry images. This issue will be problematic when the contents of each channel of the input image carry different structures, such as multimodal PET-CT images. Accordingly, to maintain the edges and synthesize images with details as much as possible, the last term includes the Laplacian (lap) pyramid loss:

$$\mathcal{L}_{lap(I_{out}, I_{gt})} = \sum_j 2^{2j} \|L^j(I_{out}) - L^j(I_{gt})\|_1$$

Where $L^j(x)$ refers to the j th level of the Laplacian pyramid representation of input x . In this study, the parameter j was set to 3, i.e., three levels of pyramid representations were computed.

Therefore, the overall objective function is the combination of all the mentioned loss terms:

$$\mathcal{L}_{total} = 30\mathcal{L}_{valid} + 240\mathcal{L}_{hole} + 0.2\mathcal{L}_{perceptual} + 0.05(\mathcal{L}_{style_{out}} + \mathcal{L}_{style_{comp}}) + 250\mathcal{L}_{tv} + 20\mathcal{L}_{lap}$$

The coefficient of each term was calculated after conducting an extensive ablation study over 2000 test images (see section 2.1 in Supplementary Materials).

2.4. Learning the appearance of normal anatomies

The proposed inpainting model was employed to learn the attributes of healthy anatomical structures by learning to fill the irregular holes with the characteristics of healthy structures. In other words, healthy image slices corrupted with irregular random holes are used to train the inpainting model. Having the corrupted healthy images as input to the model on one side and the original healthy images as the grand truth on the other side, the inpainting model is trained to smoothly replace the holes with semantically meaningful patterns in order to synthesize realistic-looking images while preserving fine-grained details and anatomical constraints. With this strategy, the inpainting model is assumed to estimate the distribution of healthy anatomies.

Considering the fact that tumors appear with irregular shapes and different sizes at different locations, the corrupting holes should be generated in a way to imitate the visual attributes of the tumors. Accordingly, irregular holes were synthesized by carefully combining the ordinary regular geometric shapes, including circles, ellipses, and lines. Thus, the simulated holes were distributed randomly over different spatial coordinates of the image space to occupy, on average per batch, 25 to 30 percent of the image size. With this approach, two models were trained separately for NSCLC and HN datasets. In specific, 6233 healthy images from the NSCLC dataset and 12171 healthy slices from the HN dataset were extracted to train the inpainting model. An additional 2000 slices from each dataset were used as the validation set.

Each model was trained for 300 epochs with Adam optimizer and a batch size of 8. The presence of the holes in the image causes issues for the BN parameters updating because the zero values inside the holes will contribute to updating the mean and variance of BN. Accordingly, it sounds rational to disable the calculation of the BN inside the holes. On the other hand, the training procedure forces the model to gradually fill the holes until they completely disappear so that they can potentially contribute to the BN parameter updating. Hence, the training was done in two phases. In the first phase, the models were trained for 150 epochs with a learning rate of 0.0001 and enabled all the BN layers. In the second phase, the model continues training for another 150 epochs with a learning rate of 0.00005. In this phase, the BN layers within the encoder path were disabled while they were enabled for the decoder stage. This fine-tuning strategy is not only beneficial to speed up the convergence but also to avoid the incorrect calculations of the mean and variance parameters of the BN operator [43][51]. The accuracy metrics over the validation set were monitored, and a certain epoch that resulted in the best accuracy metrics was used for the testing phase. It is worth mentioning that the described training procedure was performed independently for each of the examined imaging modalities, i.e., CT, PET, and PET-CT scans. Figure 3 demonstrates the qualitative performance of the model in replacing the irregular holes with the appearance of normal anatomical regions.

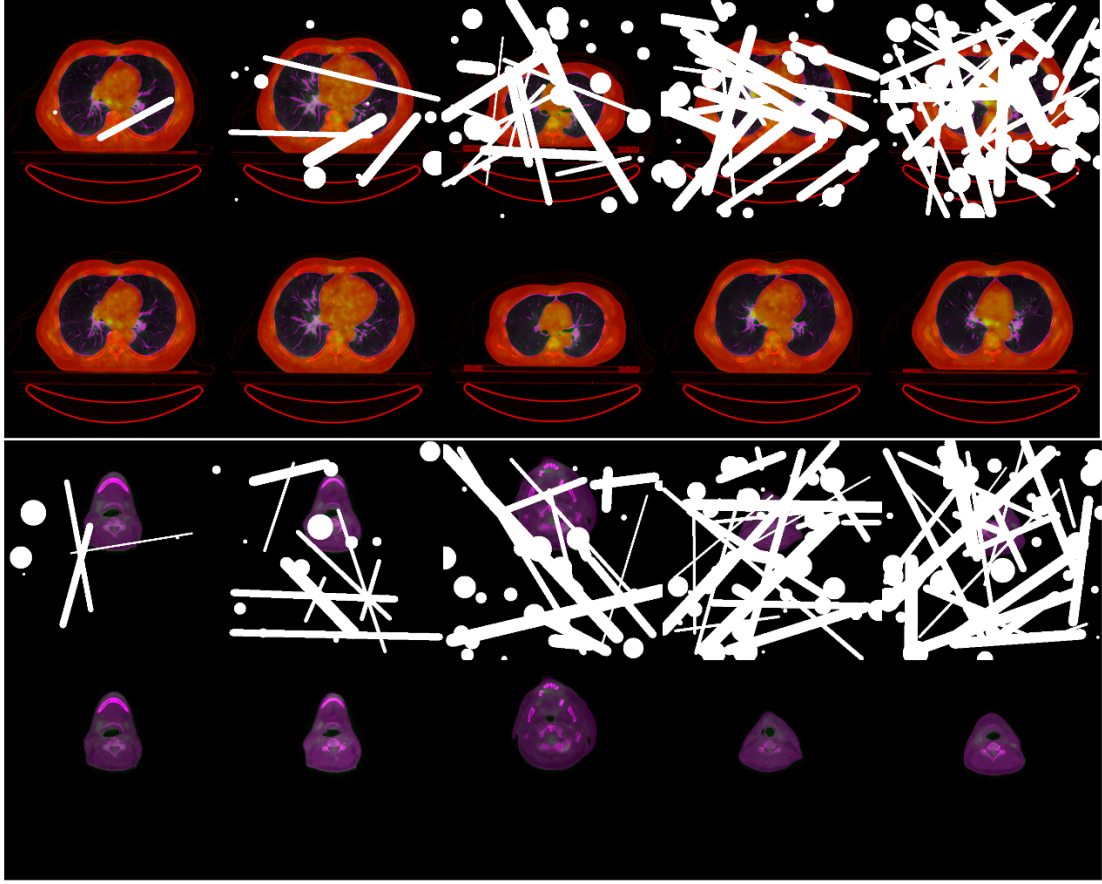


Figure 3. The inpainting model could successfully replace the irregular random holes with the appearance of healthy anatomies while preserving the anatomical constraints. For each set of NSCLC and HN tumors, the first row shows the corrupted images with random holes, and the second row illustrates the inpainted results of the model.

2.5. Autoinpainting for unsupervised tumor segmentation

The trained inpainting model learns to synthesize semantically correct and contextually smooth contents in the predefined missing regions. Training the model only with healthy slices reinforces the model to replace the missing healthy tissues with the appearance of healthy tissues. This strategy enables the inpainting network to model the distribution of healthy anatomical structures that can be further utilized to detect the anomalies as outliers from the learned normative distribution. In other words, replacing the tumor with the appearance of already learned healthy tissues leads to synthesizing tumor-free images from which the tumoral regions can be detected by calculating the differences between the original and synthesized images. Accordingly, the learned inpainting network, which was trained only with random holes, can function as a UAD model, given that no segmentation label is required to localize the tumor location. That being the case, a pipeline is proposed to turn the manual inpainting network into an autoinpainting model to segment the tumors in an unsupervised fashion.

The underlying idea thereby is to replace the random holes with a sliding window to sweep different anatomical regions for the inpainting process. Therefore, if the sliding window covers healthy regions, the inpainting network will replace the appearance of healthy structures with learned healthy structures; thus, the newly generated images remain intact. On the other hand, if the sliding window encounters tumoral regions, it substitutes the textures of the tumors with the appearance of already learned healthy tissues. Accordingly, for each original tumoral slice, a fake tumor-free image can be generated without needing any kind of supervised signal. Hence, a pipeline is proposed to efficiently inpaint the tumoral regions while preserving the appearance of healthy tissues with anatomical constraints. This pipeline functions

in the following four folds: I) preparing the input slices, II) detecting the candidate regions, III) determining the target region, IV) segmenting the target tumor:

Preparing the input slices

The images within the NSCLC dataset covered only the chest region, while the HN images covered not only the HN region but also brain and chest organs as well. Thus, to concentrate the analyses within the target organs, lung field masks for NSCLC data and HN masks for HN data were delineated. In specific, a pretrained Progressive Holistically-Nested Networks(P-HNNs) [52] was used for the CT volumes to accurately segment the lung fields in the presence of pathologies. The segmentation masks were visually examined, and manual refinements were needed only for a very limited number of cases. Then, the binary lung field masks were applied to the corresponding PET images as well. The CT images of the HN dataset, on the other hand, were visually examined to manually crop out the brain and lung tissues from the volumes. The cropped binary masks were later applied to the corresponding PET images. This step assures us that all the further analyses will be performed within the Organ Of Interest (OOI) where the tumors are presented. The last step includes the extraction of all the axial slices from the OOIs.

Detecting the candidate regions

Depending on the size of the OOIs, a certain number of subregions is determined with the help of a sliding window strategy for further analyses. In particular, a sliding circle with a radius of 27 pixels and an interval distance of 15 pixels sweeps over the OOIs in each of the axial slices. The already trained network is employed as an inference model to inpaint each of the moving circles independently. In other words, the sliding window scans each slice to produce several candidate circles to be inpainted by the trained network. The inpainting model, therefore, replaces the contents of the coordinates occupied by the circles with the textural patterns it learns from the healthy images in the training phase. As a result, for each of the circles within one slice, there will be a new synthesized image. If the moving circle masks a healthy subregion, the inpainting model replaces it with the texture of healthy tissues, and therefore there will be no remarkable intensity/textural differences between the original and the synthesized images. On the other hand, if the moving circle masks a tumoral region, the learned inpainting model changes the textures of the tumor with the patterns of healthy tissues. In this case, remarkable intensity/textural differences between the input slice and the generated slice can be observed. Accordingly, to identify which of the moving windows could cover the tumor(s), the sum of the intensity differences between the input slices and the inpainted slices was calculated for each of the moving circles. These values were then sorted, and only the top few values with notable differences w.r.t the other values were kept as these larger intensity difference values represent notable changes between the input image and the synthesized one, which could potentially imply the tumor location.

It should be emphasized that if the size of the moving window is too small, the inpainting model will not be able to completely replace the tumoral regions. On the other hand, if this size is too large, it may slightly change so many tiny details, which would slightly change the general context. Thus, this size should be defined as a trade-off between the largest and smallest possible tumors within the datasets. In this study, based on the diversity of tumor sizes, a range of potential values were examined in an ablation study which yields set the radius of the moving window equal to 27 as the optimal value (See section 2.2 in Supplementary Materials).

Determining the target region

The identified top candidate regions either masked one single tumor or covered different anomalies related to multi-focal tumors. To automatically find out whether the top candidate regions share the same tumor or they focus on various subregions, the union of the top candidate binary masks is calculated. To this extent, if the top candidate regions overlap each other, their union will form a larger binary mask; however, if they don't share even a single pixel, the outcome of the union calculation will not differ from the originally separated masks. This simple scheme ensures us whether only one tumor or several tumors are presented in the slice. Then, the updated union mask will be ready to perform the final inpainting step. Considering the possibility of presence of extremely large-size tumors, this final mask may not be large enough to cover the whole abnormalities. Accordingly, the size of this binary mask needs to be enlarged without compromising the efficacy of small-size anomalies. To do so, an incremental morphological dilation approach is adopted

in order to dilate the updated binary mask with square-shaped structural elements of the width of [7,9,11,13,15]. Simply explaining, in addition to the updated union mask, five other dilated versions of this mask will be generated to conduct a total number of six final inpaintings sequentially. For each of them, the intensity differences between the input slice and the inpainted slices will be quantified, and if no changes are observed between the sequential orders, then the mask with the smaller size is selected; otherwise, the one with the larger size will be set as the final candidate mask(s). In this way, the small-size tumors will not be affected by this incremental dilation strategy as they remain inpainted with the updated union mask, while the extremely large-size tumors can be covered more efficiently by the dilated masks.

Segmenting the target tumor

The proposed pipeline analyzes all the axial OOI slices; however, not all the slices contain tumors. Therefore, to prevent the model from detecting small deviations in healthy slices as anomalies, a size-based criterion is included in the pipeline. In fact, the radius of the smallest tumor in the studied dataset is 6 pixels. Having known the minimum value, any detected abnormalities with sizes smaller than the minimum radius can be recognized as a false positive and skipped from the further steps. To implement this concept, first, the residual images are calculated as the differences between the input and the final inpainted images. Connect Components (CCs) of the residual images are computed, and the size of the largest CC at each slice is compared against the minimum radius of the tumors. If the condition is satisfied, the output of the algorithm will become the final inpainted slice; otherwise, the input slice will be directly set as the output. The latter case necessarily means that either the model could not detect the tumor(s) or the image slice does not contain any tumors. Figure 4 illustrates a general schematic presentation of the autoinpainting pipeline. Please note that the segmentation of NSCLC tumors in CT images is more challenging than the multimodal PET-CT images; therefore, the graphic illustration in figure 4 is depicted on a CT slice to accentuate the abilities of the proposed pipeline.

The mentioned process is repeated for all the axial slices from which a stack of volume can be formed from the algorithm outputs. Therefore, for each input volume, there will be a synthesized autoinpainting volume. The intensity range of both input and synthesized one is scaled between 0 to 1. The final residual volume is then computed as intensity differences between the two volumes. Finally, to quantify the segmentation performance, two approaches were followed. First, a conventional quantification was done by setting a single threshold value to binarize all the residual volumes. Second, a variable threshold in the range of 0 to 0.8 with an incremental rate of 0.02 was used to binarize the residuals for further quantifications, from which the threshold that leads to the best segmentation accuracy was selected. Therefore, for each subject, a different threshold value was used for the quantification. These metrics are reported by the [] notations.

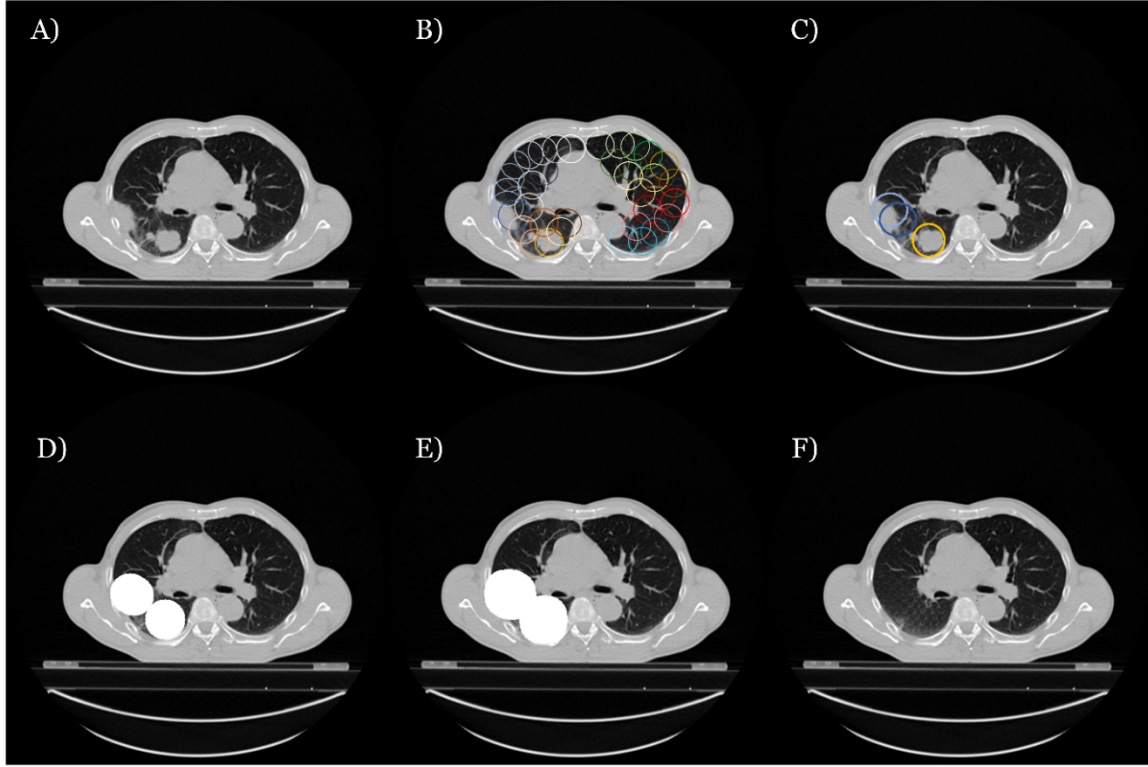


Figure 4. The autoinpainting pipeline employs the moving window strategy to adaptively inpaint the tumoral regions in a pure unsupervised approach. A) the original slice consists of multi-focal tumors; B) the determined circles to be inpainted independently are presented with different colors; C) the top three candidate circles detected the two different tumors; D) the union of candidate regions was used to corrupt the image for inpainting process; E) incrementally increasing the size of the detected regions better cover the tumoral zones and F) final inpainted image does not contain the tumors anymore.

2.6. External validation

To benchmark the efficacy of the proposed model, we compare its performance to State-Of-The-Art (SOTA) models in two folds: 1) A supervised segmentation model was employed to find out what optimal performance can be achieved over the investigated tasks. In specific, the self-configuring nnU-Net model [53] as a powerful segmentation framework was utilized to estimate the maximum achievable segmentation accuracy of the studied dataset. This model was trained with a 5-fold cross-validation fashion for each dataset separately. The default settings of the nnU-Net framework were adopted without further modifications, and the models were trained for 1000 epochs. 2) A set of recently developed deep UAD models were analyzed as well to objectively compare the segmentation accuracy of the proposed unsupervised model against the relevant UAD references. In this context, the following models were examined [32]: dense AE (dAE), spatial AE (sAE), context-encoding AE (ceAE), Variational AE (VAE), context-encoding Variational AE (ceVAE), Gaussian Mixture Variational AE (GMVAE), Fast-AnomalyGAN (F-AnoGAN), and Adversarial AE (AAE). Similar to the proposed autoinpainting model, for each of the datasets, healthy slices were used to train these UAD models, and the pathological slices were employed in the test phase. It is worth mentioning that standard implementations of these models were used for fair comparisons [54].

2.7. Quantitative Evaluation

To assess the performance of the proposed inpainting model and the segmentation pipelines, two sets of quantitative metrics were examined.

The first group of metrics includes Mean Square Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM). These metrics are measured to quantitatively evaluate the performance of the inpainting network

by directly comparing the original image to the synthesized one. The MSE metric measures the amount of changes per pixel between the two images; therefore, the smaller value of this measure represents more similarity between the two images. PSNR is another quality assessment measure between the two images where the higher PSNR value indicates the better quality of the synthesized image. SSIM assesses the perceptual image quality to quantify the visible differences between the two images. Let the original image be I_{org} , and I_{out} shows the synthesized image with equal matrix sizes of $m \times n$ and the maximum possible intensity value of R ; then, the metrics can be mathematically defined as:

$$MSE(I_{org}, I_{out}) = \frac{1}{m \times n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |I_{org}(i, j) - I_{out}(i, j)|^2$$

$$PSNR(I_{org}, I_{out}) = 10 \log_{10} \left(\frac{R^2}{MSE(I_{org}, I_{out})} \right)$$

$$SSIM(I_{org}, I_{out}) = \frac{(2\mu_{I_{org}}\mu_{I_{out}} + c_1)(2\sigma_{I_{org}I_{out}} + c_2)}{(\mu_{I_{org}}^2 + \mu_{I_{out}}^2 + c_1)(\sigma_{I_{org}}^2 + \sigma_{I_{out}}^2 + c_2)}$$

Where $\mu_{I_{org}}$, and $\mu_{I_{out}}$ are average intensities; $\sigma_{I_{org}}^2$ and $\sigma_{I_{out}}^2$ are variance values and $\sigma_{I_{org}I_{out}}$ represents the covariance of the two images. Parameters c_1 and c_2 are two variables that ensure stability when the denominator becomes 0.

The second group of metrics is used to quantify the segmentation accuracy of the proposed pipeline. These metrics include Dice coefficient (DSC), Precision, and Recall. While DSC measures the overlap between the target masks and model predictions, Precision and Recall metrics demonstrate the accuracy of pixel classifications. Given that P represents the segmentation output of the model and G refers to the grand truth mask, T_p , F_p , F_N show true positive, false positive, and false negative, respectively, calculated from the confusion matrix, the definitions of the metrics are formulated as follows:

$$DSC = \frac{2|S \cap G|}{|S| + |G|}$$

$$Recall = \frac{T_p}{T_p + F_N}$$

$$Precision = \frac{T_p}{T_p + F_p}$$

3. Results

In this section, the performance of the proposed method for unsupervised tumor segmentation is presented in two folds: (1) the quality of the inpainting model, (2) the segmentation accuracy of the autoinpainting pipeline.

3.1. Inpainting Quality

There exist many possible solutions to quantify the performance of inpainting models; therefore, no specific numerical metrics were designed for this task. Nevertheless, we employed the described MSE, PSNR, and SSIM metrics as conventionally have been used by other studies [43], [44]. Furthermore, qualitative comparisons are included by demonstrating both the corrupted and inpainted images. In the followings, $G_{convLap}$ denotes the proposed method, which is compared against P_{conv} and G_{conv} models.

Tables 1, and 2 represent the comparison results between the performance of the models for each of the PET-CT images, CT channel, and PET channel of multimodal images for the NSCLC dataset and HN dataset separately. In

specific, the already trained models were used in the test phase to inpaint the corrupted input images. Original images were then compared against the model predictions using the three quantitative metrics.

Table 1 – Numerical comparison between the performance of inpainting models on the NSCLC dataset

Model-Data	Quantitative Metrics ($\mu \pm \sigma$)		
	<i>MSE</i>	<i>PSNR</i>	<i>SSIM</i>
Pconv-CT	123.401 \pm 66.536	27.915 \pm 2.623	0.908 \pm 0.033
Gconv-CT	67.098 \pm 48.486	31.311 \pm 4.022	0.939 \pm 0.031
Gconv _{Lap} -CT	66.041 \pm 47.330	31.495 \pm 4.332	0.943 \pm 0.030
Pconv-PET	22.722 \pm 22.925	35.981 \pm 3.413	0.961 \pm 0.014
Gconv-PET	21.931 \pm 28.111	37.449 \pm 5.094	0.973 \pm 0.015
Gconv _{Lap} -PET	21.888 \pm 31.336	38.070 \pm 5.836	0.977 \pm 0.013
Pconv-Multi	69.428 \pm 37.546	30.385 \pm 2.530	0.947 \pm 0.019
Gconv-Multi	45.850 \pm 32.813	32.814 \pm 3.682	0.960 \pm 0.018
Gconv _{Lap} -Multi	44.290 \pm 33.785	33.271 \pm 4.267	0.966 \pm 0.018

From Table 1, we can infer that the proposed Gconv_{Lap} model could inpaint the corrupted images more accurately than the other two models. In particular, the numerical metrics obtained from the proposed Gconv_{Lap} indicate fewer errors in terms of MSE metric and higher similarity in terms of PSNR and SSIM for all the experiments regardless of the type of the input images. As expected, quantitative values of the PET image show higher accuracy compared to those of the CT and multimodal images for all the experiments.

Table 2 – Numerical comparison between the performance of inpainting models on the HN dataset

Model-Data	Quantitative Metrics ($\mu \pm \sigma$)		
	<i>MSE</i>	<i>PSNR</i>	<i>SSIM</i>
Pconv-CT	9.934 \pm 8.367	39.922 \pm 4.561	0.985 \pm 0.012
Gconv-CT	7.136 \pm 7.504	42.295 \pm 5.868	0.988 \pm 0.011
Gconv _{Lap} -CT	5.744 \pm 6.177	43.622 \pm 6.396	0.991 \pm 0.009
Pconv-PET	5.370 \pm 10.208	45.476 \pm 6.732	0.992 \pm 0.006
Gconv-PET	4.270 \pm 8.621	46.462 \pm 6.660	0.991 \pm 0.007
Gconv _{Lap} -PET	3.130 \pm 6.199	48.530 \pm 7.579	0.995 \pm 0.005
Pconv-Multi	8.412 \pm 7.457	40.689 \pm 4.560	0.986 \pm 0.010
Gconv-Multi	6.155 \pm 6.536	42.828 \pm 5.659	0.989 \pm 0.00
Gconv _{Lap} -Multi	4.851 \pm 5.241	44.268 \pm 6.287	0.991 \pm 0.008

Similar to the NSCLC experiments, for the HN dataset, the proposed Gconv_{Lap} model outperforms the other methods with respect to the quality of the inpainted images. It should be noted that both NSCLC and HN datasets were trained and tested under similar conditions, including the network parameters, shape, and size of the irregular holes. Therefore, the only reason that the range of the reported numerical values is different between the two datasets is related to the fact that the HN images entail fewer contents and textures compared to NSCLC images. In addition to assessing the inpainting models with multimodal datasets, the models were trained and tested with single modality images as well. In other words, for each of the NSCLC and HN datasets, CT images and PET images were independently used to train and test the quality of the inpainting models (Tables 1.1. and 1.2. in Supplementary Materials). Similar to multimodal inpainting networks, even for the single modality images, Gconv_{Lap} outperformed the other methods with a rather remarkable margin. To test the statistical significant difference between the performance of the Gconv_{Lap} model and the two other inpainting baselines, Wilcoxon signed rank test as a non-parametric method was applied on the calculated image quality metrics (see Table 1.3. in Supplementary Material).

Figure 5 demonstrates the qualitative comparisons between the functionality of the inpainting models in filling the random holes with meaningful patterns in the multimodal NSCLC dataset. The irregular holes were randomly distributed

over different locations on the image plane to learn the heterogeneous appearance of anatomical structures such as ribs, cardiac muscle, aorta, arteries, chest wall, etc.

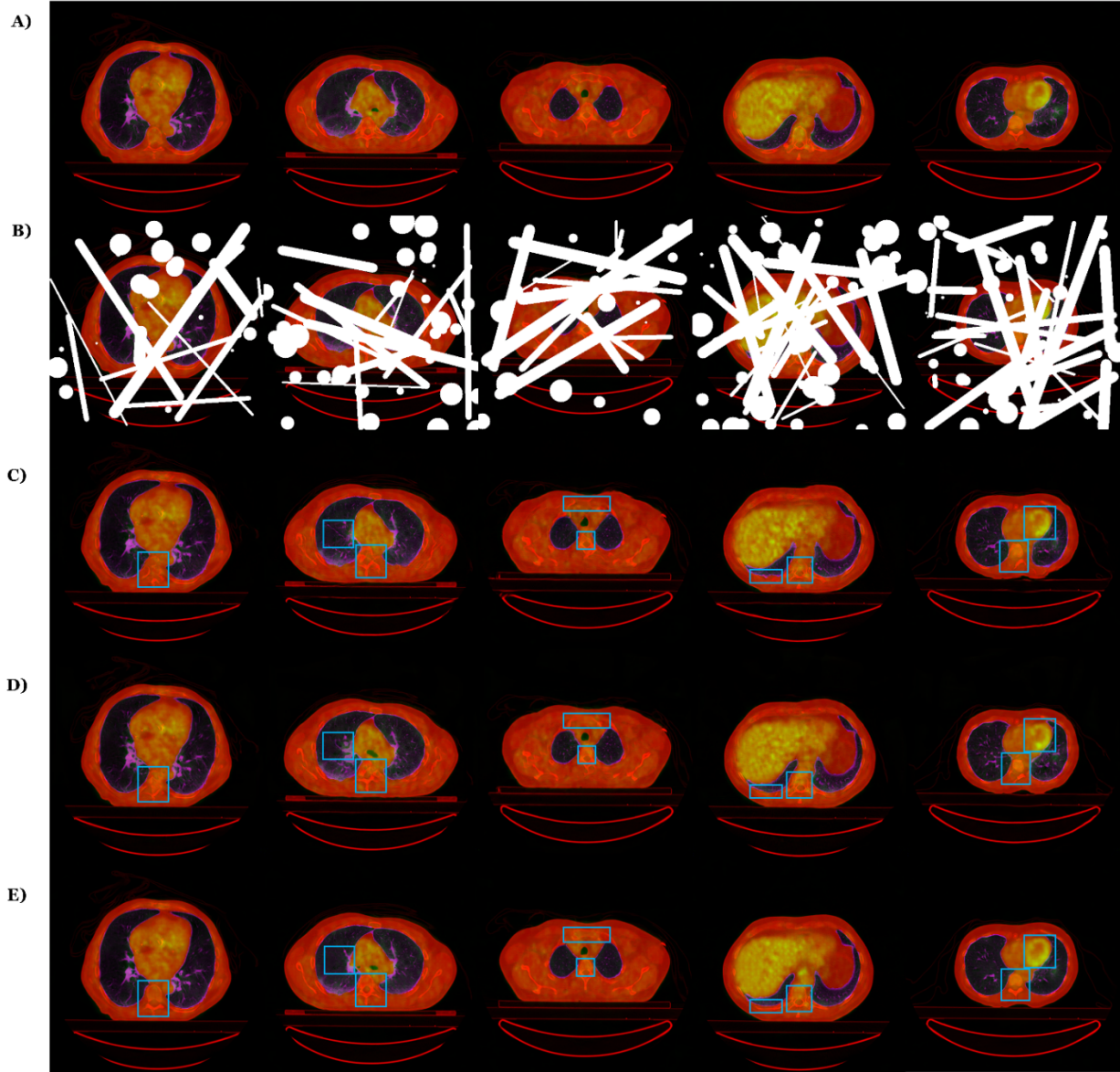


Figure 5. Qualitative comparisons of image inpainting performance. Row A: original PET-CT slices; row B: corrupted slices with random holes; row C: inpainted results by Pconv model; row D: inpainted results by Gconv model; and row E: inpainted results by the proposed GconvLap model. As can be seen, the proposed GconvLap model could replace the irregular holes with meaningful anatomical patterns and preserve the anatomical constraints far better than the other two methods. The blue bounding boxes highlighted the regions where the inpainted patterns by the proposed model are much more meaningful anatomically than the other models.

For both the NSCLC and HN datasets, quantitative values show that the performance of the proposed $Gconv_{Lap}$ model is far better than the Pconv model and slightly better than the Gconv model. Nonetheless, the capability of the proposed $Gconv_{Lap}$ model in preserving the anatomical constraints is highlighted in Figure 5. In specific, while the Pconv and Gconv models filled the random holes with semantic image contents, they were not able to synthesize anatomically meaningful contents. From the qualitative comparisons between the anatomical regions highlighted with the blue boxes in Figure 4, it can be understood that the proposed $Gconv_{Lap}$ model synthesized plausible image contents with highly realistic anatomical details. Therefore, both the image details and cotextual patterns of the inpainted images synthesized

by the proposed model are more similar to those of the original images, which in return leads to reducing the reconstruction errors.

3.2. Autoinpainting for Tumor Segmentation

The performance of the proposed autoinpainting pipeline for tumor segmentation is quantified by finding the agreement between the segmented volumes and the label masks. The same autoinpainting pipeline was applied to all the three inpainting models, followed by the same postprocessing steps for tumor segmentation. Tables 3 and 4 represent the segmentation accuracy of the proposed autoinpainting strategy for NSCLC and HN tumors, respectively.

Table 3 – Numerical results of NSCLC tumor segmentation with autoinpainting pipeline

Model-Data	Quantitative Metrics ($\mu \pm \sigma$)			<i>Dice</i>
	[<i>Dice</i>]	[<i>Precision</i>]	[<i>Recal</i>]	
Pconv-CT	0.382 \pm 0.157	0.408 \pm 0.186	0.389 \pm 0.151	0.353 \pm 0.111
Gconv-CT	0.423 \pm 0.180	0.463 \pm 0.199	0.411 \pm 0.178	0.398 \pm 0.124
Gconv _{Lap} -CT	0.442 \pm 0.176	0.482 \pm 0.192	0.426 \pm 0.176	0.410 \pm 0.134
Pconv-PET	0.709 \pm 0.215	0.793 \pm 0.196	0.669 \pm 0.221	0.654 \pm 0.132
Gconv-PET	0.750 \pm 0.176	0.792 \pm 0.192	0.747 \pm 0.189	0.690 \pm 0.184
Gconv _{Lap} -PET	0.746 \pm 0.196	0.822 \pm 0.169	0.706 \pm 0.217	0.686 \pm 0.121
Pconv-Multi	0.673 \pm 0.245	0.771 \pm 0.219	0.622 \pm 0.252	0.625 \pm 0.122
Gconv-Multi	0.747 \pm 0.172	0.799 \pm 0.178	0.718 \pm 0.183	0.692 \pm 0.136
Gconv _{Lap} -Multi	0.766 \pm 0.171	0.832 \pm 0.158	0.726 \pm 0.184	0.708 \pm 0.118

From table 3, we can observe that the segmentation accuracy achieved by the proposed Gconv_{Lap} model is remarkably higher than that of the PConv model, regardless of the type of input images. The same trend can be seen when comparing the Gconv_{Lap} model with the ordinary Gconv model for the CT and multimodal images though the Gconv model slightly performs better on the PET images.

Table 4 – Numerical results of HN tumor segmentation with autoinpainting pipeline

Model-Data	Quantitative Metrics ($\mu \pm \sigma$)			<i>Dice</i>
	[<i>Dice</i>]	[<i>Precision</i>]	[<i>Recal</i>]	
Pconv-CT	NA	NA	NA	NA
Gconv-CT				
Gconv _{Lap} -CT				
Pconv-PET	0.412 \pm 0.190	0.541 \pm 0.230	0.462 \pm 0.172	0.389 \pm 0.132
Gconv-PET	0.445 \pm 0.188	0.557 \pm 0.407	0.408 \pm 0.181	0.407 \pm 0.130
Gconv _{Lap} -PET	0.453 \pm 0.196	0.550 \pm 0.236	0.414 \pm 0.181	0.405 \pm 0.130
Pconv-Multi	0.408 \pm 0.241	0.511 \pm 0.274	0.360 \pm 0.224	0.344 \pm 0.100
Gconv-Multi	0.462 \pm 0.202	0.539 \pm 0.233	0.443 \pm 0.189	0.418 \pm 0.133
Gconv _{Lap} -Multi	0.465 \pm 0.198	0.541 \pm 0.233	0.435 \pm 0.188	0.422 \pm 0.135

Similar to the NSCLC tumors, the segmentation accuracy of HN tumors achieved by the proposed Gconv_{Lap} outperformed the Pconv model with a relatively large margin and performed slightly better than the ordinary Gconv model on the PET and multimodal images. The appearance, textural distributions, and Hounsfield values of the HN tumors are very similar to those of the surrounding soft tissues (see Figure 1.1. in Supplementary Materials). Hence, the HN tumors in CT images cannot be distinguished directly from the nearby structures due to the lack of visible contrasts. Therefore, none of the inpainting approaches is able to detect abnormal tumoral tissues. In this domain, it is worth mentioning that even the supervised segmentation methods can hardly detect the HN tumors in full resolution CT images. For instance, in reference [55], a promising Dice score of 0.48 was reported for the HN tumor segmentation in CT images when only a cropped region around the tumors was analyzed. The proposed unsupervised autoinpainting pipeline was not able to detect the HN tumors in CT images; therefore, the notation of “NA” was used in the relevant row of Table 4. Table

1.4. in Supplementary Materials shows the results of the applied Wilcoxon signed rank test on the Dice values achieved by the autoinpainting pipeline.

Figure 6 illustrates the capability of the proposed pipeline in segmenting the tumors in multimodal images. Figure 1.2. in Supplementary Materials depicts the same illustration in single modality images.

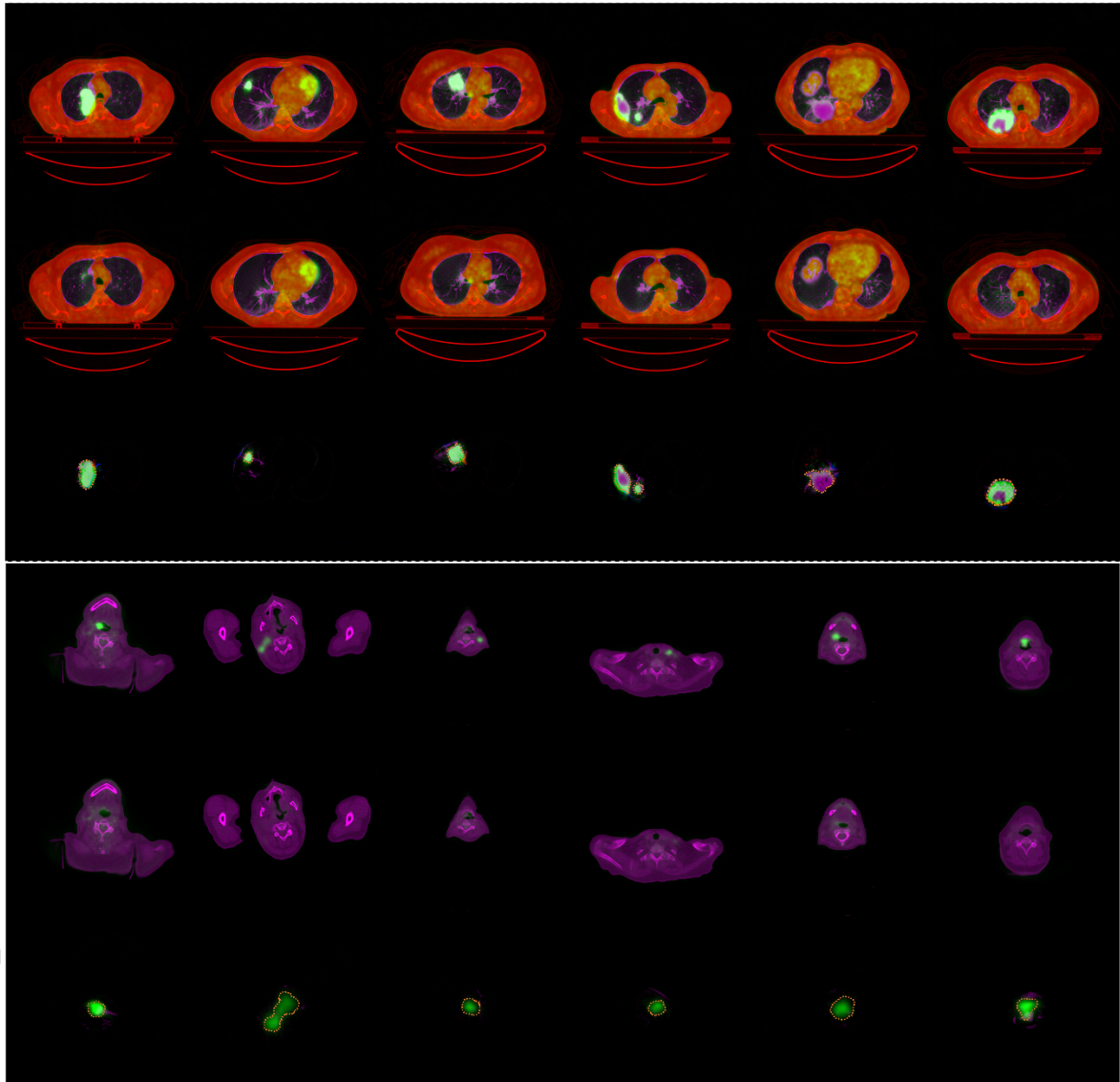


Figure 6. Visualization of the segmentation performance of the proposed autoinpainting pipeline. For each of the NSCLC and HN images, the first row shows the original tumoral slices, the second row depicts the result of the proposed autoinpainting model, and the last row demonstrates the residuals between the two images. Please note that residual images were zoomed around the tumoral candidates to better visualize the qualitative comparison between the detected tumors and the grand truth (dashed orange contours).

Of course, there are certain cases in which the proposed unsupervised method faces some difficulties in segmenting the tumors. Figures 1.3. to 1.6. in Supplementary Materials depict different examples of challenging cases where the proposed pipeline failed to completely remove the tumors.

3.3. Supervised Tumor Segmentation

Table 5 presents the segmentation accuracy of the supervised nnU-Net model, which was trained with a 5-fold cross-validation resampling strategy for each of the NSCLC and HN tumors independently.

Table 5 – Numerical results of supervised segmentation accuracy achieved by the nn-UNet model

Tumor-Data	Quantitative Metrics ($\mu \pm \sigma$)		
	<i>Dice</i>	<i>Precision</i>	<i>Recall</i>
NSCLC-CT	0.707 \pm 0.224	0.762 \pm 0.238	0.713 \pm 0.258
NSCLC-PET	0.802 \pm 0.177	0.802 \pm 0.231	0.854 \pm 0.174
NSCLC-Multi	0.802 \pm 0.179	0.847 \pm 0.182	0.812 \pm 0.231
HN-CT	0.293 \pm 0.208	0.275 \pm 0.232	0.269 \pm 0.230
HN-PET	0.641 \pm 0.177	0.636 \pm 0.200	0.704 \pm 0.215
HN-Multi	0.660 \pm 0.179	0.653 \pm 0.209	0.731 \pm 0.212

Similar to the autoinpainting results, the supervised segmentation accuracy over the PET images is higher than CT images for both NSCLC and HN tumors. Moreover, integrating both modalities together into the segmentation pipeline yielded the best results, which were even more accurate than PET images alone.

As was expected, the supervised models segment the tumors more accurately than the proposed unsupervised pipeline. However, carefully comparing the results, we can observe that the performance of the unsupervised autoinpainting models is not far behind the powerful supervised nnU-Net models, especially in the cases of multimodal and PET images. For instance, the Dice score achieved by the proposed $G_{convLap}$ model for multimodal NSCLC tumors is 0.708, which is around 10 percent lower than that of the nnU-Net model (Dice = 0.802). For the case of HN tumors, the nnU-Net model outperformed the unsupervised approach with remarkable margins ($Dice_{G_{convLap}multi} = 0.422$ vs. $Dice_{nnU-Netmulti} = 0.660$). However, as was already described in section 2.6, comparing the differences between the supervised and unsupervised methods is not fair. In fact, the only reason that the supervised nnU-Net model was examined is to estimate the maximum accuracy which can be achieved on the same datasets.

3.4. Tumor Segmentation with UAD methods

The segmentation accuracy of the employed UAD methods in multimodal images is presented in tables 6 and 7. Tables 1.5. to 1.7. in Supplementary Materials show the same evaluations for single modality images. In fact, eight conventional UAD models have been examined to benchmark the performance of the proposed unsupervised autoinpainting.

Table 6 – Segmentation accuracy of unsupervised anomaly detection models on multimodal images of NSCLC tumors

Model	Quantitative Metrics ($\mu \pm \sigma$)			
	[<i>Dice</i>]	[<i>Precision</i>]	[<i>Recal</i>]	<i>Dice</i>
dAE	0.305 \pm 0.122	0.270 \pm 0.132	0.405 \pm 0.147	0.285 \pm 0.068
sAE	0.097 \pm 0.047	0.064 \pm 0.038	0.249 \pm 0.072	0.094 \pm 0.030
ceAE	0.346 \pm 0.129	0.330 \pm 0.1464	0.407 \pm 0.144	0.314 \pm 0.078
VAE	0.311 \pm 0.132	0.271 \pm 0.142	0.421 \pm 0.158	0.282 \pm 0.068
ceVAE	0.254 \pm 0.109	0.228 \pm 0.126	0.320 \pm 0.119	0.242 \pm 0.069
GMVAE	0.023 \pm 0.016	0.012 \pm 0.008	0.583 \pm 0.117	0.023 \pm 0.004
F-AnoGAN	0.262 \pm 0.133	0.286 \pm 0.158	0.390 \pm 0.180	0.262 \pm 0.073
AAE	0.277 \pm 0.129	0.284 \pm 0.167	0.335 \pm 0.159	0.237 \pm 0.059

Comparing the numerical values of table 6 to those in table 3, one can obviously observe that the proposed autoinpainting pipeline significantly outperformed all the UAD models on NSCLC tumors. In specific, the best Dice score in the UAD family achieved by the dAE model is 0.285, which is 0.42 inferior to the $G_{convLap}$ model (Dice=0.708). The same trend can be observed for the single modality images when comparing the segmentation accuracy of the proposed autoinpainting pipeline against the UAD models.

Table 7 – Segmentation accuracy of unsupervised anomaly detection models on multimodal images of HN tumors

Model	Quantitative Metrics ($\mu \pm \sigma$)			
	[Dice]	[Precision]	[Recal]	Dice
dAE	0.126 \pm 0.071	0.162 \pm 0.148	0.179 \pm 0.111	0.101 \pm 0.034
sAE	0.080 \pm 0.046	0.068 \pm 0.064	0.257 \pm 0.226	0.066 \pm 0.019
ceAE	0.128 \pm 0.072	0.167 \pm 0.149	0.174 \pm 0.105	0.101 \pm 0.034
VAE	0.148 \pm 0.086	0.196 \pm 0.179	0.197 \pm 0.104	0.120 \pm 0.033
ceVAE	0.119 \pm 0.082	0.176 \pm 0.163	0.218 \pm 0.176	0.109 \pm 0.023
GMVAE	0.049 \pm 0.028	0.026 \pm 0.016	0.479 \pm 0.090	0.049 \pm 0.008
F-AnoGAN	0.134 \pm 0.092	0.164 \pm 0.160	0.190 \pm 0.103	0.111 \pm 0.025
AAE	0.136 \pm 0.099	0.199 \pm 0.195	0.187 \pm 0.130	0.112 \pm 0.028

The UAD models were not able to deal with even more challenging HN tumors. In other words, while the proposed Gconv_{Lap} model could achieve a segmentation accuracy of 0.422 in multimodal HN tumors, the examined UAD models barely obtained a Dice score of 0.120. Similar behavior was observed with PET images, where the proposed autoinpainting model could outperform the UAD models. However, it should be noted that both UAD models and autoinpainting pipeline were failed to segment the HN tumors in CT images.

Figure 7 visualizes a qualitative comparison between the proposed autoinpainting approach and the employed UAD models. Such comparisons signify the superiority of the proposed unsupervised autoinpainting approach over the conventional UAD models. In fact, the ability of the Gconv_{Lap} model to reconstruct high-resolution images by preserving the anatomical constraints on one side and its potential to detect and remove the tumors without corrupting the remaining anatomical structures on the other side boost the performance of the autoinpainting approach. On the other hand, the UAD models can neither preserve the anatomical constraints nor completely replace the tumors with healthy tissues. Figures 7 and 8 in Supplementary Materials show the same concept for the PET-CT images of HN tumors and CT images of NSCLC tumors.

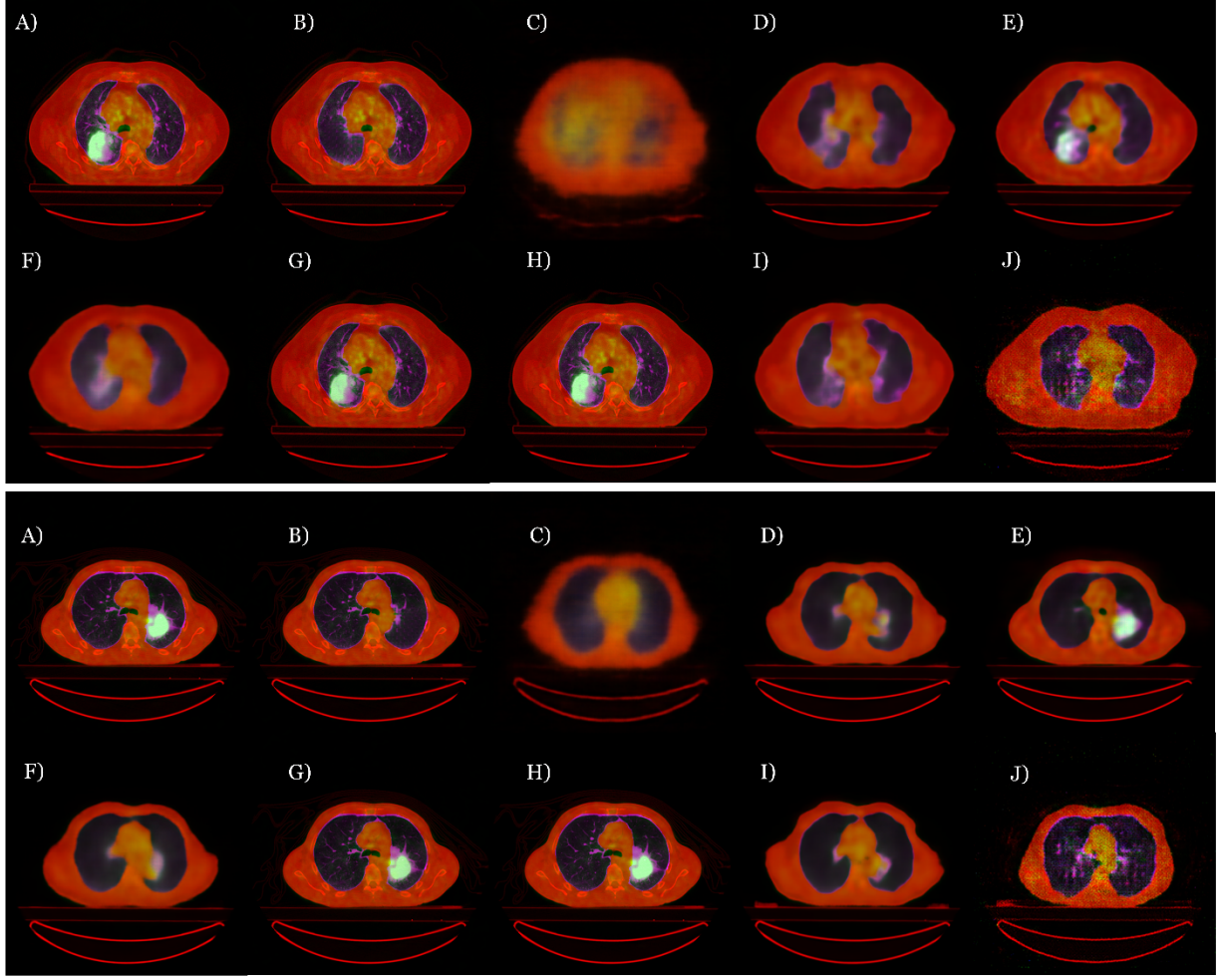


Figure 7. Qualitatively comparing the performance of the proposed autoinpainting pipeline against eight UAD models in learning the appearance of healthy lungs. Each set of images consists of: A) original tumor slice, A) original tumor slice, B) proposed autoinpainting image, C) adversarial autoencoder result, D) dense autoencoder result, E) spatial autoencoder result, F) variational autoencoder result, G) context-encoding variational autoencoder result, H) Gaussian mixture variational autoencoder result, I) context-encoding autoencoder result, and J) Fast-Anomaly GAN.

4. Discussion and conclusion

The detection and segmentation of tumors in medical images support a series of important clinical tasks, including diagnosis, prognosis, treatment, and surgery planning. The development of accurate computerized methods for automatic tumor segmentation has become a major endeavor in medical image analysis communities. Recent advances in deep learning-based methods have led to the development of robust models which could achieve even expert-level performance in some applications. However, most of the developed models depend on an explicitly defined target class for their supervised training procedures. This dependency, in general, increases the sensitivity to the quality and quantity of the available labeled data, which in turn limits the generalization power of the models. Recently, to overcome the necessity of expensive labeled data, UAD methods have emerged as promising tools to detect pathologies from arbitrary types. These methods aim to resemble how radiologists examine imaging scans. In fact, expert radiologists are trained to learn the appearance of healthy anatomical regions. Therefore, they do not need data with pixel-level annotations because they can detect arbitrary abnormalities as outliers with respect to the healthy anatomies [32], [56]. However, one of the limitations of conventional UAD models is that they hardly learn the appearance of healthy anatomical structures with fine-grained details. In specific, they often tend to learn a general representation of anatomical structures without

preserving the details of anatomical constraints. The main objective of this study has been focused on developing an autoinpainting model to segment the tumors by generating high-resolution medical images without the tumors while preserving the anatomical details in the process of representation learning. Specifically, we proposed a robust image inpainting model, $Gconv_{Lap}$, which is capable of capturing the appearance of normal anatomies and can synthesize high-resolution medical images by preserving the fine-grained anatomical details. This inpainting model was trained with healthy image slices to model the characteristics of healthy anatomies by learning to fill the irregular random holes with semantically and anatomically meaningful patterns. Then, an autoinpainting pipeline was developed to automatically inpaint the tumoral regions and synthesize high-quality tumor-free images. In fact, we hypothesized that the well-trained inpainting model would replace the tumoral tissues with the characteristics of already learned healthy structures and leave the healthy parts of the images intact. Therefore, the differences between the original tumoral images and the synthesized inpainted images can be used to segment the tumoral regions.

The conventional AE-based models are often trained by optimizing per-pixel loss functions that tend to reconstruct blurry images. One potential approach is to modify the objective function in order to improve the quality of the reconstructed images. Therefore, more advanced loss functions such as perceptual loss and style loss can potentially increase the conceptual and textural quality of the generated images. However, integrating these objective functions into the conventional representation learning models such as AE models would degrade their ability to learn the latent characteristics of the healthy anatomies. In other words, such modified models tend to learn a wide range of image-based details and hardly can discriminate normal structures from anomalies. In fact, such fortified objective functions increase the risk of model overfitting with respect to representation learning tasks. However, limiting the convolutional operators with image subregions can regularize the learning process of representation learning models and avoid the overfitting problem. In particular, while the powerful objective function is prone to overfit on the details of anatomical structures, localizing the functionality of convolutional operators can potentially counteract this unwanted behavior. Accordingly, considering the functionality of the $Gconv$ operators, they can be a perfect choice for this problem as they deal with local convolutions instead of ordinary global convolutions. As a result, the representation learning process in this study was turned from conventional AE and GAN-based models into an image inpainting problem. In practice, leveraging the inpainting model with multi-term objective function as an optimization algorithm and $Gconv$ operator as localized convolutional backbones could successfully enforce the model to synthesize the high fidelity realistic-looking medical images while preserving the anatomical constraints regardless of the imaging modality. In practice, integrating the $Gconv$ operator into a U-Net-like architecture optimized by a multi-term objective function that is fortified by the Laplacian loss could successfully improve the quality of the inpainted images regardless of the imaging modality. In particular, the quantified metrics of tables 1, 2, and tables 1.1. and 1.2. in Supplementary Materials verify the superiority of the proposed $Gconv_{Lap}$ model. The learnable soft mask updating procedure of the $Gconv$ operator heuristically updates the invalid pixels, which leads to reconstructing images with more fidelities compared to the hard-gating rules embedded in the $Pconv$ operator. This effect is more evident by comparing the quality of the inpainted images by the three models when multimodal PET-CT images were used (such as figure 5). Besides that, employing an encoder-decoder network architecture with skip connections could propagate the detailed color and textural information to the decoding path and fill the hole boundaries with smooth patterns. In addition, leveraging the objective function with Laplacian loss was a beneficial strategy to preserve the edges and synthesize images with fine-grained details as much as possible. In fact, one of the limitations of the $Pconv$ and $Gconv$ model is to maintain the anatomical constraints, especially in the edges, such as transitions between soft and hard tissues or sharp intensity changes within soft tissues. As can be seen in figure 5, both $Pconv$ and $Gconv$ were unable to reconstruct meaningful anatomical details, while the proposed $Gconv_{Lap}$ model synthesized images with the highest similarity with respect to the original image slices regardless of the level of corruptions applied to the images. Such qualitative comparison is consistent with the numerical values in tables 1 and 2, which point to the advantages of the proposed inpainting model.

The proposed autoinpainting pipeline for tumor segmentation yielded interesting results in the context of unsupervised segmentation. In fact, the segmentation accuracy of the proposed unsupervised pipeline was not far behind the performance of the supervised nnU-Net model when the PET images were analyzed as multimodal or single-modality image data. In specific, the performance of the examined supervised model over the multimodal NSCLC dataset is 4 percent, and for the multimodal HN dataset is 19 percent higher than the proposed unsupervised approach. This can be explained by the fact that the hyper signal intensity in PET images caused by tumoral uptakes facilitates tumor localization. Nevertheless, this should be noted that not all the hyperactive regions are related to cancerous tissues. In

other words, other healthy tissues such as cardiac muscle and lymph nodes uptake high levels of injected FDGs and often appear with hyperintensity patterns. Therefore, localization and segmentation of tumors in PET and multimodal PET-CT images is not a trivial task. In addition, the capabilities of the proposed inpainting model were not limited only to hyperintensity signals of PET images, as the pipeline could detect and inpaint the challenging NSCLC tumors in CT images as well. Highly similar visual attributes of NSCLC tumors with respect to the surrounding soft tissues make them challenging for segmentation models, even for the supervised ones. Nevertheless, the proposed autoinpainting strategy could inpaint and segment the challenging cases and lead to rather promising results. Comparing the segmentation accuracy of $Gconv_{Lap}$ model with $Pconv$ and the ordinary $Gconv$ model within the proposed autoinpainting framework signifies the superiority of the proposed inpainting model (tables 3, 4). In particular, the advantage of $Gconv$ operator over the $Pconv$ module on one side and the ability of the proposed model to preserve the anatomical constraints on the other side lead to inpainting the tumoral regions while retaining the healthy structures intact as much as possible. Therefore, tumoral tissues were removed by the proposed autoinpainting while the healthy structures were not manipulated, which resulted in remarkably fewer false positives. As expected, the tumor segmentation in PET images resulted in more accurate results than in CT images. In specific, while a Dice score of 0.442 was achieved by the $Gconv_{Lap}$ model for NSCLC tumor segmentation in CT images, this metric improved to 0.746 for the PET images on the same dataset. In this domain, it should be noted that the proposed unsupervised model failed to detect the HN tumors in CT images. As can be seen in figure 1.1. in Supplementary Materials, the lack of intensity and the textural contrast between the tumors and nearby soft tissues prevent the autoinpainting approach from recognizing the tumoral regions as anomalies. Such a limitation can be observed in the NSCLC dataset as well when the lung collapses or the tumors appear in the middle of soft tissues (figure 1.3. in Supplementary Materials). Nevertheless, analyzing the PET-CT images together could improve the segmentation accuracy for both the NSCLC and HN tumors.

Comparing the segmentation accuracy of the proposed pipeline against the conventional UAD methods can highlight the great potential of the autoinpainting model. Numerically, the best Dice performance achieved by the examined UAD models is 0.311 for multimodal NSCLC tumors and 0.120 for multimodal HN tumors, which are 0.465 and 0.345 inferior to the corresponding Dice metrics achieved by the proposed $Gconv_{Lap}$ model. In practice, the UAD models failed to reconstruct healthy images from tumoral slices while preserving anatomical structures. In other words, they either removed the tumors and synthesized new images with meaningless anatomical structures or preserved the anatomical structures but could not remove the tumors. It should be emphasized that even when the UAD models managed to remove the tumors, they corrupted many other healthy structures, which resulted in a high rate of false positives. Such results challenge the underlying hypothesis of such UAD models, which aim to model the distribution of healthy data. Carefully examining the images (figure 7 and figures 1.7. and 1.8. in Supplementary Materials) generated by the best performing UAD models such as VAE, ceVAE, and F-AnoGAN, one can deduce that such models reconstructed texture-free images which do not hold meaningful anatomical details. Therefore, the tumors can be detected from the residual images only because of their hyperintensity patterns with respect to the nearby tissues. Such a major limitation of the current UAD methods was highlighted in a recent study [40] in which the authors showed that even with simple image processing techniques such as thresholding, competitive results could be achieved. Other types of UAD methods aim to detect the anomalies but not directly from the residual maps between the original and the reconstructed images [39], [57]; therefore, such models do not aim to produce high-quality anomaly-free images either. In contrast to these methods, the proposed autoinpainting-based anomaly detection pipeline can capture the normal anatomies and generate high-resolution anomaly-free images by retaining fine-grained anatomical details.

The strategy of moving window for detecting and inpainting the tumors is implemented to resemble the way human experts look at different regions of medical images to identify the abnormalities. In fact, once the model detects the candidate regions, it will proceed with the inpainting steps; otherwise, it returns the original image as a healthy one.

Finally, despite the efficacy of the proposed autoinpainting-based UAD model for segmenting tumors in multimodal and single-modal images, there exist some limitations within the proposed pipeline, which will be investigated in our future studies. In particular, the underlying idea of tumor segmentation is based on the pixel-wise differences between the inpainted and original images. This error-prone strategy would be replaced by comparing the learned distribution of healthy anatomies with the distribution of tumoral slices and fine-tuning the trained models to minimize the distribution differences instead of intensity differences. Furthermore, extending the 2D autoinpainting pipeline into a 3D approach

requires the development of a robust 3D inpainting model, which may further improve the accuracy of inpainting by incorporating the volumetric contexts.

While the unsupervised segmentation methods aim to overcome the disadvantages of supervised models, the current UAD models have not been robust enough to yield as accurate results as supervised models. In this study, an inpainting-based UAD method was proposed to segment the NSCLC and HN tumors in multimodal and single-modal images. To the best knowledge of the author, it has been the first attempt to segment such challenging tumors with unsupervised methods. The quantitative results show the potential of the proposed pipeline with superior performance over the conventional UAD models.

5. Acknowledgment

This study was supported by the Swedish Childhood Cancer Foundation (grant no. MT2019-0019), the Swedish innovation agency Vinnova (grant no. 2017-01247), and the Swedish Research Council (VR) (grant no. 2018-04375). We also thank Stockholm Medical Image Laboratory and Education (SMILE) for giving us access to their Nvidia DGX-1 server.

References

- [1] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges," *Journal of Digital Imaging*, vol. 32, no. 4, pp. 582–596, Aug. 2019, doi: 10.1007/S10278-019-00227-X/TABLES/2.
- [2] J. Fournel *et al.*, "Medical image segmentation automatic quality control: A multi-dimensional approach," *Medical Image Analysis*, vol. 74, p. 102213, Dec. 2021, doi: 10.1016/J.MEDIA.2021.102213.
- [3] A. Wadhwa, A. Bhardwaj, and V. Singh Verma, "A review on brain tumor segmentation of MRI images," *Magnetic Resonance Imaging*, vol. 61, pp. 247–259, Sep. 2019, doi: 10.1016/J.MRI.2019.05.043.
- [4] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "A review of deep learning based methods for medical image multi-organ segmentation," *Physica Medica*, vol. 85, pp. 107–122, May 2021, doi: 10.1016/J.EJMP.2021.05.003.
- [5] A. Thakur and R. Shyam Anand, "A local statistics based region growing segmentation method for ultrasound medical images," *statistics*, vol. 12, 2004.
- [6] C. C. Benson, V. L. Lajish, and K. Rajamani, "Brain tumor extraction from MRI brain images using marker based watershed algorithm," *2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015*, pp. 318–323, Sep. 2015, doi: 10.1109/ICACCI.2015.7275628.
- [7] M. Astaraki *et al.*, "Evaluation of localized region-based segmentation algorithms for CT-based delineation of organs at risk in radiotherapy," *Physics and Imaging in Radiation Oncology*, vol. 5, pp. 52–57, Jan. 2018, doi: 10.1016/J.PHRO.2018.02.003.
- [8] S. R. T. J. Goubalan, Y. Goussard, and H. Maaref, "Unsupervised malignant mammographic breast mass segmentation algorithm based on pickard Markov random field," *Proceedings - International Conference on Image Processing, ICIP*, vol. 2016-August, pp. 2653–2657, Aug. 2016, doi: 10.1109/ICIP.2016.7532840.

- [9] X. Chen and L. Pan, "A Survey of Graph Cuts/Graph Search Based Medical Image Segmentation," *IEEE Reviews in Biomedical Engineering*, vol. 11, pp. 112–124, Jan. 2018, doi: 10.1109/RBME.2018.2798701.
- [10] S. Candemir *et al.*, "Atlas-based rib-bone detection in chest X-rays," *Computerized Medical Imaging and Graphics*, vol. 51, pp. 32–39, Jul. 2016, doi: 10.1016/J.COMPAMEDIMAG.2016.04.002.
- [11] N. Chowdhury *et al.*, "Concurrent segmentation of the prostate on MRI and CT via linked statistical shape models for radiotherapy planning," *Medical Physics*, vol. 39, no. 4, pp. 2214–2228, Apr. 2012, doi: 10.1118/1.3696376.
- [12] T. Siriapisith, W. Kusakunniran, and P. Haddawy, "Pyramid graph cut: Integrating intensity and gradient information for grayscale medical image segmentation," *Computers in Biology and Medicine*, vol. 126, p. 103997, Nov. 2020, doi: 10.1016/J.COMPBIOMED.2020.103997.
- [13] Z. Zheng, X. Zhang, H. Xu, W. Liang, S. Zheng, and Y. Shi, "A Unified Level Set Framework Combining Hybrid Algorithms for Liver and Liver Tumor Segmentation in CT Images," *BioMed Research International*, vol. 2018, 2018, doi: 10.1155/2018/3815346.
- [14] H. R. Torres, S. Queirós, P. Morais, B. Oliveira, J. C. Fonseca, and J. L. Vilaça, "Kidney segmentation in ultrasound, magnetic resonance and computed tomography images: A systematic review," *Computer Methods and Programs in Biomedicine*, vol. 157, pp. 49–67, Apr. 2018, doi: 10.1016/J.CMPB.2018.01.014.
- [15] F. Khalifa, A. Soliman, A. Elmaghraby, G. Gimel'farb, and A. El-Baz, "3D Kidney Segmentation from Abdominal Images Using Spatial-Appearance Models," *Computational and Mathematical Methods in Medicine*, vol. 2017, 2017, doi: 10.1155/2017/9818506.
- [16] G. Delpon *et al.*, "Comparison of automated atlas-based segmentation software for postoperative prostate cancer radiotherapy," *Frontiers in Oncology*, vol. 6, no. AUG, p. 178, Aug. 2016, doi: 10.3389/FONC.2016.00178/BIBTEX.
- [17] W. K. H. Wong, L. H. T. Leung, and D. L. W. Kwong, "Evaluation and optimization of the parameters used in multiple-atlas-based segmentation of prostate cancers in radiation therapy," *British Journal of Radiology*, vol. 89, no. 1057, Dec. 2016, doi: 10.1259/BJR.20140732/ASSET/IMAGES/LARGE/BJR.20140732.G002.JPEG.
- [18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.
- [19] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, Apr. 2019, doi: 10.1016/J.MEDIA.2019.01.012.
- [20] M. Islam, V. S. Vibashan, V. J. M. Jose, N. Wijethilake, U. Utkarsh, and H. Ren, "Brain Tumor Segmentation and Survival Prediction Using 3D Attention UNet," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11992 LNCS, pp. 262–272, 2020, doi: 10.1007/978-3-030-46640-4_25.
- [21] Z. Zhang, H. Fu, H. Dai, J. Shen, Y. Pang, and L. Shao, "ET-Net: A Generic Edge-attention Guidance Network for Medical Image Segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11764 LNCS, pp. 442–450, 2019, doi: 10.1007/978-3-030-32239-7_49.
- [22] C. Szegedy *et al.*, "Going Deeper with Convolutions," Sep. 2015. [Online]. Available: <http://arxiv.org/abs/1409.4842>

- [23] H. Cheng, Y. Zhu, and H. Pan, "Modified U-Net block network for lung nodule detection," *Proceedings of 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, ITAIC 2019*, pp. 599–605, May 2019, doi: 10.1109/ITAIC.2019.8785445.
- [24] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, Aug. 2016, Accessed: Feb. 03, 2021. [Online]. Available: <http://arxiv.org/abs/1608.06993>
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2016. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [26] M. Baldeon-Calisto and S. K. Lai-Yuen, "AdaResU-Net: Multiobjective adaptive convolutional neural network for medical image segmentation," *Neurocomputing*, vol. 392, pp. 325–340, Jun. 2020, doi: 10.1016/J.NEUCOM.2019.01.110.
- [27] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM U-net with densely connected convolutions," *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, pp. 406–415, Oct. 2019, doi: 10.1109/ICCVW.2019.00052.
- [28] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11045 LNCS, pp. 3–11, 2018, doi: 10.1007/978-3-030-00889-5_1.
- [29] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, "SegAN: Adversarial Network with Multi-scale L1 Loss for Medical Image Segmentation," *Neuroinformatics 2018 16:3*, vol. 16, no. 3, pp. 383–392, May 2018, doi: 10.1007/S12021-018-9377-X.
- [30] S. Hansen, S. Gautam, R. Jenssen, and M. Kampffmeyer, "Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels," *Medical Image Analysis*, vol. 78, p. 102385, May 2022, doi: 10.1016/J.MEDIA.2022.102385.
- [31] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Medical Image Analysis*, vol. 54, pp. 30–44, May 2019, doi: 10.1016/J.MEDIA.2019.01.010.
- [32] C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni, "Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study," *Medical Image Analysis*, vol. 69, p. 101952, Apr. 2021, doi: 10.1016/J.MEDIA.2020.101952.
- [33] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Sep. 2019, vol. 11383 LNCS, pp. 161–169. doi: 10.1007/978-3-030-11723-8_16.
- [34] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein, "Unsupervised anomaly localization using variational auto-encoders," in *Lecture Notes in Computer Science*, Oct. 2019, vol. 11767 LNCS, pp. 289–297. doi: 10.1007/978-3-030-32251-9_32.
- [35] Y. Tian *et al.*, "Constrained Contrastive Distribution Learning for Unsupervised Anomaly Detection and Localisation in Medical Images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12905 LNCS, pp. 128–140, Sep. 2021, doi: 10.1007/978-3-030-87240-3_13.

- [36] C. Baur, R. Graf, B. Wiestler, S. Albarqouni, and N. Navab, “SteGANomaly: Inhibiting CycleGAN Steganography for Unsupervised Anomaly Detection in Brain MRI,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12262 LNCS, pp. 718–727, Oct. 2020, doi: 10.1007/978-3-030-59713-9_69.
- [37] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, “Scale-Space Autoencoders for Unsupervised Anomaly Segmentation in Brain MRI,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12264 LNCS, pp. 552–561, Oct. 2020, doi: 10.1007/978-3-030-59719-1_54.
- [38] S. Naval Marimont and G. Tarroni, “Implicit Field Learning for Unsupervised Anomaly Detection in Medical Images,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12902 LNCS, pp. 189–198, Sep. 2021, doi: 10.1007/978-3-030-87196-3_18.
- [39] R. Dey and Y. Hong, “ASC-Net: Adversarial-Based Selective Network for Unsupervised Anomaly Segmentation,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12905 LNCS, pp. 236–247, 2021, doi: 10.1007/978-3-030-87240-3_23.
- [40] F. Meissen, G. Kaissis, and D. Rueckert, “Challenging Current Semi-Supervised Anomaly Segmentation Methods for Brain MRI,” Sep. 2021, doi: 10.48550/arxiv.2109.06023.
- [41] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, “Image Inpainting: A Review,” *Neural Processing Letters 2019 51:2*, vol. 51, no. 2, pp. 2007–2028, Dec. 2019, doi: 10.1007/S11063-019-10163-0.
- [42] J. Jam, C. Kendrick, K. Walker, V. Drouard, J. G. S. Hsu, and M. H. Yap, “A comprehensive review of past and present image inpainting methods,” *Computer Vision and Image Understanding*, vol. 203, p. 103147, Feb. 2021, doi: 10.1016/J.CVIU.2020.103147.
- [43] G. Liu, F. A. Reda, K. J. Shih, T. C. Wang, A. Tao, and B. Catanzaro, “Image Inpainting for Irregular Holes Using Partial Convolutions,” in *Lecture Notes in Computer Science*, Sep. 2018, vol. 11215 LNCS, pp. 89–105. doi: 10.1007/978-3-030-01252-6_6.
- [44] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, “Free-form image inpainting with gated convolution,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, pp. 4470–4479, Oct. 2019, doi: 10.1109/ICCV.2019.00457.
- [45] N. Wang, Y. Zhang, and L. Zhang, “Dynamic Selection Network for Image Inpainting,” *IEEE Transactions on Image Processing*, vol. 30, pp. 1784–1798, 2021, doi: 10.1109/TIP.2020.3048629.
- [46] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, and J. Liao, “PD-GAN: Probabilistic Diverse GAN for Image Inpainting,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9367–9376, 2021, doi: 10.1109/CVPR46437.2021.00925.
- [47] C. Wang, H. Frimmel, and Ö. Smedby, “Fast level-set based image segmentation using coherent propagation,” *Medical Physics*, vol. 41, no. 7, p. 073501, Jul. 2014, doi: 10.1118/1.4881315.
- [48] M. Vallières *et al.*, “Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer,” *Scientific Reports 2017 7:1*, vol. 7, no. 1, pp. 1–14, Aug. 2017, doi: 10.1038/s41598-017-10371-5.
- [49] M. Vallières *et al.*, “Data from Head-Neck-PET-CT,” *The Cancer Imaging Archive*, 2017.
- [50] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv:1409.1556*, Sep. 2015, [Online]. Available: <http://arxiv.org/abs/1409.1556>

- [51] M. Gruber, “Image Inpainting for Irregular Holes Using Partial Convolutions Keras Implementation,” 2019. <https://github.com/MathiasGruber/PConv-Keras>
- [52] A. P. Harrison, Z. Xu, K. George, L. Lu, R. M. Summers, and D. J. Mollura, “Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10435 LNCS, pp. 621–629, 2017, doi: 10.1007/978-3-319-66179-7_71/TABLES/1.
- [53] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods* 2020 18:2, vol. 18, no. 2, pp. 203–211, Dec. 2020, doi: 10.1038/s41592-020-01008-z.
- [54] “Autoencoders for Unsupervised Anomaly Segmentation.” [Online]. Available: https://github.com/StefanDenn3r/Unsupervised_Anomaly_Detection_Brain_MRI
- [55] V. Andrearczyk *et al.*, “Automatic Segmentation of Head and Neck Tumors and Nodal Metastases in PET-CT scans,” in *Medical Imaging with Deep Learning*, 2020, pp. 33–43. [Online]. Available: <https://proceedings.mlr.press/v121/andrearczyk20a.html>
- [56] W. H. L. Pinaya *et al.*, “Unsupervised Brain Imaging 3D Anomaly Detection and Segmentation with Transformers,” *Medical Image Analysis*, p. 102475, May 2022, doi: 10.1016/J.MEDIA.2022.102475.
- [57] K. M. van Hespen, J. J. M. Zwanenburg, J. W. Dankbaar, M. I. Geerlings, J. Hendrikse, and H. J. Kuijf, “An anomaly detection approach to identify chronic brain infarcts on MRI,” *Scientific Reports* 2021 11:1, vol. 11, no. 1, pp. 1–10, Apr. 2021, doi: 10.1038/s41598-021-87013-4.

Supplementary Materials

Unsupervised Tumor Segmentation

Mehdi Astaraki, Francesca De Benetti, Yousef Yeganeh, Iuliana Toma-Dasu, Örjan Smedby, Chunliang Wang, Nassir Navab, Thomas Wendler

1) Supplementary Tables and Figures

Table 1.1. The numerical comparison between the performance of inpainting models trained and tested on NSCLC single modality images

Model-Data	Quantitative Metrics ($\mu \pm \sigma$)		
	<i>MSE</i>	<i>PSNR</i>	<i>SSIM</i>
Pconv-CT	109.883±63.296	28.572±2.966	0.921±0.034
Gconv-CT	67.955±51.888	31.770±5.161	0.943±0.033
Gconv _{Lap} -CT	62.061±47.406	32.096±5.011	0.949±0.030
Pconv-PET	16.336±17.698	38.382±4.494	0.980±0.012
Gconv-PET	16.040±21.927	39.351±6.115	0.981±0.013
Gconv _{Lap} -PET	15.668±19.547	39.312±6.112	0.982±0.012

Table 1.2. The numerical comparison between the performance of inpainting models trained and tested on HN single modality images

Model-Data	Quantitative Metrics ($\mu \pm \sigma$)		
	<i>MSE</i>	<i>PSNR</i>	<i>SSIM</i>
Pconv-CT	9.533±9.009	40.429±5.066	0.985±0.013
Gconv-CT	6.815±6.944	42.757±6.414	0.989±0.011
Gconv _{Lap} -CT	5.542±6.686	44.072±6.917	0.991±0.009
Pconv-PET	6.073±7.484	47.142±7.363	0.993±0.007
Gconv-PET	5.852±17.400	48.533±6.909	0.993±0.008
Gconv _{Lap} -PET	3.413±10.395	50.123±9.312	0.995±0.006

Table 1.3. Statistical comparison of image quality metrics between the three inpainting models using the Wilcoxon signed rank test

<i>Data-Tumor</i>	<i>Models</i>	p-value		
		<i>MSE</i>	<i>PSNR</i>	<i>SSIM</i>
CT-NSCLC	Pconv vs. Geonv	< 0.0001	< 0.0001	< 0.0001
	Pconv vs. Gconv _{Lap}	< 0.0001	< 0.0001	< 0.0001
	Gconv vs. Gconv _{Lap}	0.665	0.430	0.0007
PET-NSCLC	Pconv vs. Geonv	0.516	0.005	0.0005
	Pconv vs. Gconv _{Lap}	< 0.0001	< 0.0001	< 0.0001
	Gconv vs. Gconv _{Lap}	< 0.0001	< 0.0001	< 0.0001
Multi-NSCLC	Pconv vs. Geonv	< 0.0001	< 0.0001	< 0.0001
	Pconv vs. Gconv _{Lap}	< 0.0001	< 0.0001	< 0.0001
	Gconv vs. Gconv _{Lap}	0.001	0.0139	< 0.0001
CT-HN	Pconv vs. Geonv	< 0.0001	< 0.0001	< 0.0001
	Pconv vs. Gconv _{Lap}	< 0.0001	< 0.0001	< 0.0001
	Gconv vs. Gconv _{Lap}	< 0.0001	< 0.0001	< 0.0001
PET-HN	Pconv vs. Geonv	< 0.0001	< 0.0001	< 0.0001
	Pconv vs. Gconv _{Lap}	< 0.0001	< 0.0001	< 0.0001
	Gconv vs. Gconv _{Lap}	< 0.0001	< 0.0001	< 0.0001
Multi-HN	Pconv vs. Geonv	< 0.0001	< 0.0001	< 0.0001
	Pconv vs. Gconv _{Lap}	< 0.0001	< 0.0001	< 0.0001
	Gconv vs. Gconv _{Lap}	< 0.0001	< 0.0001	< 0.0001

Supplementary Materials

Table 1.4. Statistical comparison of the achieved Dice scores by the autoinpainting method applied on the three inpainting models

<i>Data-Tumor</i>	<i>Models</i>	<i>p-value</i>
CT-NSCLC	Pconv vs. Gconv	< 0.0001
	Pconv vs. Gconv _{Lap}	< 0.0001
	Gconv vs. Gconv _{Lap}	< 0.0001
PET-NSCLC	Pconv vs. Gconv	0.083
	Pconv vs. Gconv _{Lap}	< 0.0001
	Gconv vs. Gconv _{Lap}	< 0.0001
Multi-NSCLC	Pconv vs. Gconv	< 0.0001
	Pconv vs. Gconv _{Lap}	< 0.0001
	Gconv vs. Gconv _{Lap}	< 0.0001
PET-HN	Pconv vs. Gconv	0.354
	Pconv vs. Gconv _{Lap}	< 0.0001
	Gconv vs. Gconv _{Lap}	< 0.0001
Multi-HN	Pconv vs. Gconv	< 0.0001
	Pconv vs. Gconv _{Lap}	< 0.0001
	Gconv vs. Gconv _{Lap}	0.019

Table 1.5. Segmentation accuracy of unsupervised anomaly detection models on CT images of NSCLC tumors

Model	Quantitative Metrics ($\mu \pm \sigma$)			
	[Dice]	[Precision]	[Recal]	Dice
dAE	0.225±0.123	0.302±0.126	0.992±0.009	0.200±0.127
sAE	0.022±0.013	0.134±0.062	0.950±0.046	0.013±0.008
ceAE	0.221±0.123	0.295±0.122	0.991±0.018	0.2018±0.1419
VAE	0.210±0.114	0.287±0.107	0.990±0.020	0.182±0.121
ceVAE	0.102±0.059	0.220±0.118	0.979±0.044	0.072±0.050
GMVAE	0.013±0.007	0.519±0.108	0.721±0.018	0.006±0.003
F-AnoGAN	0.084±0.071	0.376±0.183	0.960±0.034	0.051±0.047
AAE	0.083±0.072	0.422±0.162	0.938±0.060	0.053±0.055

Table 1.6. Segmentation accuracy of unsupervised anomaly detection models on PET images of NSCLC tumors

Model	Quantitative Metrics ($\mu \pm \sigma$)			
	[Dice]	[Precision]	[Recal]	Dice
dAE	0.583±0.160	0.533±0.159	0.998±0.000	0.658±0.184
sAE	0.623±0.256	0.677±0.201	0.996±0.009	0.634±0.275
ceAE	0.564±0.142	0.515±0.145	0.998±0.000	0.641±0.162
VAE	0.660±0.173	0.634±0.174	0.999±0.000	0.702±0.190
ceVAE	0.478±0.161	0.473±0.155	0.997±0.002	0.525±0.205
GMVAE	0.026±0.017	0.653±0.094	0.814±0.030	0.013±0.009
F-AnoGAN	0.633±0.199	0.593±0.197	0.999±0.000	0.695±0.215
AAE	0.593±0.194	0.579±0.199	0.999±0.000	0.628±0.192

Table 1.7. Segmentation accuracy of unsupervised anomaly detection models on PET images of HN tumors

Model	Quantitative Metrics ($\mu \pm \sigma$)			
	[Dice]	[Precision]	[Recal]	Dice
dAE	0.390±0.157	0.477±0.214	0.359±0.138	0.330±0.104
sAE	0.158±0.109	0.371±0.289	0.149±0.137	0.137±0.038
ceAE	0.402±0.153	0.481±0.211	0.376±0.129	0.337±0.107
VAE	0.437±0.207	0.471±0.240	0.455±0.174	0.391±0.113

Supplementary Materials

ceVAE	0.271 ± 0.180	0.469 ± 0.308	0.339 ± 0.229	0.260 ± 0.050
GMVAE	0.064 ± 0.037	0.033 ± 0.021	0.756 ± 0.120	0.064 ± 0.011
F-AnoGAN	0.426 ± 0.203	0.461 ± 0.239	0.417 ± 0.171	0.388 ± 0.108
AAE	0.413 ± 0.200	0.406 ± 0.237	0.453 ± 0.157	0.361 ± 0.116

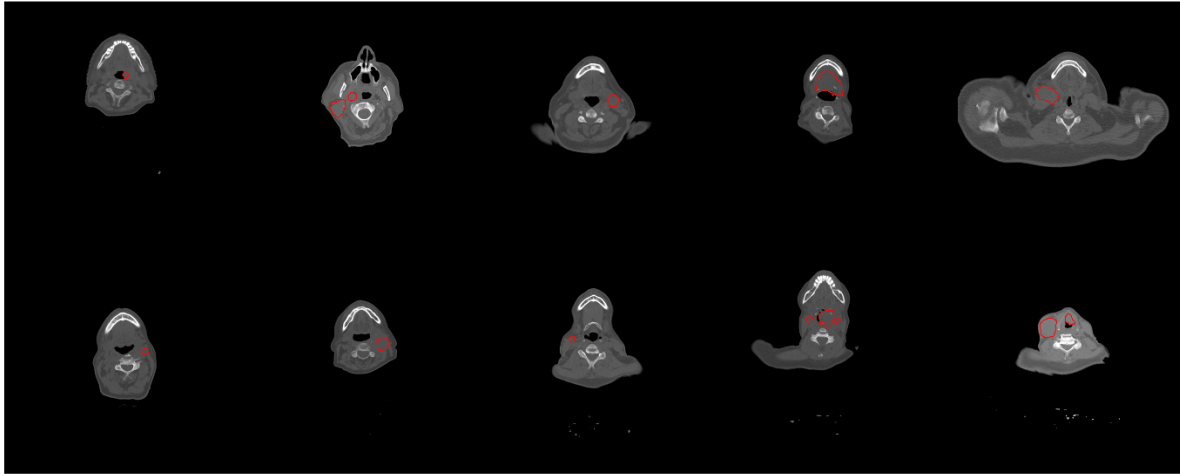


Figure 1.1. Examples of the appearance of HN tumors, highlighted in red contours, in CT images. The presence of tumors among the densely connected soft tissues with a similar range of Hounsfield values makes the segmentation of HN tumors a challenging problem.

Supplementary Materials

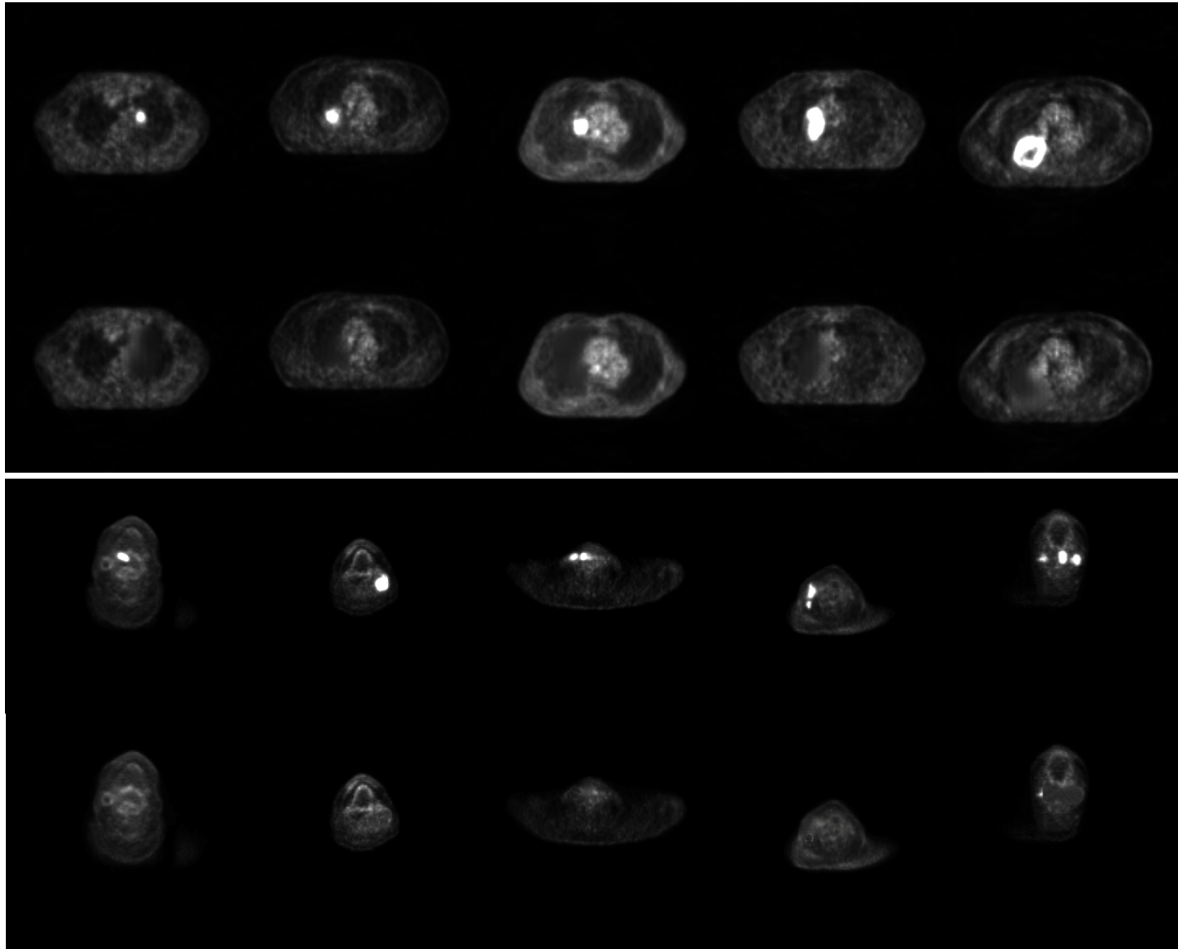


Figure 1.2. Visualization of the segmentation performance of the proposed autoinpainting pipeline for single modality PET images. For each of the NSCLC and HN images, the first row shows the original tumoral slices, and the second row depicts the result of the proposed autoinpainting model where the tumors were replaced by healthy tissues and fake tumor-free images were synthesized.

Supplementary Materials

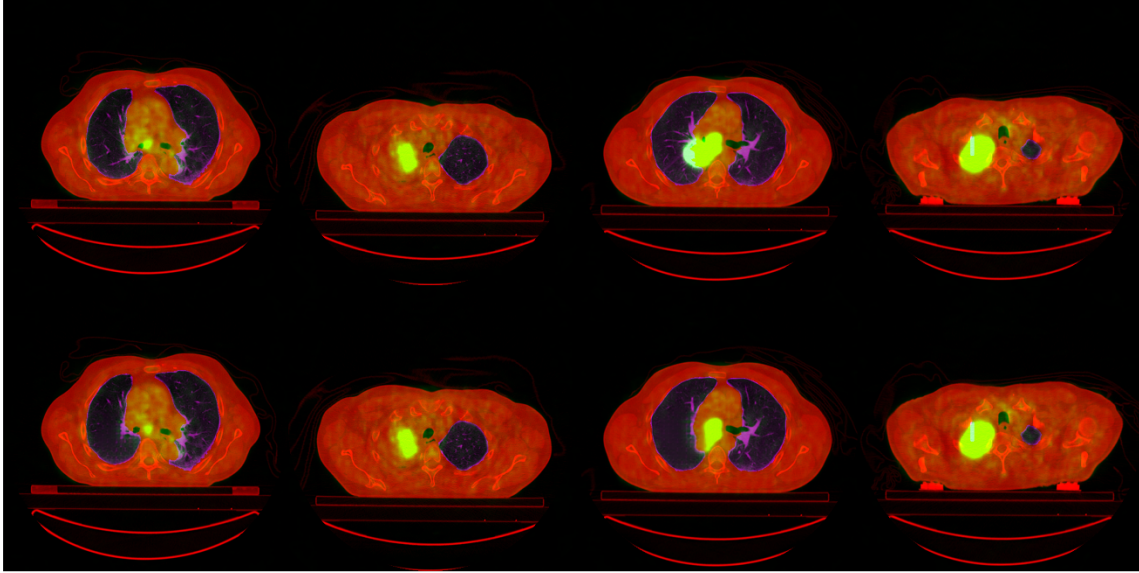


Figure 1.3. Examples of challenging NSCLC tumors in multimodal PET-CT images where the proposed autoinpainting pipeline either failed to detect the tumors or could only partially remove the tumors. The first row shows the original tumoral slices, and the second row depicts the results of autoinpainting pipeline

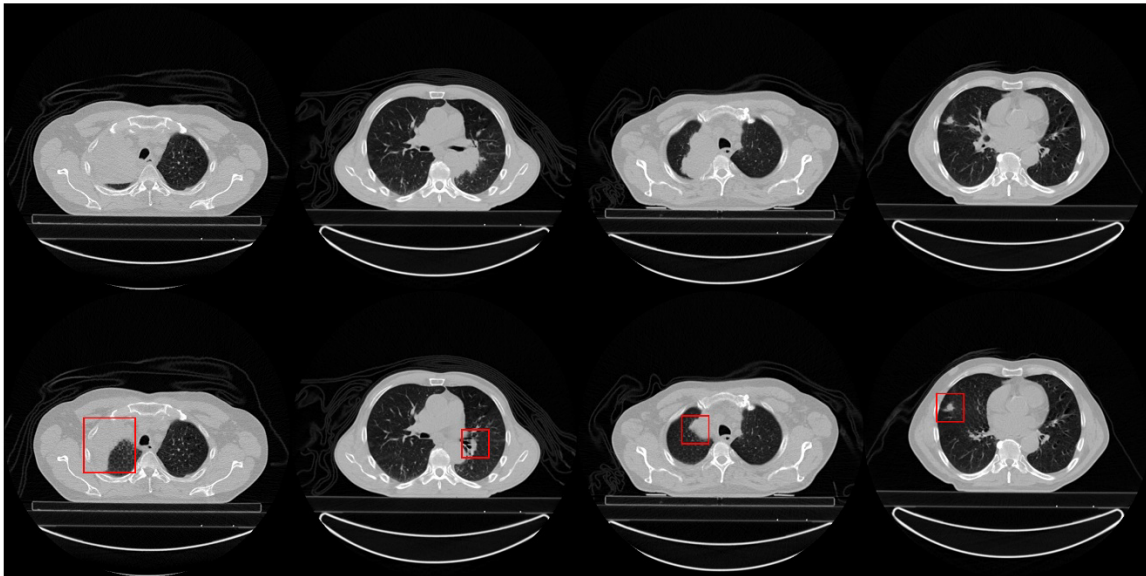


Figure 1.4. Examples of challenging NSCLC tumors in CT images where the proposed autoinpainting pipeline could partially remove the tumors (the first three examples from left to right), or it mistakenly removes the healthy structures (the last image). The first row shows the original tumoral slices, and the second row depicts the results of autoinpainting pipeline. The parts of the tumors which were not removed are highlighted in red bounding boxes.

Supplementary Materials

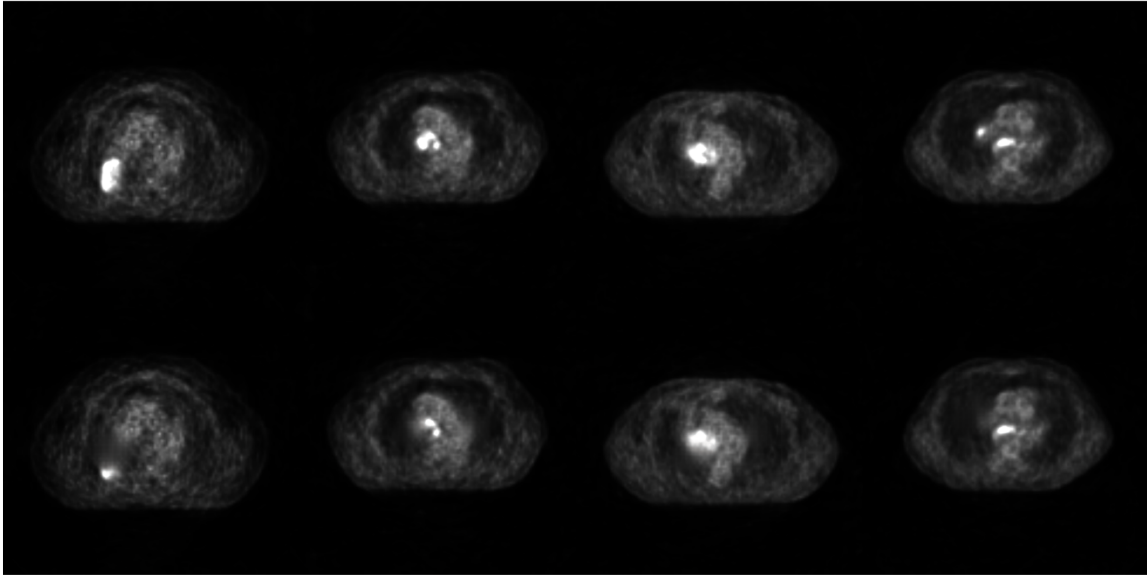


Figure 1.5. Examples of PET images where the tumors were partially removed by the proposed autoinpainting pipeline. The first row shows the original tumoral slices, and the second row depicts the results of autoinpainting pipeline.

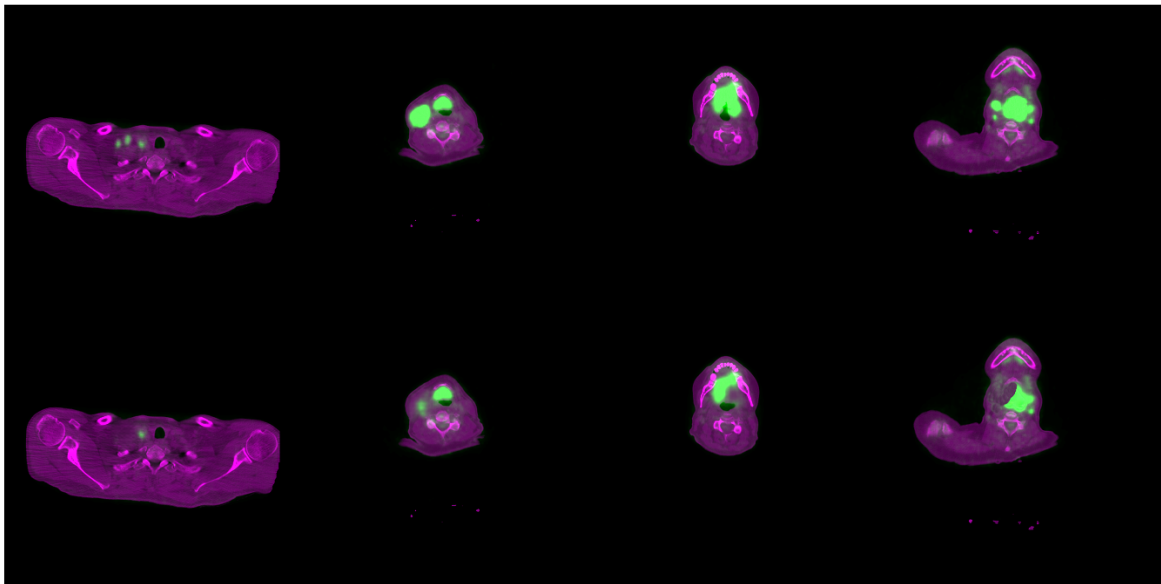


Figure 1.6. Examples of HN tumors in PET-CT images where the autoinpainting pipeline failed to completely substitute the tumoral regions with healthy anatomies. The first row shows the original tumoral slices, and the second row depicts the results of autoinpainting pipeline.

Supplementary Materials

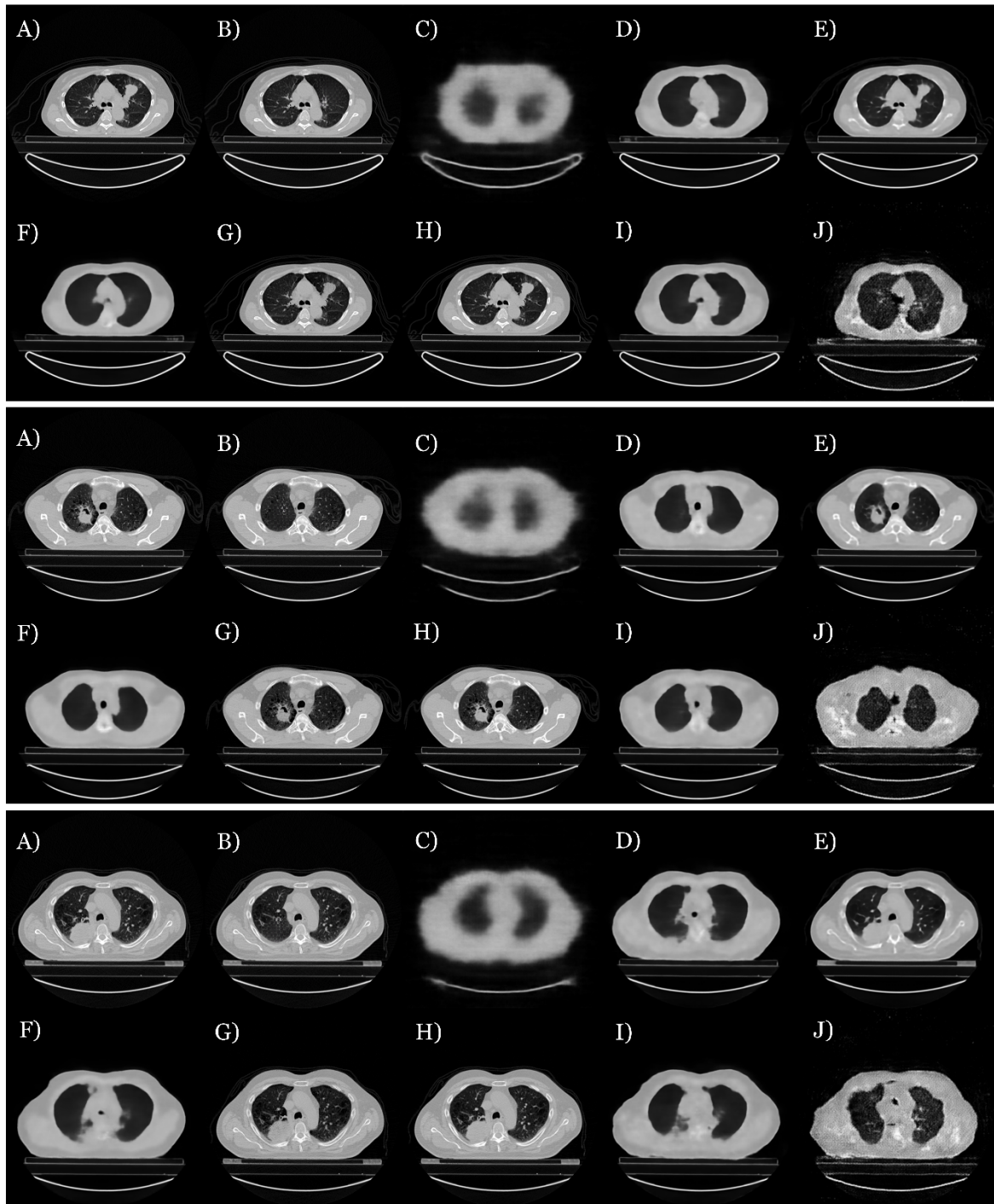


Figure 1.7. Qualitative comparison between the performance of the proposed autoinpainting pipeline and the other eight UAD models to learn the appearance of healthy anatomy in CT images and reconstruct NSCLC tumor-free images. Each three set of images include: A) original tumor slice, B) proposed autoinpainting image, C) adversarial autoencoder result, D) dense autoencoder result, E) spatial autoencoder result, F) variational autoencoder result, G) context-encoding variational autoencoder result, H) Gaussian mixture variational autoencoder result, I) context-encoding autoencoder result, and J) Fast-Anomaly GAN.

Supplementary Materials

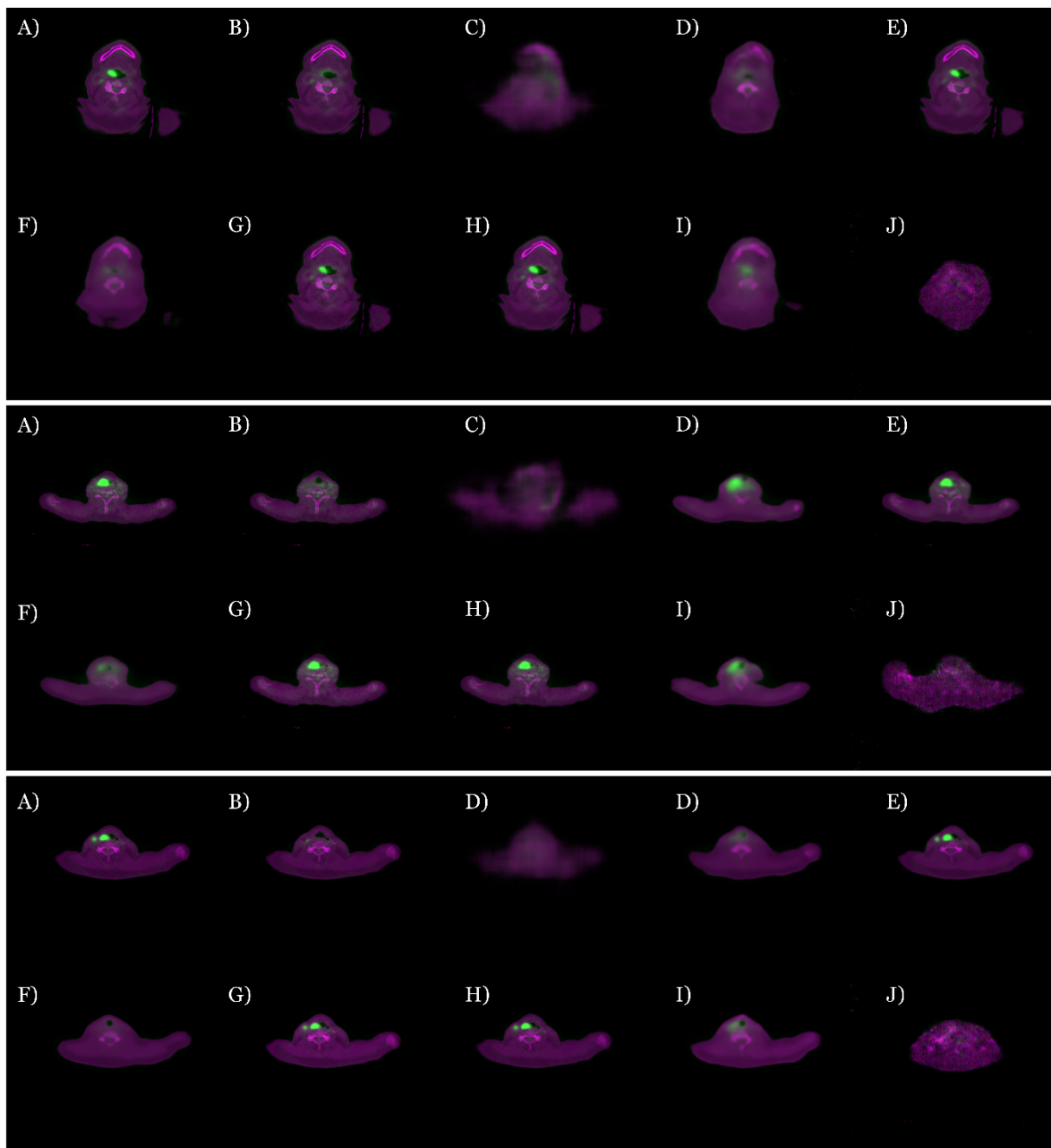


Figure 1.8. Qualitative comparison between the performance of the proposed autoinpainting pipeline and the other eight UAD models to learn the appearance of healthy anatomy in PET-CT images and reconstruct HN tumor-free images. Each three set of images include: A) original tumor slice, B) proposed autoinpainted image, C) adversarial autoencoder result, D) dense autoencoder result, E) spatial autoencoder result, F) variational autoencoder result, G) context-encoding variational autoencoder result, H) Gaussian mixture variational autoencoder result, I) context-encoding autoencoder result, and J) Fast-Anomaly GAN.

Supplementary Materials

2) Ablation Study

2.1) Objective Function

The original objective function which was used for the Pconv model is:

$$\mathcal{L}_{total} = 1\mathcal{L}_{valid} + 6\mathcal{L}_{hole} + 0.05\mathcal{L}_{perceptual} + 120(\mathcal{L}_{style_{out}} + \mathcal{L}_{style_{comp}}) + 0.1\mathcal{L}_{tv}$$

In this study, the general framework of the objective function developed for the proposed Gconv_{Lap} model is:

$$\mathcal{L}_{total} = C_1\mathcal{L}_{valid} + C_2\mathcal{L}_{hole} + C_3\mathcal{L}_{perceptual} + C_4(\mathcal{L}_{style_{out}} + \mathcal{L}_{style_{comp}}) + C_5\mathcal{L}_{tv} + C_6\mathcal{L}_{lap}$$

And the specific weighting coefficients of different terms are:

$$\mathcal{L}_{total} = 30\mathcal{L}_{valid} + 240\mathcal{L}_{hole} + 0.2\mathcal{L}_{perceptual} + 0.05(\mathcal{L}_{style_{out}} + \mathcal{L}_{style_{comp}}) + 250\mathcal{L}_{tv} + 20\mathcal{L}_{lap}$$

An extensive ablation study was conducted to determine the proper coefficients (C_i) of the above-mentioned loss function. In the following, a summary of the results is reported.

The coefficient of different terms of the objective function was, first, determined based on the idea that all the terms should contribute to the optimization equally so that none of them could outweigh the others. Accordingly, the following loss function was initiated:

$$\mathcal{L}_{total} = 20\mathcal{L}_{valid} + 120\mathcal{L}_{hole} + 0.1\mathcal{L}_{perceptual} + 0.015(\mathcal{L}_{style_{out}} + \mathcal{L}_{style_{comp}}) + 250\mathcal{L}_{tv} + 20\mathcal{L}_{lap}$$

Then, for each of the C_i a set of values were set around the initial coefficients, and the model was trained 100 epochs for each setting independently. The image similarity metrics over a test set of 2000 images were quantified to determine the optimal values.

Table 2.1 shows the results of comparing the numerical metrics between the original Pconv model and the initial guess for the coefficients of the proposed model:

Table 2.1. The numerical comparison between the proposed model with the initial guess of loss term coefficients and the original Pconv model

Experiment Name	Quantitative Metrics ($\mu \pm \sigma$)		
	MSE	PSNR	SSIM
Original Pconv coef.	117.460±59.841	28.004±2.277	0.884±0.021
All coef. set to 1	78.775±47.747	30.046±2.925	0.930±0.022
Initial guess without Laplasian loss	64.009±44.861	31.317±3.598	0.943±0.021
Initial guess with Laplassian loss	61.922±44.460	31.471±3.586	0.946±0.020

Table 2.2 shows the effect of changing the coefficient C_1 i.e.:

$$\mathcal{L}_{total} = C_1\mathcal{L}_{valid} + 120\mathcal{L}_{hole} + 0.1\mathcal{L}_{perceptual} + 0.015(\mathcal{L}_{style_{out}} + 2\mathcal{L}_{style_{comp}}) + 250\mathcal{L}_{tv} + 20\mathcal{L}_{lap}$$

Table 2.2. The impact of changing the C_1 coefficients on the model performance. The candidate values for further experiments are marked in bold.

C_1 variable	Quantitative Metrics ($\mu \pm \sigma$)		
	MSE	PSNR	SSIM
0	61.666±39.895	31.613±3.478	0.952±0.020
1	61.009±47.244	31.751±3.928	0.951±0.019
10	66.820±45.263	31.025±3.438	0.952±0.020
20	71.907±50.929	30.834±3.634	0.945±0.021
30	62.626±40.826	31.257±3.361	0.947±0.020

Supplementary Materials

50	65.056±43.748	31.140±3.431	0.949±0.019
----	---------------	--------------	-------------

Table 2.3. shows the effect of changing the coefficient C_2 i.e.:

$$\mathcal{L}_{total} = 20\mathcal{L}_{valid} + C_2\mathcal{L}_{hole} + 0.1\mathcal{L}_{perceptual} + 0.015(\mathcal{L}_{style_{out}} + 2\mathcal{L}_{style_{comp}}) + 250\mathcal{L}_{tv} + 20\mathcal{L}_{lap}$$

Table 2.3. The impact of changing the C_2 coefficients on the model performance. The candidate values for further experiments are marked in bold.

C_2 variable	Quantitative Metrics ($\mu \pm \sigma$)		
	MSE	PSNR	SSIM
0	61.368±43.172	31.430±3.449	0.951±0.019
1	63.964±44.135	31.246±3.473	0.950±0.021
10	61.220±45.115	31.573±3.605	0.949±0.020
60	59.938±40.708	31.434±3.290	0.948±0.020
120	72.804±46.925	30.600±3.446	0.946±0.020
240	58.955±41.414	31.624±3.484	0.950±0.020
500	62.696±40.600	31.229±3.545	0.949±0.019

Table 2.4. shows the effect of changing the coefficient C_3 i.e.:

$$\mathcal{L}_{total} = 20\mathcal{L}_{valid} + 120\mathcal{L}_{hole} + C_3\mathcal{L}_{perceptual} + 0.015(\mathcal{L}_{style_{out}} + 2\mathcal{L}_{style_{comp}}) + 250\mathcal{L}_{tv} + 20\mathcal{L}_{lap}$$

Table 2.4. The impact of changing the C_3 coefficients on the model performance. The candidate values for further experiments are marked in bold.

C_3 variable	Quantitative Metrics ($\mu \pm \sigma$)		
	MSE	PSNR	SSIM
0	70.880±46.704	30.678±3.244	0.934±0.021
0.05	61.982±44.739	31.468±3.595	0.947±0.019
0.1	62.816±43.824	31.368±3.542	0.950±0.021
0.2	58.224±44.404	31.950±3.935	0.953±0.021

Table 2.5. shows the effect of changing the coefficient C_4 i.e.:

$$\mathcal{L}_{total} = 20\mathcal{L}_{valid} + 120\mathcal{L}_{hole} + 0.1\mathcal{L}_{perceptual} + C_4(\mathcal{L}_{style_{out}} + \mathcal{L}_{style_{comp}}) + 250\mathcal{L}_{tv} + 20\mathcal{L}_{lap}$$

Table 2.5. The impact of changing the C_4 coefficients on the model performance. The candidate values for further experiments are marked in bold.

C_4 variable	Quantitative Metrics ($\mu \pm \sigma$)		
	MSE	PSNR	SSIM
0	66.653±39.837	31.003±3.487	0.945±0.019
0.05	67.448±45.412	30.956±3.384	0.949±0.020
0.1	66.550±44.078	31.056±3.481	0.943±0.021
0.2	80.045±53.061	30.166±3.272	0.942±0.021
1	83.138±53.245	29.926±3.127	0.928±0.022
10	143.102±62.211	26.998±1.969	0.874±0.021
60	138.358±67.401	27.279±2.292	0.894±0.021

Table 2.6 shows the effect of changing the coefficient C_5 i.e.:

$$\mathcal{L}_{total} = 20\mathcal{L}_{valid} + 120\mathcal{L}_{hole} + 0.1\mathcal{L}_{perceptual} + 0.015(\mathcal{L}_{style_{out}} + \mathcal{L}_{style_{comp}}) + C_5\mathcal{L}_{tv} + 20\mathcal{L}_{lap}$$

Supplementary Materials

Table 2.6. The impact of changing the C_5 coefficients on the model performance. The candidate values for further experiments are marked in bold.

C_5 variable	Quantitative Metrics ($\mu \pm \sigma$)		
	MSE	PSNR	SSIM
0	65.480 \pm 38.833	31.203 \pm 3.218	0.941 \pm 0.018
0.1	61.467 \pm 41.109	31.374 \pm 3.420	0.950 \pm 0.020
1	64.616 \pm 43.502	31.052 \pm 3.167	0.945 \pm 0.019
10	65.461 \pm 50.113	31.419 \pm 3.835	0.949 \pm 0.021
100	59.500\pm41.086	31.639\pm3.738	0.949\pm0.021
250	58.340\pm41.086	31.769\pm3.738	0.949\pm0.019
500	65.933 \pm 48.985	31.373 \pm 3.926	0.948 \pm 0.020

Finally, Table 2.7. shows the effect of changing the coefficient C_6 i.e.:

$$\mathcal{L}_{total} = 20\mathcal{L}_{valid} + 120\mathcal{L}_{hole} + 0.1\mathcal{L}_{perceptual} + 0.015(\mathcal{L}_{style_{out}} + \mathcal{L}_{style_{comp}}) + 250\mathcal{L}_{tv} + C_6\mathcal{L}_{lap}$$

Table 2.7. The impact of changing the C_6 coefficients on the model performance. The candidate values for further experiments are marked in bold.

C_6 variable	Quantitative Metrics ($\mu \pm \sigma$)		
	MSE	PSNR	SSIM
0	65.924 \pm 42.228	31.213 \pm 3.533	0.940 \pm 0.020
1	63.426\pm44.382	31.289\pm3.453	0.946\pm0.019
10	68.367 \pm 41.182	30.699 \pm 3.061	0.943 \pm 0.018
20	63.121\pm45.666	31.394\pm3.585	0.950\pm0.020
50	66.587 \pm 44.098	30.945 \pm 3.246	0.947 \pm 0.019
100	64.476 \pm 40.940	31.268 \pm 3.407	0.948 \pm 0.020

From the described conducted experiments, the final weight candidates for each of the loss terms will be:

$$\mathcal{L}_{total} = \begin{cases} 0 \\ 1 \\ 30 \end{cases} \mathcal{L}_{valid} + \begin{cases} 60 \\ 240 \end{cases} \mathcal{L}_{hole} + 0.2\mathcal{L}_{perceptual} + \begin{cases} 0 \\ 0.05 \\ 0.1 \end{cases} (\mathcal{L}_{style_{out}} + \mathcal{L}_{style_{comp}}) + \begin{cases} 100 \\ 250 \end{cases} \mathcal{L}_{tv} + \begin{cases} 1 \\ 20 \end{cases} \mathcal{L}_{lap}$$

Therefore, to determine the optimal values of the weight coefficients, a set of independent experiments was examined by setting the different combinations of the weight candidates (Table 2.8).

Table 2.8. The impact of different combinations of weighting coefficients on the model performance. The final candidate values are marked in bold.

Experiment	Quantitative Metrics ($\mu \pm \sigma$)		
	MSE	PSNR	SSIM
$0\mathcal{L}_{valid}+60\mathcal{L}_{hole}+0.2\mathcal{L}_{per.}+0(\mathcal{L}_{style_{out}}+\mathcal{L}_{style_{comp}})+100\mathcal{L}_{tv}+1\mathcal{L}_{lap}$	56.235 \pm 36.519	31.749 \pm 3.614	0.954 \pm 0.027
$1\mathcal{L}_{valid}+60\mathcal{L}_{hole}+0.2\mathcal{L}_{per.}+0(\mathcal{L}_{style_{out}}+\mathcal{L}_{style_{comp}})+100\mathcal{L}_{tv}+1\mathcal{L}_{lap}$	53.175 \pm 29.341	31.837 \pm 4.267	0.957 \pm 0.031
$1\mathcal{L}_{valid}+60\mathcal{L}_{hole}+0.2\mathcal{L}_{per.}+0.1(\mathcal{L}_{style_{out}}+\mathcal{L}_{style_{comp}})+100\mathcal{L}_{tv}+1\mathcal{L}_{lap}$	48.591 \pm 38.947	32.830 \pm 4.101	0.960 \pm 0.019
$1\mathcal{L}_{valid}+60\mathcal{L}_{hole}+0.2\mathcal{L}_{per.}+0.1(\mathcal{L}_{style_{out}}+\mathcal{L}_{style_{comp}})+100\mathcal{L}_{tv}+20\mathcal{L}_{lap}$	48.413 \pm 38.700	32.959 \pm 4.339	0.961 \pm 0.020
$1\mathcal{L}_{valid}+240\mathcal{L}_{hole}+0.2\mathcal{L}_{per.}+0.1(\mathcal{L}_{style_{out}}+\mathcal{L}_{style_{comp}})+100\mathcal{L}_{tv}+20\mathcal{L}_{lap}$	52.182 \pm 29.381	31.862 \pm 3.962	0.961 \pm 0.041
$30\mathcal{L}_{valid}+60\mathcal{L}_{hole}+0.2\mathcal{L}_{per.}+0.1(\mathcal{L}_{style_{out}}+\mathcal{L}_{style_{comp}})+100\mathcal{L}_{tv}+20\mathcal{L}_{lap}$	50.439 \pm 36.468	32.463 \pm 3.830	0.962 \pm 0.019
$30\mathcal{L}_{valid}+240\mathcal{L}_{hole}+0.2\mathcal{L}_{per.}+0.1(\mathcal{L}_{style_{out}}+\mathcal{L}_{style_{comp}})+100\mathcal{L}_{tv}+20\mathcal{L}_{lap}$	47.938 \pm 36.107	32.768 \pm 3.960	0.963 \pm 0.018
$30\mathcal{L}_{valid}+60\mathcal{L}_{hole}+0.2\mathcal{L}_{per.}+0.05(\mathcal{L}_{style_{out}}+2\mathcal{L}_{style_{comp}})+250\mathcal{L}_{tv}+20\mathcal{L}_{lap}$	45.245 \pm 35.381	33.272 \pm 4.469	0.966 \pm 0.018
$30\mathcal{L}_{valid}+240\mathcal{L}_{hole}+0.2\mathcal{L}_{per.}+0.05(\mathcal{L}_{style_{out}}+\mathcal{L}_{style_{comp}})+250\mathcal{L}_{tv}+20\mathcal{L}_{lap}$	44.290\pm33.785	33.271\pm4.267	0.966\pm0.018
$30\mathcal{L}_{valid}+240\mathcal{L}_{hole}+0.2\mathcal{L}_{per.}+0.05(\mathcal{L}_{style_{out}}+\mathcal{L}_{style_{comp}})+100\mathcal{L}_{tv}+20\mathcal{L}_{lap}$	46.238 \pm 34.697	33.013 \pm 4.154	0.963 \pm 0.020

Supplementary Materials

Therefore, the set of weight coefficients led to inpainting the images with the highest quality was determined, and the final objective function is defined as:

$$\mathcal{L}_{total} = 30\mathcal{L}_{valid} + 240\mathcal{L}_{hole} + 0.2\mathcal{L}_{perceptual} + 0.05(\mathcal{L}_{style_{out}} + \mathcal{L}_{style_{comp}}) + 250\mathcal{L}_{tv} + 20\mathcal{L}_{lap}$$

2.2) Autoinpainting

The protocols of the proposed autoinpainting pipeline require the choice of two hyperparameters: 1) the radius of the moving windows (circles) and 2) the number of top candidate regions. Accordingly, a set of independent experiments were examined to determine the optimal values of these two parameters.

To specify the radius of the moving circles and the number of top candidates, a range of different values was studied, and the effect of these values on the segmentation accuracy was quantified (Table 2.9.).

Table 2.9. The impact of changing the radius of the moving circles and the number of top candidate regions on the segmentation accuracy. The final candidate values are marked in bold.

Top Candidates - Circle Radius	Segmentation Metrics ($\mu \pm \sigma$)		
	Dice	Sensitivity	Specificity
1-23	0.305±0.231	0.291±0.153	0.982±0.001
1-25	0.308±0.189	0.289±0.114	0.984±0.001
1-27	0.312±0.233	0.301±0.149	0.984±0.001
1-29	0.321±0.194	0.309±0.213	0.984±0.001
1-31	0.328±0.183	0.305±0.142	0.986±0.001
2-23	0.395±0.192	0.381±0.183	0.999±0.000
2-25	0.393±0.159	0.378±0.128	0.999±0.000
2-27	0.410±0.193	0.392±0.176	0.999±0.000
2-29	0.401±0.203	0.389±0.127	0.999±0.000
2-31	0.403±0.143	0.395±0.142	0.999±0.000
3-23	0.422±0.167	0.416±0.160	0.999±0.000
3-25	0.429±0.170	0.412±0.167	0.999±0.000
3-27	0.437±0.172	0.419±0.171	0.999±0.000
3-29	0.433±0.174	0.415±0.174	0.999±0.000
3-31	0.419±0.174	0.412±0.176	0.999±0.000

Therefore, the radius of the moving circles was set as 27 pixels and the number of top candidate regions was determined as 3.