SUPERPIXEL-BASED CONTEXT RESTORATION FOR PANCREAS SEGMENTATION

SANDER VAN DONKELAAR

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

2002765

COMMITTEE

dr. Sharon Ong

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

May 26, 2022

ACKNOWLEDGMENTS

I want to thank my supervisors, Sharon Ong and Lois Daamen for all the feedback and valuable insights during our meetings. Moreover I'd like to thank Ralf Zoetekouw from Datacation for all his feedback and for introducing me to UMC Utrecht. Besides, I'd like to thank my family, friends and girlfriend for feedback and support.

SUPERPIXEL-BASED CONTEXT RESTORATION FOR PANCREAS SEGMENTATION

SANDER VAN DONKELAAR

Abstract

Automatic segmentation of the pancreas can help battle pancreatic cancer and other pancreatic diseases. Quantitative measures which are extracted from the pancreas based on CT imaging provide valuable biomarkers for tracking the progression of various endocrine and exocrine diseases (Panda et al., 2021). In recent years, deep learning has proven to be a powerful tool for pancreas segmentation. However, deep learning models suffer from data scarcity: the lack of annotated data poses a significant drawback in developing new models. The current work investigates to what extent self-supervised learning (SSL) can be used to leverage unlabeled data to increase performance in pancreas segmentation. This is done by in-painting masks generated by superpixels. It is hypothesized that this task yields a more heterogeneous representation since the network is forced to learn contextual information. It is empirically shown that the current approach outperforms the baseline trained without any self-supervision. Moreover, the current approach outperforms other state-of-art SSL approaches.

1 ETHIC STATEMENT

The work on this thesis did not involve collecting data from human participants or animals. The data used in the current dissertation belongs to the cancer imaging archive (TCIA), and the author acknowledges it does not have any legal claim to this data. In this research, code is used, which is adapted from Zhou et al. (2017).

2 INTRODUCTION

It is estimated that pancreatic cancer will be one of the deadliest forms of cancer by the year 2030 (Rahib et al., 2014). Early detection is vital

for the survival of the patient. However, since physical complaints only appear at a later stage, medical experts are often too late. As a result, the five-year survival rate has remained stable at around 5 % for the last 30 years (Åkerberg, Ansari, Andersson, & Tingstedt, 2017). Automatic segmentation of the pancreas can help battle pancreatic cancer and other pancreatic diseases. Pancreas segmentation is an crucial step in medical image processing. It consists of extracting a region of interest from a CT image or MRI scan, which identifies pancreatic tissue. Quantitative measures, which are extracted from the pancreas based on CT or MRI imaging, provide valuable biomarkers not only for tracking the progression of cancer but also of various other endocrine and exocrine diseases (Panda et al., 2021). Likewise, pancreas segmentation can be crucial for tasks such as observing lesions, analyzing anatomical structures, and the tracking of a disease and predicting the prognosis of a patient. (Lim et al., 2022). It has been found that all this can greatly support diagnosis and therapy (Yao, Song, & Liu, 2019). Therefore, pancreas segmentation is an important task in medical image analysis.

Segmenting the pancreas is a long and tedious task, and there are not always enough resources available to do this. Next tot this, manual segmentation of the pancreas is very error-prone. For example, a single abdominal CT scan can easily consist of several hundreds of slices, and the radiologist must keep concentration during the entire process. Therefore, it is worthwhile to investigate methods that can automate this process. This thesis will further explore this topic and is conducted in collaboration with the cancer center of UMC Utrecht. It is part of a larger coalition aimed at detecting pancreatic cancer recurrence after resection. Several researchers at this division are investigating whether Artificial Intelligence can be used in the early detection of pancreatic cancer recurrence. Their goal is to develop a pipeline in which Artificial Intelligence is used to aid radiologists in medical image diagnosis. This thesis is exploratory and explores the possibilities of a self-supervised learning method within the domain of pancreas segmentation. Moreover, it is estimated that these techniques will eventually also improve performance in detecting pancreatic cancer recurrence.

2.1 *Recent Developments in Medical Image Analysis*

In last years, significant process has been made in medical image analysis, which is empowered by the usage of deep learning models (Ker, Wang, Rao, & Lim, 2018). Convolutional Neural Networks (CNNs), a branch of deep learning models, are are at the core of state-of-art models in multiple subfields of medical image processing, such as image segmentation, classi-

fication, and more. For example, variations of convolutional networks have been successfully used to classify brain tumors (Irmak, 2021), segment lung tumors (Kasinathan & Jayakumar, 2022) and segment pancreatic tissue- and tumors (Lim et al., 2022; Mahmoudi et al., 2022).

2.2 *Current Drawbacks*

A downside of CNNs is that they usually contain many parameters. As a result, these networks require large amounts of data to learn (Shurrab & Duwairi, 2021). However, especially in medical image processing, annotated data is rare. Therefore, one of the significant drawbacks in developing deep learning models in the medical domain is data scarcity. Moreover, data is not easily shared due to privacy concerns (Bansal, Sharma, & Kathuria, 2022). One approach to tackle this is transfer learning, in which knowledge that is learned in a particular task is used in another target task. The most common approach is to use pre-trained state-of-art models such as ResNet or VGGnet. These pre-trained models are trained on massive datasets, which usually consist of millions of natural images and more than a thousand corresponding labels (Shurrab & Duwairi, 2021). Although pre-trained models show significant performance improvements in the natural image domain, it has been found that leveraging these pre-trained models in the medical domain yields limited performance increases. Mostly because natural- and medical images significantly differ in the distribution of features, such as intensity, contrast, and more (Azizi et al., 2021). Thus, the natural- and medical domains do not share enough characteristics to yield feature representations that are valuable in both domains (ibid.). An alternative option is to use transfer learning from tasks within the same domain. Nonetheless, this is often infeasible since it also requires labeled data (ibid.). However, other promising approaches have been developed in deep learning to tackle this.

2.3 *Self-Supervised Learning*

Self-Supervised Learning (SSL) is a hybrid learning approach that combines supervised and unsupervised learning. It consists of an unsupervised pre-training stage and a supervised fine-tuning stage. The unsupervised pre-training stage is aimed at introducing a kind of common sense into a network: it leverages supervisory signals from the data itself, which allows it to learn a representation that captures the underlying structure (Shurrab & Duwairi, 2021). This representation is functional at a later stage, as the model has learned a set of features that are useful in the subsequent task

(Zhai, Oliver, Kolesnikov, & Beyer, 2019). The knowledge is transferred by initializing a part of the network for the subsequent task with the weights that are learned in the unsupervised task. This way, unstructured medical data, such as unannotated CT scans, can be utilized. It it is hypothesized that self-supervised learning can strongly accelerate developments in deep learning, especially in industries where data is scarce (T. Chen, Kornblith, Swersky, Norouzi, & Hinton, 2020).

The current research will focus on this further, and will investigate the potential of self-supervision techniques to improve pancreas segmentation. More specifically, this research builds further upon that of (Chen et al., 2019), who proposed a *context-restoration* strategy for self-supervised learning. Their self-supervised task consisted of reconstructing a distorted image. It has been found that reconstructing these images as pre-training stage has produced significantly better performance on various subsequent tasks such as fetal standard scan plane classification, abdominal multi-organ localization, and brain tumor segmentation (ibid.).

2.4 *Superpixel-based Self-Supervised Learning*

In the current paper, it is investigated to what extent this also holds for pancreas segmentation. Besides, it is examined to what extent this method can be improved by using *superpixels*. To elaborate, most self-supervised learning that rely on reconstruction of distorted images use uninformed and random regions to corrupt an image (Kayal, Chen, & de Bruijne, 2020). For example, in the context-restoration method as described in Chen et al. (2019), images are distorted by swapping sub-patches of an image. However, the boundaries of the patches do not adhere to the boundaries of the organs in the image. Consequently, the network can use information from the organ itself to reconstruct the distorted areas. Therefore, it is hypothesized that the network does not have to rely on global contextual information, such as the presence and relative position of other anatomical structures, to rebuild the image. As a result, the network is not forced to learn a representation that encapsulates global spatial relationships. However, learning this information can be especially relevant for the pancreas, since the position, shape and size of the pancreas are strongly affected by its surrounding organs, such as the liver, stomach and kidneys (Oda et al., 2016). Likewise, it has been found that learning contextual information in CT scans improves performance in deep learning networks (Petit, Thome, Rambour, & Soler, 2021; Tang et al., 2020).

Therefore, the current thesis aims to improve this, by distorting images based on superpixels. Superpixels are a subgroup of pixels in an image that

share common characteristics, such as their location and pixel intensities. This way, superpixels segment an image into subsegments by considering similarity measures. The central principle is that the areas of the segmented superpixels adhere well to organ boundaries within a CT scan, which is utilized to segment parts of the image automatically. After the superpixels have been made, several superpixel segments are randomly selected and distorted. Afterward, a neural network is trained to reconstruct the image. Since superpixel segments correlate with the object boundaries in an image, areas that contain large parts of an organ, or even the entire organ are distorted. Therefore, the network is forced to use the presence and position of other organs to recreate the image. It is hypothesized that this will increase performance in the subsequent task.

2.5 *Research Questions*

In the following thesis, it will be investigated to what extent pancreas segmentation can be improved by using superpixel-based SSL. The goal of this method, is to learn a representation during pre-training that encapsulates contextual information in the abdominal region. It is hypothesized that learning this contextual information will yield a more heterogeneous representation that encapsulates global contextual information, which will benefit performance during pancreas segmentation. Given all this, the current research revolves around the following research question:

> *To what extent can superpixel-based context restoration improve pancreas segmentation?*

In order to tackle this question. Multiple sub-questions need to be addressed. First, a baseline is established to which the effects of superpixel-based context restoration is compared. This baseline will consist of U-Net, a common segmentation architecture, which is either randomly initialized, or initialized using patch-based pre-training as described by Chen et al. (2019). As a result, the first subquestion will read as follows:

> *RQ1: To what extent does patch-based context restoration increase performance in pancreas segmentation?*

Afterward, the results both networks will be compared to a third U-Net, which is pre-trained using superpixel-based context restoration. Thus, the second subquestion will read as follows:

RQ2: Does superpixel-based context restoration improve segmentation performance, compared to patch-based context restoration?

Moreover, (Chen et al., 2019) used L2 loss as loss function during the pre-training task. However the impact of using a different loss function during pre-training is not clear. Although L2 loss has been used widely in image reconstruction tasks, it also suffers from significant drawbacks. For example, L2 assumes pixel independence, and therefore fails to take spatial relationships between pixels into account. Other loss functions, such as SSIM loss do not make this assumption (Zhao, Gallo, Frosio, & Kautz, 2017). Therefore, it is further evaluated how SSIM loss improves performance compared to L2 loss. This yields the third subqestion:

RQ3: To what extent does the usage of SSIM loss during pre-training improve performance of pancreas segmentation, compared to L2 loss?

2.6 *Contributions*

2.6.1 *Academic Contributions*

The current work will investigate how superpixel-based SSL can be utilized to improve pancreas segmentation, which has not been investigated yet. The main goal is to develop a method that can leverage unstructured data. This can tackle the problem of data scarcity, which is especially relevant for the medical domain. It is hypothesized that leveraging SSL methods in the medical domain will result in more robust models and better generalization performance. Likewise, this can accelerate the performance of deep learning models for pancreas segmentation.

2.6.2 *Societal Contributions*

The current research investigates SSL in the context of pancreas segmentation, which can be utilized as tool to battle diseases such as pancreatic cancer. However, SSL techniques are applicable to a much wider range of tasks in the medical domain, such as brain tumour classification (Kayal et al., 2020), cardiac arrhythmia classification (Kiyasseh, Zhu, & Clifton, 2021) and abdominal organ segmentation (Ouyang et al., 2020). As a result, it can help accelerate the development- and performance of deep learning models in various medical tasks. This will can help the adoption of AI in the medical domain, which can increase the efficiency and effectiveness of our healthcare systems.

2.7 *Thesis Outline*

The research is structured in the following manner. Section 3 provides an overview of relevant literature. In section 4, the data and methods are further elaborated. The performance of the methods is evaluated in section 4. In section 5, the results are discussed, In section 6, conclusions are drawn from the research.

3 RELATED WORK

3.1 *Current Advancements in pancreas segmentation*

Several methods for pancreas segmentation have been developed over the last years. The segmentation of the pancreas is a difficult task due to large anatomical differences in terms of shape, size, and location (Yao et al., 2019). However, various methods have been proposed for this task. Traditional approaches involve multi-atlas techniques, which extract statistical information regarding size, orientation, or shape from training data. However, these techniques usually fail to cover all the anatomical variability and are highly dependent on the selection of training images (Huang, Huang, Yuan, & Kong, 2021). Therefore, these techniques have shown limited performance and generalization capability. Deep learning-based approaches greatly increased performance in pancreas segmentation (Yao et al., 2019), in which CNNs are at the core of these developments. Examples of well known architectures used for pancreas segmentation include fully convolutional neural networks (FCNs) (Xue et al., 2021), U-net (M. Li, Lian, Wang, & Guo, 2021; Petit et al., 2021; Zhou et al., 2017) and V-Net (Giddwani, Tekchandani, & Verma, 2020). Moreover, CNN architectures can either be in 2D, 3D or hybrid structure. In the following sections, multiple pancreas segmentation networks will be further elaborated.

3.1.1 *2D Pancreas Segmentation*

2D CNNs operate on the single slice level. Therefore, they are computationally very efficient, making them suitable for large datasets and easy deployment in production. Therefore, they are widely used in pancreas segmentation. For example, Petit et al. (2021) augmented the U-net architecture with self-and cross attention modules in order to learn long-range contextual interactions and spatial dependencies. The model yielded good results, with a Dice score of 78.50 %. Other approaches include multi-stage U-Net models, in which a bottom-up approach is used to segment the

pancreas Zhou et al. (2017). First, a coarse model is used to roughly segment the pancreas region, which is used as input to a subsequent model to fine-tune the segmentation. This process is repeated multiple times to improve performance. The model proved to be successful in 2D pancreas segmentation, with an average Dice coefficient of 82.37 %. One of the drawbacks of this method was that the loss function was minimized for each stage individually. As a result, it was found that some dependencies between stages could not be modeled. In order to tackle this, Yu et al. (2017) enhanced this work by introducing a saliency transformation module, in which the segmentation map of the previous iteration is applied to the current iteration. This allows for joint optimization, improving segmentation performance. Moreover, in more recent work by M. Li et al. (2021), a multi-level pyramidal pooling mechanism is used, which combines multiple pooling layers to gather contextual information for segmentation. However, one of the significant drawbacks of 2D models is that they cannot model inter-slice dependencies. As a result, some spatial information gets lost. This decreases the performance compared to other methods, such as 3D convolutional neural networks.

3.1.2 *2.5D and 3D models for pancreas Segmentation*

2.5D and 3D models generally outperform 2D models for pancreas segmentation. In short, 2.5D models operate by training a segmentation network for the axial, coronal, and sagittal planes and combining their predictions by using a majority vote, used to construct a 3D pancreas volume. This approach is powerful, especially when combined with other techniques, such as described by Zhou et al. (2017). For example, Yan and Zhang (2021) combined a 2.5D segmentation network with a coarse-to-fine grained training process. The network yields impressive results, with a Dice score of 86.61 %. Although 2.5D models are less resource-intensive than 3D models, they still require a longer time to train compared to 2D (Minnema et al., 2021). When looking at 3D networks , Salanitri et al. (2021) developed a fully connected CNN for pancreas segmentation. It consists of a 3D encoder that learns to extract volumes at different scales, which are sent to multiple 3D decoders, which are combined to obtain a unique single mask. The model yields impressive performance. However, this also comes at a cost. For 3D models, both training and inference models take significantly longer. Such models are often infeasible to use in a production setting, since state-of-art computational devices are needed, which are not always available. Moreover, all three architectures are limited by the fact that they are trained in a fully-supervised manner.

3.1.3 *Limitations of Supervised Deep Learning models*

Since all three approaches are based on supervised learning, they are all limited by the amount of annotated data that is available. Since annotated data is scarce, it is harder for a network to learn a heterogeneous representation that encapsulates the variation in the data. Data augmentation is one way to create a more heterogeneous representation. In data augmentation, existing data is augmented by copies of the data, which are slightly modified. Data augmentation has proven to be powerful in increasing robustness and generalization performance. However, it still requires annotated data. Therefore, its performance gains are limited, and abundantly available unstructured data sets remain unused. Likewise, transfer learning can be effective but suffers from the same limitation, since it requires annotated medical data to be effective (Azizi et al., 2021). Learning a more heterogeneous representation is especially relevant for the pancreas, since its high anatomical variability. Therefore, SSL methods have been developed to tackle this.

3.2 *Self-Supervised Learning*

Self-Supervised learning refers to methods in which neural networks can be trained to learn from data using *self-supervision*. The most popular approach in self-supervision is a two-stage paradigm, in which a network is pre-trained on a proxy task and then fine-tuned on a downstream task. It is found that pre-training on a proxy task yields robust features that increase generalization performance and convergence and avoids overfitting. The usage of SSL approaches in the medical domain has received relatively little attention (Azizi et al., 2021). As a result, several developed frameworks have not been extensively tested in the medical domain. In general, self-supervised learning approaches can be divided into three categories: contrast-based, context-based, and generative self-supervised learning strategies.

3.2.1 *Contrast-based Self-Supervised Learning*

Contrast-based self-supervised learning aims to learn representations by comparative learning. The core idea of contrastive learning is that similar objects should have similar representations. Recent developments in contrastive learning show promising results. For example, Azizi et al. (2021) proposed *multi-instance-constrastive* learning, in which data augmentation methods such as cropping or Gaussian blur were used to create different views of the same image (see figure 1). Moreover, if multiple images of the same object are available (such as a CT scan and a follow-up scan),

the distinct images were used to create positive pairs of examples. Afterward, an encoder network was used to learn valuable representations. The network was optimized using contrastive loss, aiming to minimize the difference between positive examples and maximize the difference between negative examples. For each positive pair, negative examples were obtained by considering all other augmented examples within a minibatch as negative pairs, following the training protocol of (T. Chen, Kornblith, Norouzi, & Hinton, 2020). It has been found that this technique yields significant performance improvements, which outperforms other approaches such as supervised transfer learning from images of the natural image domain, e.g. from images such as real-world scenes (ibid.). Moreover, the self-supervised models generalize better and are more label-efficient. As a result, the downstream model achieves state-of-art performance in a dermatology condition classification task. However, these methods are severely affected by the selection of negative examples, which is not optimal and can result in varying performance depending on the task (Xu, 2021). Other approaches exist within self-supervision frameworks, such as context-based learning and generative learning, which do not require the construction of negative examples.

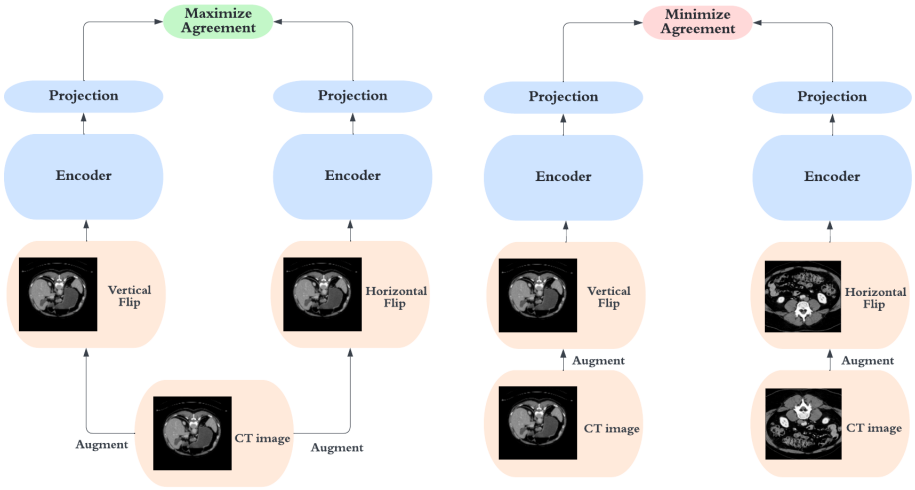

Figure 1: Schematic overview of a contrastive learning task. Image based on (Azizi et al., 2021).

3.2.2 *Context-based Self-Supervised Learning*

The primary goal of context-based SSL tasks is to learn contextual semantics. Examples include patch relative position prediction, angle prediction, or jigsaw puzzles. It has been found that it can increase performance in the subsequent task. For example, Noroozi and Favaro (2016) created a

jigsaw puzzle task in which a CNN was trained to classify nine sub-patches of an image in the correct sequence. The method proved successful in pre-training for the subsequent task but also had drawbacks. For example, since the number of possible combinations of a sequence of 9 items is exceptionally high (362880), the method was challenging in terms of model complexity and memory.

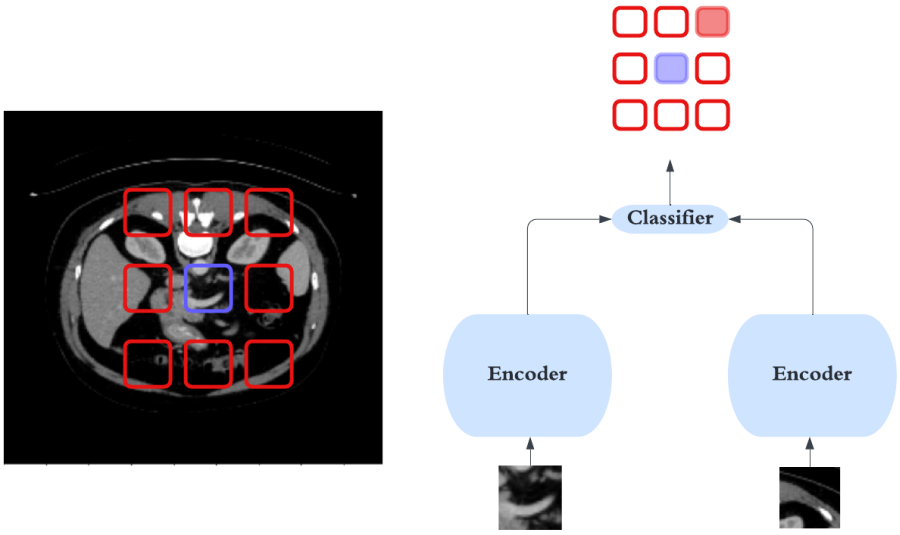

Figure 2: Schematic overview of a Context Learning task, based on the method of Doersch et al. (2015)

In order to tackle this, less computationally expensive tasks have been developed. One example is rotation prediction, in which the pre-training task consisted of predicting the angle in rotation. Although the model showed performance improvements on limited data and converged faster, the performance improvements on the whole dataset were limited (Imran et al., 2020). Other approaches to self-supervised learning include predicting the position in a 3x3 between a central patch and its surrounding patches (Doersch et al., 2015) (see figure 2). However, it has been found that the performance gains are limited since the network could complete the task using relatively trivial features (Chen et al., 2019). This emphasizes the complexity of designing a good pre-training task: it should have a good balance between simplicity and complexity. Moreover, the pre-training task must lie in the same domain as the fine-tuned task to learn semantically

relevant features. As a result, designing a pre-training task is difficult. Nonetheless, other approaches have been developed to tackle this.

3.2.3 *Generative Self-Supervised Learning*

Generative approaches are aimed at reconstructing distorted images volumes. For example, (Pathak, Krähenbühl, Donahue, Darrell, & Efros, 2016) proposed a method in which a CNN was learned to inpaint removed subpatches in an image. The authors proposed that by inpainting the image, the model had to learn the semantic context of an image to reconstruct it. However, this approach yield limited performance in medical imaging. One of the reasons for this is that removing an image patch alters the intensity distribution of an image. As a result, the resulting- and original images belong to a different domain, which yields limited performance (Chen et al., 2019). However, other generative approaches have yielded impressive results.

As noted earlier, Chen et al. (2019) constructed a reconstruction task in which the network had to reconstruct an image that was distorted by swapping sub-patches. More specifically, the images are distorted by iteratively swapping two random patches of size p x p. Repeating this process several times will yield a distorted image in which the spatial information is altered, but the intensity distribution is preserved. The context restoration task yielded impressive results and improved performance in various tasks, such as semantic segmentation, localization, and classification. However, one of the drawbacks is that the pre-training task is not optimized for the downstream task since the selection of regions to mask is uninformed and random. Thus, the boundaries of the masked regions do not adhere to the boundaries of the organs in the image. As a result, the network can use information from the organ itself to reconstruct the image. Therefore, the network is less forced to learn a representation that encapsulates the spatial relationships between organs. Other approaches have been developed to tackle this. For example, Kayal et al. (2020) constructed a region-of-interest guided super voxel inpainting task. In this task, supervoxels were used to mask regions in an image. Supervoxels best can be described as superpixels in 3D space, in which similar voxels are grouped based using similarity measures. Thus, the described approach is similar to the approach in the current paper. The selection of supervoxels to be masked, is guided by a region-of-interest (ROI). This entails that the task uses the annotated segmentation maps to select relevant areas to be masked. Thus, only regions that (partly) contain tumour tissue are masked. The results of this approach are promising. The ROI-super voxel task outperformed the baseline to a great extent in the downstream task. However, one of the significant drawbacks of this approach is that the method uses the

ROI to select relevant supervoxels. This counters one of the core ideas of self-supervised learning: learning from unlabelled data. Therefore, it is less relevant in the medical domain since annotated data is sparse. However, it also should be noted that even without using the ROI to select areas, the approach yielded significant performance improvements compared to the baseline. However, the potential of these methods for pancreas segmentation is unclear. Therefore, the current research will investigate this further and take the limitations of current research into account. In the following sections, the proposed method will be further elaborated (see figure 3 for an overview).

4 METHODS

The current study will focus on the effectiveness of superpixel-based context restoration as self-supervised learning (SSL) technique to improve U-Net performance in pancreas segmentation. In order to evaluate performance, the technique will be compared to a randomly initialized 2D U-Net (Baseline U-Net). Besides, the effectiveness will be compared to another SSL method, which is patch-based context-restoration (PB U-Net) (Chen et al., 2019). Each model will be compared using the Dice- and Jaccard coefficient. The data, network architecture, and pre-training methods are further explained in this chapter.

4.1 *Experimental Set-Up*

As noted earlier, the pre-training task is designed to yield a set of layer-weights that encapsulates useful information for the final task of pancreas segmentation. Therefore, instead of randomly initializing the weights of the segmentation network, the weights are initialized by using (a subset) of the weights which are obtained during the pre-training task. The primary difference between self-supervised learning methods lies in the design of the pre-training task. The goal is to create a useful task, which forces the network to learn domain knowledge that can be transferred.

The architectures of the networks in the pre-training tasks are exactly the same as in the fine-tuning tasks, and both consist of a 2D U-net (Ronneberger, Fischer, & Brox, 2015). However, since the pre-training tasks and fine-tuning tasks consist of seperate objectives, different loss functions and hyper-parameters are used. Likewise, the activation functions in the output layers are different. An overview of the experimental setup can be found in figure 3.

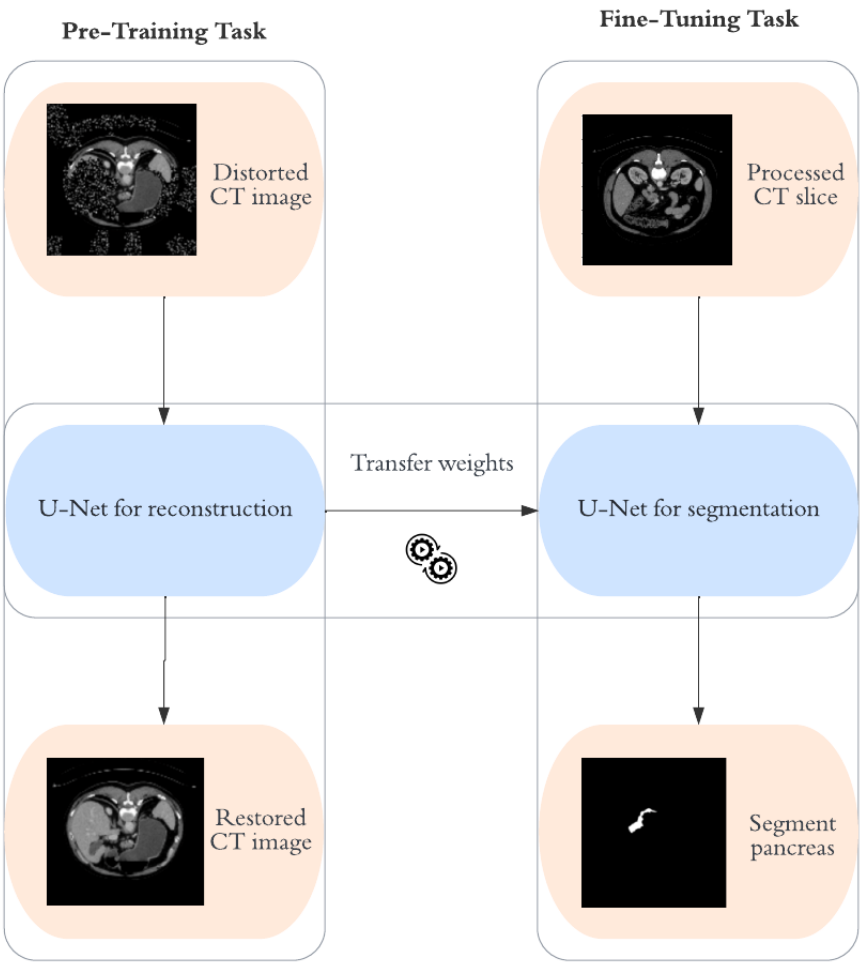

Figure 3: Overview of the experimental set-up.

4.2 *Dataset Description*

In the current study, the NIH pancreas-CT dataset is used (Roth et al., 2016) to train- and evaluate a network that can segment the pancreas. The dataset contains 82 abdominal, contrast-enhanced 3D CT scans. All scans are manually segmented by a medical student and verified by an experienced radiologist. The resolution of the CT scans is 512 x 512 x N, where N lies between [181, 466]. Moreover, the slice thickness T varies per scan where T lies between [1.5, 2.5]. Since it has been found that augmenting the data during the pre-training phase in SSL tasks leads to better performance in the subsequent task, another dataset is used during pre-training. This dataset consists of 50 abdominal CT scans from the AbdomenCT-1k dataset (Ma et al., 2021). The resolution of the CT scans is

512 x 512 x N, where N lies between [71, 113]. The slice thickness varies between [0.65, 5] cm. The AbdomenCT dataset will not be used during the subsequent task of training a model for pancreas segmentation. Therefore, it is only used in the pre-training SSL task.

4.3 *Data Preprocessing*

Pre-processing of the data consists of several steps. First, each image is clipped between [-100, 240] HU, following the protocol in (M. Li et al., 2021; Yan & Zhang, 2021). Afterward, each scan is normalized within [0,1] by using MinMax scaling. Finally, all images are cropped to the dimensions [300,300], to decrease the amount of abundant information. Afterward, the images are blurred using a Gaussian blur with a standard deviation of 0.5 to counter anti-aliasing effects. Finally, they are resized to [208, 224].

4.4 *U-Net Architecture*

The model will for both the pre-training tasks and pancreas segmentation will consist of a 2D U-Net (Ronneberger et al., 2015), which has been used extensively in medical image segmentation tasks. The U-Net architecture is based upon the *fully convolutional network* (Long, Shelhamer, & Darrell, 2014), which is adapted to work with few training data and to yield more precise segmentation. The U-Net follows an encoder-decoder-like structure, in which a contracting part consisting of various convolutional layers is followed by an expanding part that consists of various up-sampling layers. Hence the expanding layers increase the resolution of the output back to its original shape. The network differs from other encoder-decoder architectures such that the contracting and expanding part of the network is not fully decoupled. This is due to skip connections between the layers in the contrasting and expanding parts: feature maps of the convolutional layers in the contrasting part are concatenated with outputs of subsequent layers in the expanding part, which are used as input for each up-sampling layer. This allows the network to recover spatial information that is lost during down-sampling operations in the contrasting part of the network (Ronneberger et al., 2015). As a result, the number of convolutional- and up-sampling layers is equal, yielding an identical U shape. An overview of the U-Net architecture adapted for the current task of pancreas segmentation can be found in figure 4.

Since the CT scan is in grayscale, the input map is of size 208 x 224 x 1, which is followed by four encoder blocks. Each encoder block consists of two convolutional layers with a *ReLu* activation function and a kernel size of 3x3. Both layers are followed by a max-pooling operation with a kernel

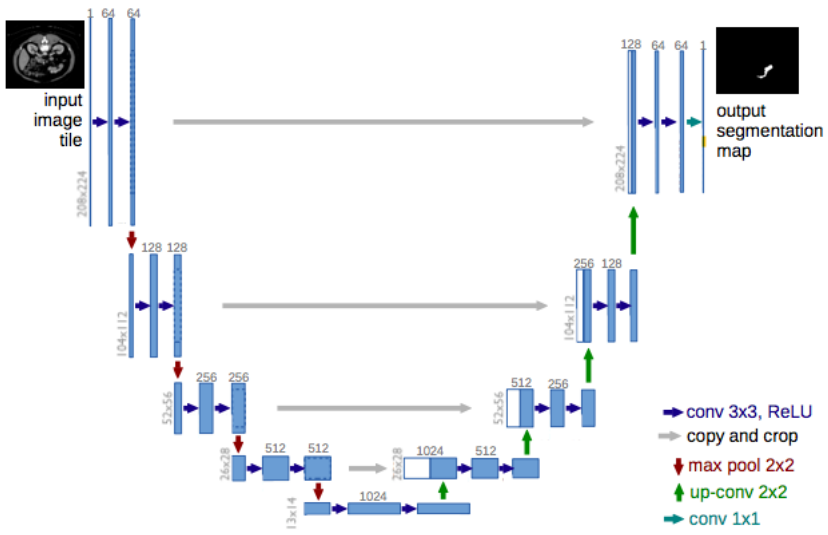

Figure 4: Overview of the used U-Net architecture, figure adapted from (Ronneberger et al., 2015)

size of 2x2. Batch Normalization is applied after each convolutional layer to make the network train faster and more stable (Ioffe & Szegedy, 2015). The number of convolutional layers in each block increases by a factor of two: the convolutional layers in the first block have 64 filters, the layers in the second block have 128 filters, the layers in the third block have 256 filters, and the layers in the fourth block have 512 filters. Afterward, the resulting feature maps are expanded by transposed convolutional layers. The expansive part of the network consists of 4 blocks that consist of one up-sampling layer, followed by two convolutional layers with *ReLu* activation function and a kernel size of 3x3. The output is passed to a concatenation layer, where the output of the subsequent layers and the corresponding output of the feature maps in the contracting path is concatenated. The amount of filters is divided by two in each block. During this process, the down-sampled representation from the contrasting part is up-sampled back to the size of the original input.

The loss during the segmentation task is minimized by using Dice Loss, which is calculated as follows:

$$\mathcal{L}_{Dice} = 1 - DSC \tag{1}$$

The Dice score coefficient (DSC) measures the overlap between the prediction and ground truth and is widely used to assess segmentation performance. The DSC is calculated by multiplying the area of intersection between A and B with two, divided by the sum of the areas of A and B.

This gives the following equation:

$$DSC = \frac{2TP}{2TP + FP + TN} \tag{2}$$

In this equation, True positive (TP) indicates the number of foreground pixels (e.g., the pancreas mask) correctly classified as pancreas by the model. The false positives (FP) are the background pixels incorrectly classified as foreground pixels. True negatives (TN) indicate the number of the background pixels correctly classified as background pixels by the model. Likewise, false negatives (FN) indicate the number of foreground pixels incorrectly classified as background pixels by the model. One of the main advantages of using Dice Loss over other loss functions in semantic segmentation is that it can handle imbalanced data (Sudre, Li, Vercauteren, Ourselin, & Cardoso, 2017). Therefore, this is especially relevant for pancreas segmentation since the pancreas only makes up a small part of each CT scan (Laoveeravat et al., 2021). Moreover, only slices that contain 50 or more pixels of the pancreas are used for training, while testing is done on all data, which helps to limit the impact of background pixels during training (Zhou et al., 2017).

4.5 *Weight Initialization*

The primary goal of pre-training is to use these weights when initializing the contractive part in the U-Net in the subsequent task. In the section below, the pre-training tasks are further explored.

4.5.1 *Standard U-Net*

If a network is not pre-trained, the weights are initialized by using the Xavier initialization (Glorot & Bengio, 2010). In this method, biases in each layer are initialized with the value 0. The weights W_{ij} are sampled from the following uniform distribution D:

$$W_{ij} \sim D[\frac{-1}{\sqrt{n}}, \frac{1}{\sqrt{n}}] \tag{3}$$

Where n is the number of outgoing connections in the previous layer. Thus, the lower- and upper bound of the distribution at layer L_i is dependent on size of the previous layer L_{i-1}.

4.5.2 *Superpixel-based pre-training*

In the superpixel-based context-restoration method, a CNN is learned to approximate the function g (x1), where x1 is the distorted image, and g(x1)

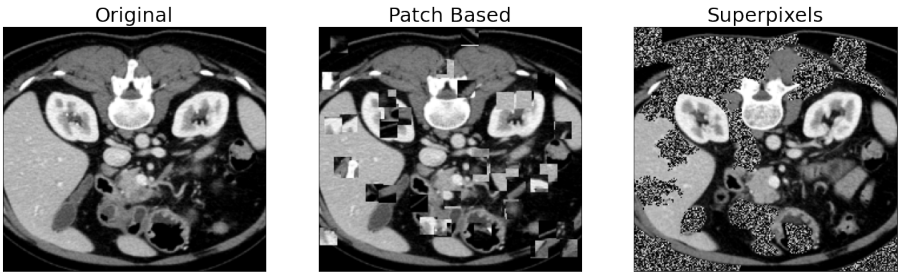

Figure 5: Example of a distorted image in both pre-training methods.

yields the original image (x). Superpixels are used to distort the image. Superpixels are a subgroup of pixels in an image that shares common characteristics, such as pixel intensities. In this way, superpixels segment an image into subsegments by considering similarity measures. The main idea is that the areas of the segmented superpixels adhere well to object boundaries within the CT scan, which can be utilized to segment parts of the image automatically. This characteristic is leveraged to create a self-supervised learning task. In short, this task is constructed as follows: first, each slice is segmented into N segments by using the SLIC algorithm (which will be discussed in more detail later on). After the image has been segmented, K segments are randomly chosen, and the intensity values are replaced with intensity values that are randomly sampled from the image. K is calculated by using the ratio parameter R. To elaborate, R can be seen as a ratio of N, the total amount of superpixel segments. For example, if N is 200 and R is 0.2, K yields 40. As a result, 40 segments will be selected to be distorted. Moreover, the reason that pixels values are sampled from the original image, is that the intensity distribution is preserved. This is important for the network to learn features belonging to a specific domain. (Chen et al., 2019).

Algorithm 1 Distort images using Superpixel method

Require: image x_o

1: Transform image x_o into N superpixel segments. Following from this is $S\epsilon[S_1, S_2...S_N]$, where S_i is a superpixel segment.
2: Randomly sample K superpixel segments into S', which yields $S'\epsilon[S_1, S_2...S_K]$.
3: Save the indices $[x_i, y_i]$ of all pixel values from the superpixels in S' into I.
4: Replace all values at indicies I with pixels randomly sampled from x_o, which gives distorted image x_d
5: Return distorted image x_d

In the current study, superpixels are generated by using the SLIC (Simple Linear Iterative Clustering) algorithm (Achanta et al., 2012). It generates superpixels by clustering pixels based on color similarity and closeness in the image plane. Since the CT images are in grayscale, clustering is performed in three-dimensional [xyi] space, [xy] is the pixel position and i is the intensity SLIC has been proven to be a fast and effective method to generate superpixels (ibid.). The algorithm works through several steps explained in more detail below.

1. The first step consists of initializing the cluster centers N. The number of cluster centers corresponds to the number of superpixels present in the image and is obtained by sampling pixels at regular grid steps S. Each pixel is represented by [I_n, X_n, Y_n]. After the cluster centers are created, they are moved to a seed location corresponding to the lowest gradient position in a 3 × 3 neighborhood to avoid placing them at an edge.

2. After the previous step, each pixel is assigned to the nearest cluster within the search area. After all, pixels are assigned, a new center is computed by taking the mean of all [ixy] vectors. This process is repeated until convergence. The algorithm converges when the residual error E is below a certain threshold.

3. After this process has been finished, connectivity is enforced by connecting disjointed pixels.

SLIC is initialized with 100 segments and a compactness of 0.05, which is the trade-off for color-similarity and proximity. An example can be found in figure 5.

4.5.3 *Patch-based pre-training*

In patch-based context restoration, images are distorted through a patch swapping method: given an image x_n, two sub-patches of dimensions (p, p) are randomly selected and swapped. This process is repeated T times, which leads to the distorted image x_d. The optimal values for the parameters p and T will be found through a random search across multiple combinations of values. An example of a distorted image can be found in figure 5.

4.6 *Loss Functions for pre-training*

In this research, the effect of two loss functions for the pre-training stage are compared: l2 Loss and SSIM loss, in order to investigate the choice of loss function on the final performance.

4.6.1 *L2 Loss*

The loss is minimized by using the L2 loss (least Squares Error) function. The L2 loss is a relatively simple loss function, as it is the sum of all the squared differences between the true and predicted values. It is calculated as:

$$\mathcal{L}2_{loss}(x, y) = \sum_{i=1}^{N} (x - y)^2 \tag{4}$$

Although L2 loss has shown powerful results, it is also known that L2 loss is not optimal for image restoration as it leads to blurred images and does not correspond well to image quality as perceived by a human observer (Zhang, Zhang, Mou, & Zhang, 2012). One of the main drawbacks of L2 loss is that it assumes that pixels are independent of each other, while in reality, this is not the case: the value for a pixel depends on the values of the pixels that surround it. However, other loss functions exist which do not make this assumption. For example, the structural similarity (SSIM) index provides a measure of similarity by comparing two images based on luminance, structural- and contrast similarity (Zhao et al., 2017), which resembles how a human would evaluate the similarity between two images.

4.6.2 *SSIM Loss*

The loss function consists of three core parts: luminance, contrast, and structure. Luminance reflects the averaged intensity values over all pixels in an image (μ_x). In order to calculate the similarity in luminance between two images (x, y) the following equation is used, where C_1 is a constant.

$$L(x, y) = \frac{2\mu_x\mu_y + C_1}{2\mu_x^2\mu_y^2 + C_1} \tag{5}$$

The second part reflects the similarity in variation in luminance, which is defined as contrast (σ_x). The similarity in contrast between the two images is calculated as follows

$$C(x, y) = \frac{2\sigma_x\sigma_y + C_2}{2\sigma_x^2\sigma_y^2 + C_2} \tag{6}$$

The third part, structure is defined as the pearson correlation of the luminance of two images. It is calculated as follows:

$$S(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x^2 \sigma_y^2 + C_3} \tag{7}$$

is defined by multiplying the three individual functions with each other, together with a corresponding weighting factor (α, β and γ)

$$SSIM(x,y) = \alpha L(x,y) \cdot \beta C(x,y) \cdot \gamma S(x,y) \tag{8}$$

Following from this, SSIM loss can be calculated as follows:

$$\mathcal{L}_{SSIM}(x,y) = 1 - SSIM \tag{9}$$

In the current work, SSIM is implemented by using the function from the Tensorflow library.

4.7 *Evaluation Metrics*

All models will be evaluated in terms of their Dice coefficient, which is explained in section 4.2. Besides, each model will also be evaluated using the Jaccard Index. For each class, it is defined as follows:

$$IoU = \frac{TP}{TP + FP + TN} \tag{10}$$

Afterward, all scores are weighted for each class. While the Dice coefficient and Jaccard index are very similar, the Dice coefficient puts more weight into the true positives (recall that for calculating the Dice coefficient, true positives are multiplied with 2 in both the numerator and denominator). On the contrary, the Jaccard index yields an even weighting for TP, FN, and TN. Therefore, it can be a more robust metric in cases where false positives or false negatives are increasingly unfavorable, such as in the healthcare domain.

4.8 *Implementation Details*

4.8.1 *Pre-training Stage*

In the pre-training stage, all experiments are conducted by training a network for 10 epochs with a learning rate of 0.0001 and a batch-size of 4.

4.8.2 *Fine-tuning Stage*

In the fine-tuning stage, all experiments are conducted by training the network for 10 epochs, which is common when training with lower batch sizes (Huang et al., 2021). Moreover, the networks are trained with a learning rate of 0.0001 and a batch-size of 4. In order to get a more robust estimate of performance, all experiments were carried out using four-fold cross-validation. Three folds of patients are used as training data set for each fold, and the remaining fold for testing. This process is repeated until all folds have been used for training- and testing.

4.8.3 *Software libraries*

All code is written in Python 3.8. The used libraries are Numpy, OpenCV2 and skimage for data processing. Besides, the Tensorflow framework is used to construct all machine learning models. Moreover, a Google Colab Pro+ instance is used to train all models, which consists of 54 GB of RAM and a Nvidia P100 GPU.

5 RESULTS

This thesis will present the results of all experiments in three parts. In the first two sections (5.1, 5.2), the results of each SSL method are explained. This consists of a qualitative- and quantitative comparison of different pre-training methods- and loss functions. In the last section (5.3), the performance of all approaches is compared. All metrics are calculated for the full CT scan (e.g. the 3D image). Moreover, the standard deviations and minimum and maximum values of each fold are reported. Note that the standard deviation is calculated over all individual predictions. Finally, in order to increase evaluation robustness, all models are compared to the baseline by using a paired t-test.

5.1 *Patch-based Context Restoration*

In the following section, the results of patch-based context restoration are presented.

5.1.1 *Pre-training task*

A qualitative overview of the results of context restoration can be found in figure 6. Both models, pre-trained with L2 and SSIM loss, yield good results. The structure of images is similar to the original image. However, some of the restored parts are blurry for both models.

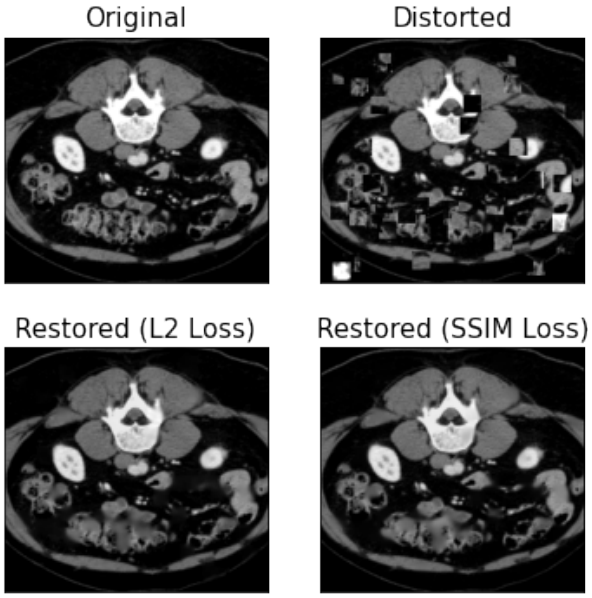

Figure 6: Example of the restored images using patch-based pre-training.

5.1.2 *Fine-tuning task*

In table 1, the results of both loss functions for patch-based context restoration are presented. Compared to the baseline, it has been found that pre-training the model using patch-based context restoration slightly improves performance. However, results from the paired t-test indicate that these improvements are not significant. As can be seen in table 1, SSIM loss yields the best performance, both in terms of Dice and Jaccard coefficients, with average scores of 75.03 and 61.06, respectively. Moreover, the standard deviations for the Dice and Jaccard scores for the model trained with SSIM Loss are 10.55 and 12.16, which are lower compared to both the baseline and the U-Net model pre-trained with L2 loss. However, the UNet pre-trained with L2 loss model has a higher minimum score for the Dice coefficient, and a higher maximum score for the Jaccard coefficient.

Loss	**Dice**	**Std.**	**Min.**	**Max.**	**Jaccard**	**Std.**	**Min.**	**Max.**
SSIM	**75.03**	**10.55**	70.99.	**78.79**	**61.06**	**12.16**	**59.95**	62.55
L2	74.49	11.51	**71.89**.	77.55	60.50	12.68	58.22	**64.09**
Baseline	*74.44*	*11.89*	*71.06*	*77.40*	*60.59*	*13.25*	*57.4*	*63.59*

Table 1: Evaluatin metrics for patch-based method as decribed by Chen et al. (2019).

In figure 7, one can see a qualitative overview in which the predictions of both loss functions for patch-based pre-training are compared for patient

70. When looking at the images, one can see that both models perform similar. However, it seems that the predictions of the pre-trained models are slightly more in line with the boundaries of the true masks. For example, when looking at figure 7, the baseline U-Net shows a less fine-grained mask of the pancreas compared to other the other models. However, it should be stressed that the performance differences in figure 7 are minor and only for a single patient.

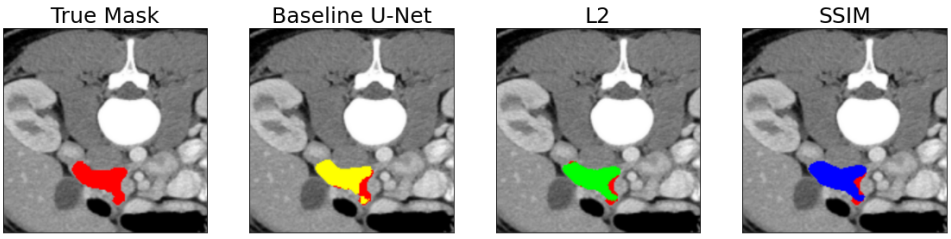

Figure 7: Example of the predicted masks of each model using patch-based pre-training. See appendix A (figure 11) for a more detailed overview.

5.2 *Superpixel-Based Context Restoration*

In the following section, the results of superpixel-based context restoration are presented.

5.2.1 *Pre-training task*

In figure 8, one can see a qualitative overview of the results of superpixel-based context restoration. As one can see, both models, pre-trained with L2 and SSIM loss, yield good results. The structure of images is similar to the original image. However, similar to the patch-based method, some of the restored parts are blurry for both models.

5.2.2 *Fine-tuning task*

In table 2, the results of superpixel-based pre-training are presented. Compared to the baseline, it has been found that pre-training the model using superpixel-based context restoration improves results. However, only for the model pre-trained with L2 loss are the results significant. Contrary to the results of patch-based context restoration, L2 loss yields the best performance, with an average Dice of 76.00 and Jaccard of 62.27 . Moreover, the standard deviation in general is lower compared to the baseline. It should be noted that although the Dice scores are slighly lower, the standard deviation of the model trained with SSIM Loss is lower compared

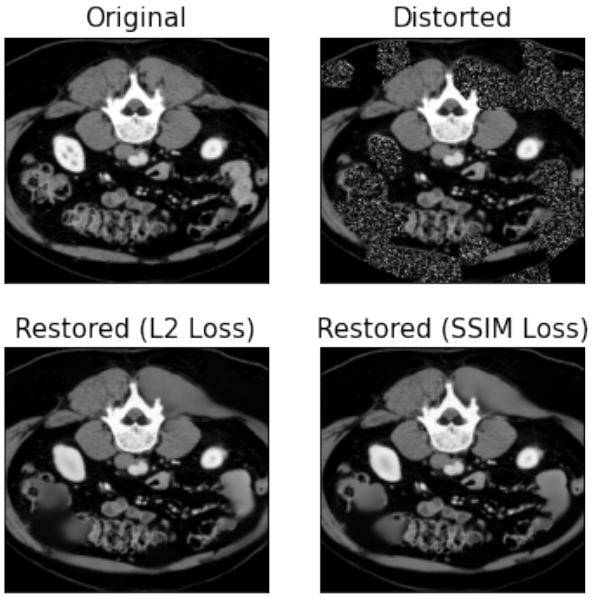

Figure 8: Example of the restored images using superpixel-based pre-training.

to the baseline and L2 loss. However, the UNet pre-trained with L2 loss model has a higher minimum score for the Dice coefficient, and a higher maximum score for the Jaccard coefficient.

Loss	Dice	Std.	Min.	Max.	Jaccard	Std.	Min.	Max.
SSIM	75.40	**10.23**	70.99.	**78.79**	61.47	**11.75**	56.74	65.50
L2*	**76.00**	10.36	**74.26**.	78.34	**62.27**	11.89	**60.22**	**64.89**
Baseline	*74.44*	*11.89*	*71.06*	*77.40*	*60.59*	*13.25*	*57.4*	*63.59*

Table 2: Results of the U-Net models pre-trained with superpixel-based context restoration. The * indicates that the results are significant compared to the baseline, for alpha 0.05.

In figure 9, one can see a qualitative overview in which the predictions of both loss functions for superpixel-based pre-training for patient 70 are compared. Similair to the results from patch-based context restoration, the results from the pre-trained models based on superpixels seem more fine-grained compared to the baseline model.

5.3 *Comparison of pre-training methods*

In table 3, all models are compared. Given these results, it is clear that pre-training with superpixels yields the best performance. Moreover,

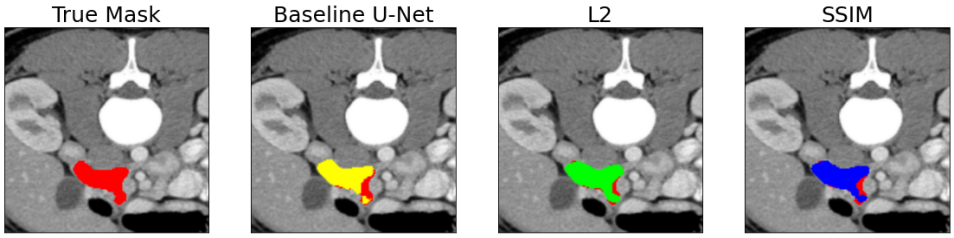

Figure 9: Example of the predicted masks of each model using superpixel-based pretraining. See appendix B (figure 12) for a more detailed overview.

pretraining with superpixels seems to result in more robust models, since the standard deviation is lower for both the Dice- and Jaccard scores.

Model	Loss	Dice	Std.	Min.	Max.	Jaccard	Std.	Min.	Max.
PB U-Net	L2	74.49	11.51	**71.89**.	77.55	60.50	12.68	58.22	64.09
PB U-Net	SSIM	75.03	10.55	70.99.	**78.79**	61.06	12.16	59.95	62.55
SP U-Net*	L2	**76.00**	10.36	**74.26**.	78.34	**62.27**	11.89	**60.22**	64.89
SP U-Net	SSIM	75.40	**10.23**	70.99.	**78.79**	61.47	**11.75**	56.74	**65.50**
Standard U-Net	-	74.44	11.89	71.06	77.40	60.59	13.25	57.4	63.59

Table 3: Comparison of all three pre-training methods. Abbreviations are used for model names such that the table can fit on one page.

This can also be seen when performing a qualtiative assesement of the results: superpixel-based pre-training significantly outperforms other methods when the data is irregular, such as being slightly rotated. For example, clear differences can be seen in terms of performance for patient 80 (figure 10). It is clear that the the irregular- and disconnected shapes of the pancreas are detected much better in comparison to other models.

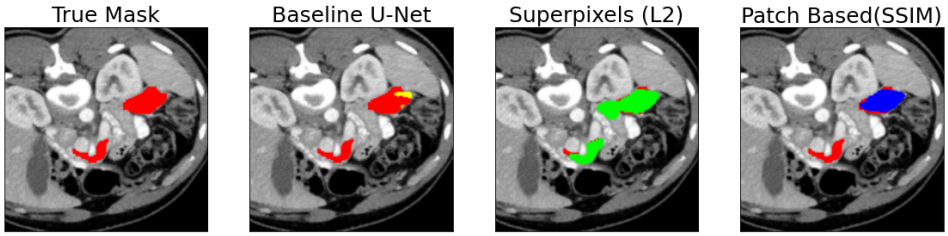

Figure 10: Example of the predicted masks of each model. See appendix C (figure 13) for a more detailed overview.

6 DISCUSSION

This research has investigated the extent to which superpixel-based pre-training self-supervised learning improves the task of pancreas segmenta-

tion. Thus, the current study revolved around the following research question: ***"To what extent can superpixel-based context restoration improve pancreas segmentation"***. Besides, it is investigated whether superpixel-based pre-training improves results compared to existing methods, and how the loss function during pre-training influences performance in the target task.

Each of the three sub-questions will be answered in the following sections to formulate an adequate answer to the overarching research question. As a first step to answer the research question, it is investigated to what extent patch-based context restoration (Chen et al., 2019) improves results (section 6.1). Besides, it has been investigated whether superpixel-based context restoration improves performance in pancreas segmentation compared to patch-based context restoration (section 6.2). Moreover, the effect of different loss functions are compared in section 6.3. In section 4, the main findings are discussed in light of the overarching research question. Finally, in section 6.5, the limitations of the current research are debated.

6.1 *Patch-based context restoration*

When looking at table 2, one can see that patch-based context restoration does not yield significant performance gains compared to the baseline. The performance gains for patch-based context restoration with L2 loss are negligible. Therefore, the differences can also be attributed to randomness during the initialization of parameters (recall that not all weights are transferred after retraining), which can also account for differences performance (Man, Huang, Feng, Li, & Wu, 2019). Therefore, when coming back to the original subquestion ***"to what extent does patch-based context restoration increase performance in pancreas segmentation?*** one cannot conclude that patch-based context restoration improves pancreas segmentation, which counters the findings presented by Chen et al. (2019). One possibility is that patch-based pre-training is not suitable for pancreas segmentation due to the nature of the pre-training task. As a result, the network does not learn semantic features relevant to the target task. This pattern can also be found in other literature. For example, Azizi et al. (2021) found that the in-painting of random patches in an image did not significantly improve brain tumor segmentation when training all data.

6.2 *Superpixel-based context restoration*

When looking at table 2, one can see that superpixel-based pre-training with L2 loss significantly outperforms the baseline. Moreover, the standard

deviations for both the Jaccard- and Dice coefficients are lower, which indicates that both models are more robust. Therefore, when coming back to the second subquestion ***"does superpixel-based context restoration improve segmentation performance, compared to patch-based context restoration"***, one can conclude that superpixel based context restoration yields better performance compared to patch-based context restoration. However, it should also be noted that only the results of superpixel-based pre-training with L2 loss are significant. Interestingly, these findings support the hypothesis that superpixel-based context restoration yields a more heterogeneous representation, which benefits pancreas segmentation.

6.3 *Effect of different loss functions*

In the current study, it has been investigated to what extent the choice of loss function affects performance in the subsequent task, which leads to the third subquestion ***To what extent does the usage of SSIM loss during pre-training improve performance of pancreas segmentation, compared to L2 loss***. When looking at superpixel-based context restoration, only the pre-training using L2 loss resulted in significant results compared to the baseline. Using SSIM loss did not result in significant performance improvements for both pre-training methods. Therefore, the current evidence indicates that L2 loss is favorable to SSIM loss. One possibility for the lower performance is the usage of a Gaussian filter during pre-process to blur the images, while it has been found that blurring negatively affects SSIM performance (C. Li & Bovik, 2010). Moreover, it can be the case that SSIM loss is not optimal for medical imaging in general, which has been supported by other studies as well (Kim et al., 2011)

6.4 *Main Findings*

The current research is the first to investigate the effect of superpixel-based context restoration in the context of pancreas segmentation. Coming back to the main research question, ***to what extent can superpixel-based context restoration improve pancreas segmentation***, the results indicate that pre-training a model using superpixel-based context restoration with L2 yields the best results. It is found that pre-training the model results in performance gains of up to 1.5 %. Besides, the standard deviation is also lower, which indicates that the model is more robust. Thus, the results suggest that superpixel-based pre-training tasks are promising for pancreas segmentation and self-supervised learning in general, which extends the findings of other literature (Kayal et al., 2020).

6.5 *Limitations*

The methodological limitations are further elaborated in the following subsections.

6.5.1 *Choice of hyperparameters*

Due to limited time, the selection of hyper-parameters could have received more attention, which can limit the robustness of the findings. For example, while training the model in both the pre-training- and fine-tuning stage, the effect of different batch sizes or the number of epochs has not been investigated. Likewise, it can be that some parts of the restored images in figure 6 and 8 are blurry simply because all models have not fully converged. Likewise, it has not been well investigated how different hyper-parameters generating distorted data can affect model performance. For the patch-based pre-training, this includes the size of the patches or the number of patches that are swapped. For superpixel-based pre-training, this consists of the number of segments when initializing the superpixel centers or the number of segments that are distorted.

6.5.2 *High Variance*

Another limitation is that the standard deviation of all models is high, which indicates a high variation in performance depending on the subject. Especially for critical applications, such as healthcare, results should be robust and consistent. Likewise, since the dataset is relatively small (82 patients), it should be further investigated how the findings generalize to a broader public.

6.5.3 *Effects of model complexity or data augmentation*

Moreover, the current study focuses on self-supervised learning and uses a relatively simple network architecture (standard U-Net) to investigate this. It is hypothesized that pre-training the model yields a richer, more heterogeneous representation, which is helpful in the subsequent task. However, can these heterogenous features *only* be learned by using pre-training? Or can these features also be learned when using more complex network architectures- and training setups, as described in the current state-of-art in the literature (M. Li et al., 2021; Petit et al., 2021; Zhou et al., 2017). This remains unclear from the current research. Moreover, the possible benefits of data augmentation have not been taken into account in the current research. Therefore, it more research is needed to investigate whether performance gains of SSL and data augmentation would yield similar representations, which would limit the total performance gains, or

whether the learned representations are complementary and both increase performance when used together.

6.5.4 *Lack of validation data*

In current validation set has not been used in the c. The main reason for this is that data is limited, and the validation set will decrease the (already small) training data size. Moreover, since the dataset is small, a validation set of 10 % will only result in eight CT scans for validation, which is simply too small for a robust estimate. Therefore, it has not been well investigated to what extent the models have converged, since callbacks such as *early stopping* are no implemented.

6.6 *Suggestions for further research*

The current work yields several opportunities for further research. For example, one of which is to compare the results of the current self-supervised learning approaches to other SSL paradigms, such as methods from contrast- and context-based SSL. Besides, the effect of different hyperparameters should be further investigated. Another interesting topic which builds further upon this, would be to investigate the effects of increasing the size of the dataset during pre-training. To elaborate, currently only 50 extra scans are used during pre-training. However, other studies use substantially more data during pre-training (Azizi et al., 2021). It is definitely possible that this yields a more heterogeneous representation which is useful for the subsequent task. Finally, it is worthwhile to investigate how the superpixel-based context restoration can be used together with coarse-to-fine methods, as described in M. Li et al. (2021); Zhou et al. (2017). For example, first a network can be used to extract a coarse segmentation of the pancreas, which is used in a subsequent superpixel-based pre-training task following the current approach. Afterward, the weights can be shared with a second segmentation network to improve the segmentation performance during fine-grained pancreas segmentation.

7 CONCLUSION

In summary, the current work explored the usage of superpixels to construct a pre-training task for self-supervised learning. During the task, superpixels are used to distort areas of an image, which the network has to reconstruct during the pre-training task. It has been found that superpixel-based context restoration adds a significant increase in performance compared to the baseline. Moreover, it outperforms existing

methods. The results indicate that superpixels can be promising in the development of pre-training tasks. However, more research is needed to take the limitations of the current research into account.

8 SELF-REFLECTION

Writing this thesis was a very valuable experience. Not only have I learned a lot in the domain of self-supervised learning and medical image analysis, I also learned practical things such as the importance of structuring your projects. Moreover, working together with the UMC was a valuable experience.

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012, November). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(11), 2274–2282. Retrieved from https://doi.org/10.1109/tpami.2012.120 doi: 10.1109/tpami.2012.120

Åkerberg, D., Ansari, D., Andersson, R., & Tingstedt, B. (2017). The effects of surgical exploration on survival of unresectable pancreatic carcinoma: A retrospective case-control study. *Journal of Biomedical Science and Engineering*, *10*(01), 1–9. Retrieved from https://doi.org/10.4236/jbise.2017.101001 doi: 10.4236/jbise.2017.101001

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., ... Norouzi, M. (2021). *Big self-supervised models advance medical image classification.* arXiv. Retrieved from https://arxiv.org/abs/2101.05224 doi: 10.48550/ARXIV.2101.05224

Bansal, M. A., Sharma, D. R., & Kathuria, D. M. (2022, January). A systematic review on data scarcity problem in deep learning: Solution and applications. *ACM Computing Surveys*. Retrieved from https://doi.org/10.1145/3502287 doi: 10.1145/3502287

Chen, Bentley, P., Mori, K., Misawa, K., Fujiwara, M., & Rueckert, D. (2019). Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, *58*, 101539. Retrieved from https://www.sciencedirect.com/science/article/pii/S1361841518304699 doi: https://doi.org/10.1016/j.media.2019.101539

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A simple framework for contrastive learning of visual representations.* arXiv. Retrieved from https://arxiv.org/abs/2002.05709 doi: 10.48550/ARXIV.2002.05709

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. E. (2020). Big self-supervised models are strong semi-supervised learners. *CoRR*, *abs/2006.10029*. Retrieved from https://arxiv.org/abs/2006.10029

Doersch, C., Gupta, A., & Efros, A. A. (2015, December). Unsupervised visual representation learning by context prediction. In *2015 IEEE international conference on computer vision (ICCV)*. IEEE. Retrieved from https://doi.org/10.1109/iccv.2015.167 doi: 10.1109/iccv.2015.167

Giddwani, B., Tekchandani, H., & Verma, S. (2020). Deep dilated v-net for 3d volume segmentation of pancreas in ct images. In *2020 7th international conference on signal processing and integrated networks (spin)* (p. 591-596). doi: 10.1109/SPIN48934.2020.9071339

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Aistats*.

Huang, M., Huang, C., Yuan, J., & Kong, D. (2021, July). A semiautomated deep learning approach for pancreas segmentation. *Journal of Healthcare Engineering*, *2021*, 1–10. Retrieved from https://doi.org/10.1155/2021/3284493 doi: 10.1155/2021/3284493

Imran, A.-A.-Z., Huang, C., Tang, H., Fan, W., Xiao, Y., Hao, D., ... Terzopoulos, D. (2020, December). Partly supervised multi-task learning. In *2020 19th IEEE international conference on machine learning and applications (ICMLA)*. IEEE. Retrieved from https://doi.org/10.1109/icmla51294.2020.00126 doi: 10.1109/icmla51294.2020.00126

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, *abs/1502.03167*. Retrieved from http://arxiv.org/abs/1502.03167

Irmak, E. (2021, April). Multi-classification of brain tumor MRI images using deep convolutional neural network with fully optimized framework. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, *45*(3), 1015–1036. Retrieved from https://doi.org/10.1007/s40998-021-00426-9 doi: 10.1007/s40998-021-00426-9

Kasinathan, G., & Jayakumar, S. (2022, January). Cloud-based lung tumor detection and stage classification using deep learning techniques. *BioMed Research International*, *2022*, 1–17. Retrieved from https://doi.org/10.1155/2022/4185835 doi: 10.1155/2022/4185835

Kayal, S., Chen, S., & de Bruijne, M. (2020). Region-of-interest guided supervoxel inpainting for self-supervision. In A. L. Martel et al. (Eds.), *Medical image computing and computer assisted intervention – miccai 2020* (pp. 500–509). Cham: Springer International Publishing.

Ker, J., Wang, L., Rao, J., & Lim, T. (2018). Deep learning applications in medical image analysis. *IEEE Access*, *6*, 9375-9389. doi: 10.1109/ACCESS.2017.2788044

Kim, B., Lee, H., Kim, K. J., Seo, J., Park, S., Shin, Y.-G., ... Lee, K. H. (2011, January). Comparison of three image comparison methods for the visual assessment of the image fidelity of compressed computed tomography images. *Medical Physics*, *38*(2), 836–844. Retrieved from https://doi.org/10.1118/1.3538925 doi: 10.1118/1.3538925

Kiyasseh, D., Zhu, T., & Clifton, D. A. (2021, 18–24 Jul). Clocs: Contrastive learning of cardiac signals across space, time, and patients. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (Vol. 139, pp. 5606–5615). PMLR. Retrieved from https://proceedings.mlr.press/v139/kiyasseh21a.html

Laoveeravat, P., Abhyankar, P. R., Brenner, A. R., Gabr, M. M., Habr, F. G., & Atsawarungruangkit, A. (2021, April). Artificial intelligence for pancreatic cancer detection: Recent development and future direction. *Artificial Intelligence in Gastroenterology*, *2*(2), 56–68. Retrieved from https://doi.org/10.35712/aig.v2.i2.56 doi: 10.35712/aig.v2.i2.56

Li, C., & Bovik, A. C. (2010, August). Content-partitioned structural similarity index for image quality assessment. *Signal Processing: Image Communication*, *25*(7), 517–526. Retrieved from https://doi.org/10.1016/j.image.2010.03.004 doi: 10.1016/j.image.2010.03.004

Li, M., Lian, F., Wang, C., & Guo, S. (2021, November). Accurate pancreas segmentation using multi-level pyramidal pooling residual u-net with adversarial mechanism. *BMC Medical Imaging*, *21*(1). Retrieved from https://doi.org/10.1186/s12880-021-00694-1 doi: 10.1186/s12880-021-00694-1

Lim, S.-H., Kim, Y. J., Park, Y.-H., Kim, D., Kim, K. G., & Lee, D.-H. (2022, March). Automated pancreas segmentation and volumetry using deep neural network on computed tomography. *Scientific Reports*, *12*(1). Retrieved from https://doi.org/10.1038/s41598-022-07848-3 doi: 10.1038/s41598-022-07848-3

Long, J., Shelhamer, E., & Darrell, T. (2014). Fully convolutional networks for semantic segmentation. *CoRR*, *abs/1411.4038*. Retrieved from http://arxiv.org/abs/1411.4038

Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., ... Yang, X. (2021). Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi: 10.1109/TPAMI.2021.3100536

Mahmoudi, T., Kouzahkanan, Z. M., Radmard, A. R., Kafieh, R., Salehnia, A., Davarpanah, A. H., ... Ahmadian, A. (2022, February). Segmentation of pancreatic ductal adenocarcinoma (PDAC) and surrounding vessels in CT images using deep convolutional neural networks and texture descriptors. *Scientific Reports*, *12*(1). Re-

trieved from https://doi.org/10.1038/s41598-022-07111-9 doi: 10.1038/s41598-022-07111-9

Man, Y., Huang, Y., Feng, J., Li, X., & Wu, F. (2019). Deep q learning driven ct pancreas segmentation with geometry-aware u-net. *IEEE Transactions on Medical Imaging*, *38*(8), 1971-1980. doi: 10.1109/TMI.2019.2911588

Minnema, J., Wolff, J., Koivisto, J., Lucka, F., Batenburg, K. J., Forouzanfar, T., & van Eijnatten, M. (2021, August). Comparison of convolutional neural network training strategies for cone-beam CT image segmentation. *Computer Methods and Programs in Biomedicine*, *207*, 106192. Retrieved from https://doi.org/10.1016/j.cmpb.2021.106192 doi: 10.1016/j.cmpb.2021.106192

Noroozi, M., & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. *CoRR*, *abs/1603.09246*. Retrieved from http://arxiv.org/abs/1603.09246

Oda, M., Shimizu, N., Karasawa, K., Nimura, Y., Kitasaka, T., Misawa, K., ... Mori, K. (2016). Regression forest-based atlas localization and direction specific atlas generation for pancreas segmentation. In *Medical image computing and computer-assisted intervention – MICCAI 2016* (pp. 556–563). Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-46723-8_64 doi: 10.1007/978-3-319-46723-8_64

Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., & Rueckert, D. (2020). Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. *CoRR*, *abs/2007.09886*. Retrieved from https://arxiv.org/abs/2007.09886

Panda, A., Korfiatis, P., Suman, G., Garg, S. K., Polley, E. C., Singh, D. P., ... Goenka, A. H. (2021, March). Two-stage deep learning model for fully automated pancreas segmentation on computed tomography: Comparison with intra-reader and inter-reader reliability at full and reduced radiation dose on an external dataset. *Medical Physics*, *48*(5), 2468–2481. Retrieved from https://doi.org/10.1002/mp.14782 doi: 10.1002/mp.14782

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. *CoRR*, *abs/1604.07379*. Retrieved from http://arxiv.org/abs/1604.07379

Petit, O., Thome, N., Rambour, C., & Soler, L. (2021). *U-net transformer: Self and cross attention for medical image segmentation.* arXiv. Retrieved from https://arxiv.org/abs/2103.06104 doi: 10.48550/ARXIV.2103.06104

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., & Matrisian, L. M. (2014, May). Projecting cancer incidence and

deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the united states. *Cancer Research*, *74*(11), 2913–2921. Retrieved from https://doi.org/10.1158/0008-5472.can-14-0155 doi: 10.1158/0008-5472.can-14-0155

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical image computing and computer-assisted intervention – miccai 2015* (pp. 234–241). Cham: Springer International Publishing.

Roth, H., Farag, A., Turkbey, E. B., Lu, L., Liu, J., & Summers, R. M. (2016). *Data from pancreas-ct.* The Cancer Imaging Archive. Retrieved from https://wiki.cancerimagingarchive.net/x/eIlXAQ doi: 10.7937/K9/TCIA.2016.TNB1KQBU

Salanitri, F. P., Bellitto, G., Irmakci, I., Palazzo, S., Bagci, U., & Spampinato, C. (2021). Hierarchical 3d feature learning forPancreas segmentation. In *Machine learning in medical imaging* (pp. 238–247). Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-030-87589-3_25 doi: 10.1007/978-3-030-87589-3_25

Shurrab, S., & Duwairi, R. (2021). *Self-supervised learning methods and applications in medical imaging analysis: A survey.* arXiv. Retrieved from https://arxiv.org/abs/2109.08685 doi: 10.48550/ARXIV.2109.08685

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 240–248). Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-67558-9_28 doi: 10.1007/978-3-319-67558-9_28

Tang, H., Liu, X., Han, K., Sun, S., Bai, N., Chen, X., ... Xie, X. (2020). *Spatial context-aware self-attention model for multi-organ segmentation.* arXiv. Retrieved from https://arxiv.org/abs/2012.09279 doi: 10.48550/ARXIV.2012.09279

Xu, J. (2021, August). A review of self-supervised learning methods in the field of medical image analysis. *International Journal of Image, Graphics and Signal Processing*, *13*(4), 33–46. Retrieved from https://doi.org/10.5815/ijigsp.2021.04.03 doi: 10.5815/ijigsp.2021.04.03

Xue, J., He, K., Nie, D., Adeli, E., Shi, Z., Lee, S.-W., ... Shen, D. (2021, April). Cascaded MultiTask 3-d fully convolutional networks for pancreas segmentation. *IEEE Transactions on Cybernetics*, *51*(4), 2153–2165. Retrieved from https://doi.org/10.1109/tcyb.2019.2955178 doi: 10.1109/tcyb.2019.2955178

Yan, Y., & Zhang, D. (2021, May). Multi-scale u-like network with attention

mechanism for automatic pancreas segmentation. *PLOS ONE*, *16*(5), e0252287. Retrieved from https://doi.org/10.1371/journal.pone.0252287 doi: 10.1371/journal.pone.0252287

Yao, X., Song, Y., & Liu, Z. (2019, December). Advances on pancreas segmentation: a review. *Multimedia Tools and Applications*, *79*(9-10), 6799–6821. Retrieved from https://doi.org/10.1007/s11042-019-08320-7 doi: 10.1007/s11042-019-08320-7

Yu, Q., Xie, L., Wang, Y., Zhou, Y., Fishman, E. K., & Yuille, A. L. (2017). Saliency transformation network: Incorporating multi-stage visual cues for pancreas segmentation. *CoRR*, *abs/1709.04518*. Retrieved from http://arxiv.org/abs/1709.04518

Zhai, X., Oliver, A., Kolesnikov, A., & Beyer, L. (2019). S4l: Self-supervised semi-supervised learning. *CoRR*, *abs/1905.03670*. Retrieved from http://arxiv.org/abs/1905.03670

Zhang, L., Zhang, L., Mou, X., & Zhang, D. (2012, December 1). A comprehensive evaluation of full reference image quality assessment algorithms. In *2012 ieee international conference on image processing, icip 2012 - proceedings* (pp. 1477–1480). (2012 19th IEEE International Conference on Image Processing, ICIP 2012 ; Conference date: 30-09-2012 Through 03-10-2012) doi: 10.1109/ICIP.2012.6467150

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2017, March). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, *3*(1), 47–57. Retrieved from https://doi.org/10.1109/tci.2016.2644865 doi: 10.1109/tci.2016.2644865

Zhou, Y., Xie, L., Shen, W., Wang, Y., Fishman, E., & Yuille, A. (2017, 09). A fixed-point model for pancreas segmentation in abdominal ct scans. In (p. 693-701). doi: 10.1007/978-3-319-66182-7_79

Appendix A

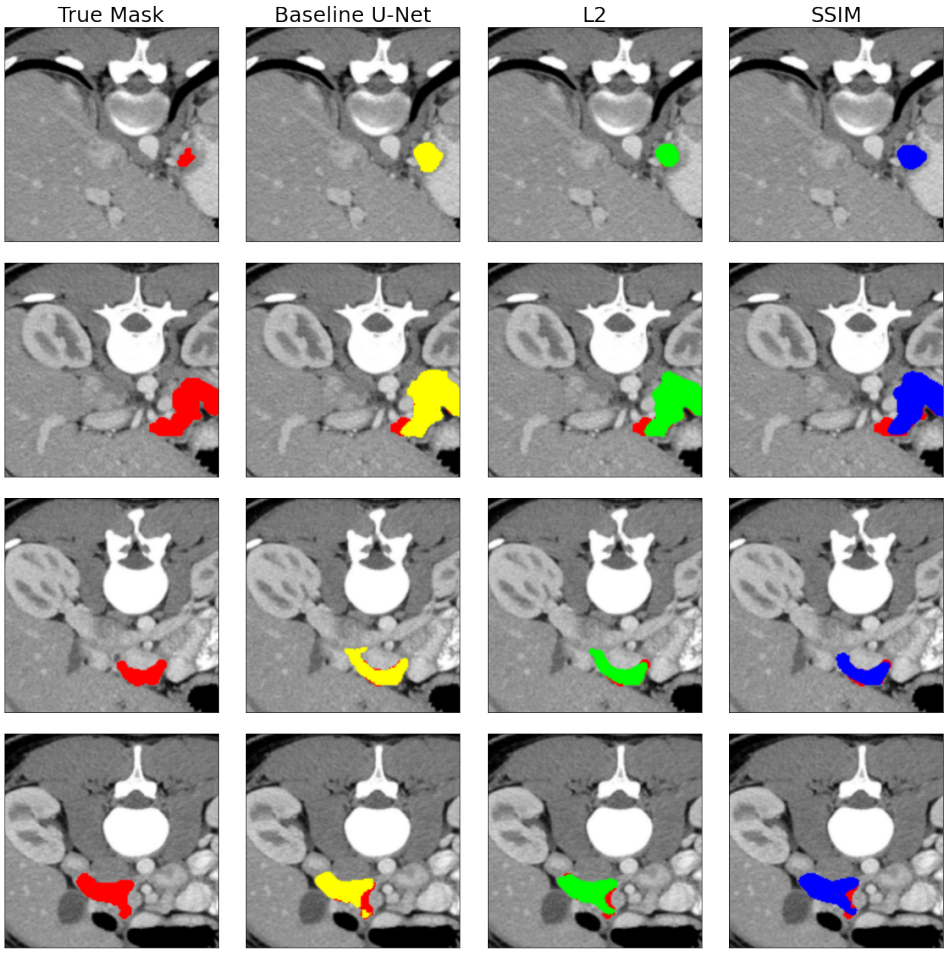

Figure 11: Example of the predicted masks of each model using patch-based pretraining.

Appendix B

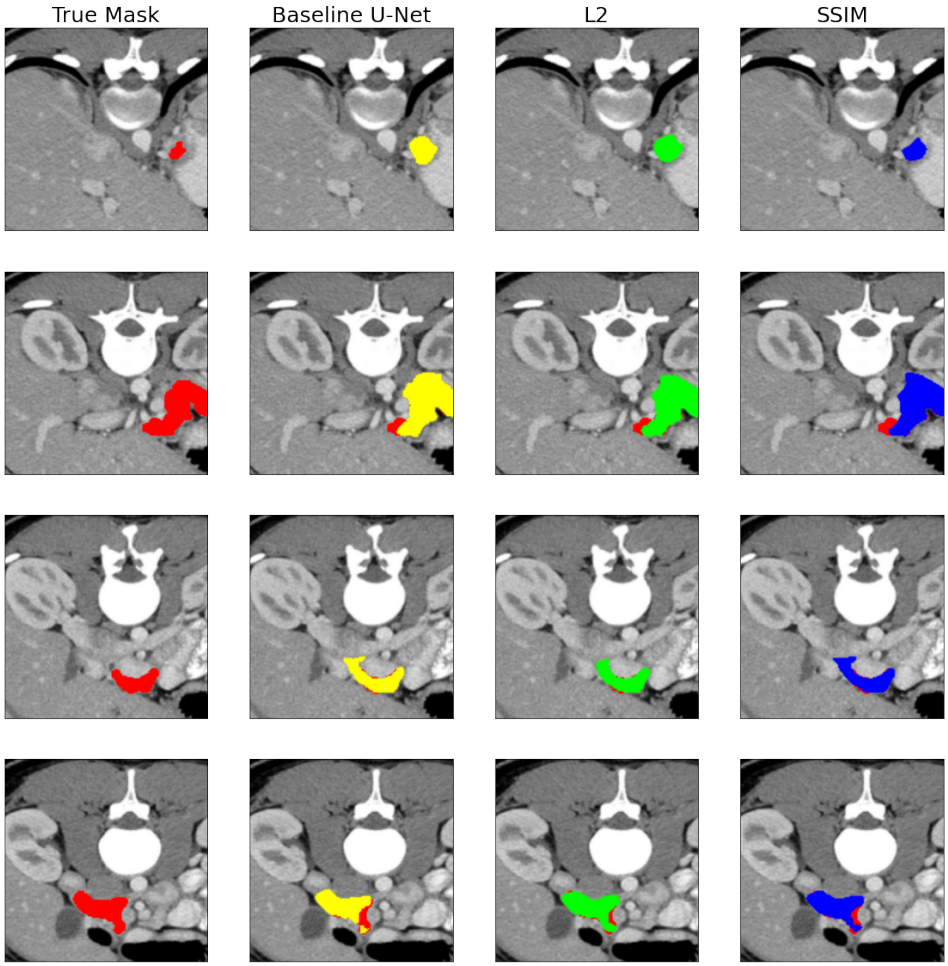

Figure 12: Example of the predicted masks of each model using superpixel-based pretraining.

Appendix C

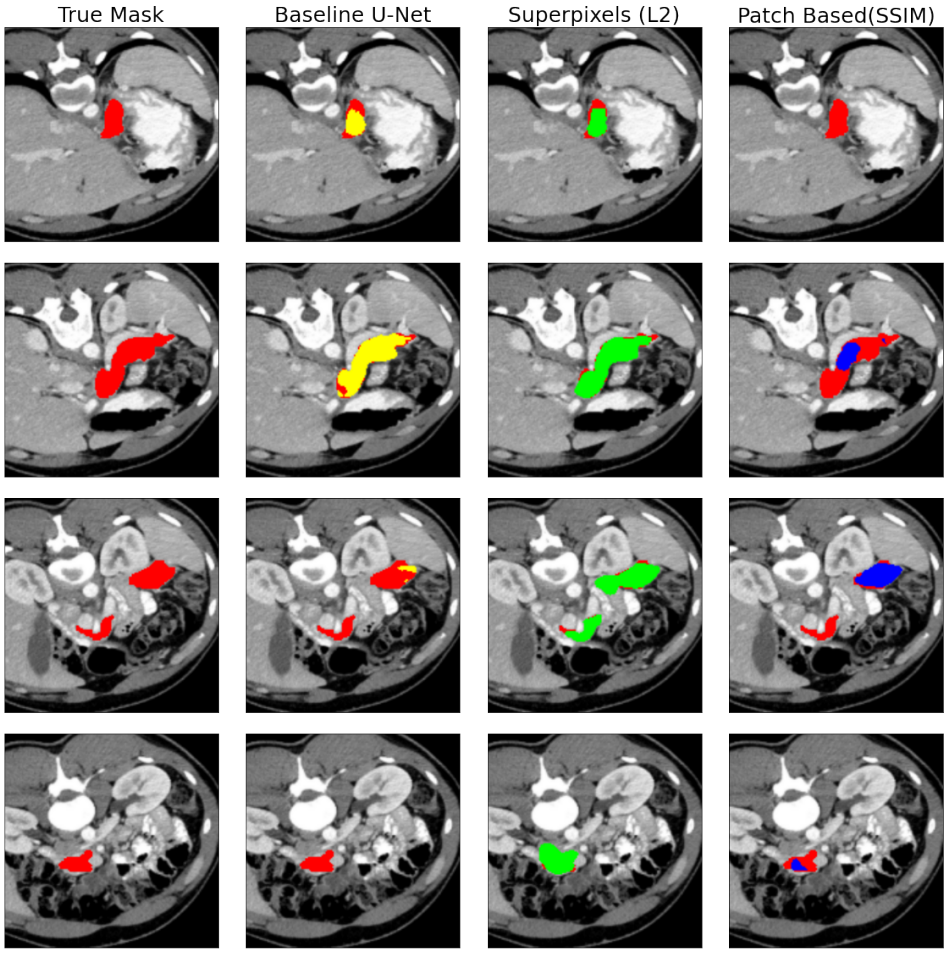

Figure 13: Comparison of predicted masks of different models.