Superpixel-based Context Restoration for Self-supervised Pancreas Segmentation from CT scans

Sander van Donkelaar[1], Lois Daamen[2], Paul Andel[2], Ralf Zoetekouw[3] and L.L. Sharon Ong[1]

[1] Dept. of CSAI, Tilburg University, Tilburg, the Netherlands
{s.van Donkelaar,l.l.ong}@tilburguniversity.edu
[2] Regional Academic Cancer Center Utrecht, University Medical Center Utrecht, the Netherlands.
{l.a.daamen-3,p.c.m.Andel-2}@umcutrecht.nl
[3] Datacation B.V., Eindhoven, the Netherlands
{r.zoetekouw}@datacation.nl

Abstract. Automatic segmentation of the pancreas can help research in 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. In recent years, deep learning has proven to be a powerful tool for pancreas segmentation. However, deep learning models in medical image analysis suffer from data scarcity: the lack of annotated data poses a significant drawback in developing new models. One possible solution is self-supervised learning which comprises of an unsupervised pre-training stage followed by a subsequent supervised learning stage. This paper presents a superpixel based approach to construct a pre-training task for self-supervised learning for pancreas segmentation. We corrupt the CT images segmented with superpixels by replacing random segments with intensity values randomly sampled from the image. The weights learnt when the model reconstructs the image are used to initialize network weights in the subsequent segmentation task. We used 59 CT scans from the AbdomenCT-1k dataset for pre-training and 82 CT scans from the NIH pancreas-CT dataset for the segmentation task. We achieved an increase in performance with our approach compared to the randomly initalized weights, as contextual image features are learnt via this context restoration. Moreover, our approach outperforms existing context restoration approaches using patch based methods.

Keywords: Context restoration · Self-supervised learning · Superpixels · Pancreas Segmentation.

1 Introduction

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 [27]. 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 [7]. Therefore, these techniques have shown limited performance and generalization capability. Deep learning-based approaches greatly increased performance in pancreas segmentation [27], 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) [25], U-net [12, 18, 31] and V-Net [6]. These supervised CNN architectures can either be in 2D, 3D or hybrid structure.

Supervised models are all limited by the amount of annotated data that is available. When annotated data is scarce, it is harder for a network to learn a heterogeneous representation that encapsulates the variation in the data. This is especially relevant for the pancreas due to its high anatomical variability. Self supervised methods have been developed to tackle this.

Self-supervised learning (SSL) is a hybrid learning approach, comprising of an unsupervised pre-training stage followed by supervised fine-tuning stage. The unsupervised pre-training stage leverages supervisory signals from the data itself, which allows it to learn a representation that captures the underlying structure [21]. This representation is functional at a later stage, as the model has learnt a set of features that are useful in the subsequent task [28]. The knowledge is transferred by initializing a part of the network for the subsequent task with the weights that are learned in the unsupervised pre-training task. This way, unstructured medical data, such as unannotated CT scans, can be utilized.

This work investigates a *context-restoration* strategy for self-supervised learning with superpixels to improve pancreas segmentation. Most self-supervised learning techniques that rely on reconstruction of distorted images use uninformed and random regions to corrupt an image [10]. For example, in the context-restoration method as described in [3], images are distorted by swapping subpatches 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. 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 [16]. Likewise, it has been found that learning contextual information in CT scans improves performance in deep learning networks [18, 23].

Superpixels are a subgroup of pixels in an image that share common characteristics, such as their location and pixel intensities. 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. In our superpixel-based SSL strategy for pancreas segmentation, superpixels are created from CT scans from the AbdomenCT-1k dataset [14]. Several superpixel segments are randomly selected and distorted. Afterward, a neural network is trained to reconstruct the image. As 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. The weights learnt when the model reconstructs the image are used to initialize network weights in the downstream segmentation task. We used a separate dataset, the NIH pancreas-CT dataset [20] to evaluate our segmentation task. We show our approach outperformed randomly initialized weights and a patch-based context restoration approach to SSL [3].

2 Related Work on Self-Supervised Learning

The usage of SSL approaches in the medical domain has received relatively little attention [2]. 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.

Contrast-based Self-Supervised 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, [2] 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.

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 [4]. 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-the-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 [24]. Other approaches exist within self-supervision frame-

works, such as context-based learning and generative learning, which do not require the construction of negative examples.

Generative Self-Supervised Learning Generative approaches are aimed at reconstructing distorted images volumes. For example, [17] proposed a method in which a CNN was learned to inpaint removed sub-patches 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 [3].

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, which have been found to increase performance in the subsequent task. For example, [15] proposed 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 high (362880), the method was challenging in terms of model complexity and memory.

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 [8]. Other approaches to self-supervised learning include predicting the position in a 3×3 grid between a central patch and its surrounding patches [5]. However, it has been found that the performance gains are limited since the network could complete the task using relatively trivial features [3]. 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.

A region-of-interest guided supervoxel inpainting task was proposed by [10]. 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 this 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-supervoxel 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 investigates this further and take the limitations of current research into account.

3 Dataset and Data Preparation

In the current study, the NIH pancreas-CT dataset [20] is used 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 $512 \times 512 \times N$, 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 [14]. The resolution of the CT scans is $512 \times 512 \times N$, 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. Both datasets are publicly available.

Pre-processing of the data consists of several steps. First, each image is clipped between $[-100, 240]$ HU (Housfield Units), following the protocol in [26, 12]. 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 Methods

Our self-supervised learning framework comprises of a pre-training task with a deep learning model (a 2D U-Net) for superpixel context restoration. The weights are transferred to another U-Net to perform segmentation. The pre-training task is designed to yield a set of layer weights that encapsulate useful information for the final task of segmentation. A subset of the weights is then used to initialize the weights of the downstream segmentation model. Figure 1 shows an overview of our framework.

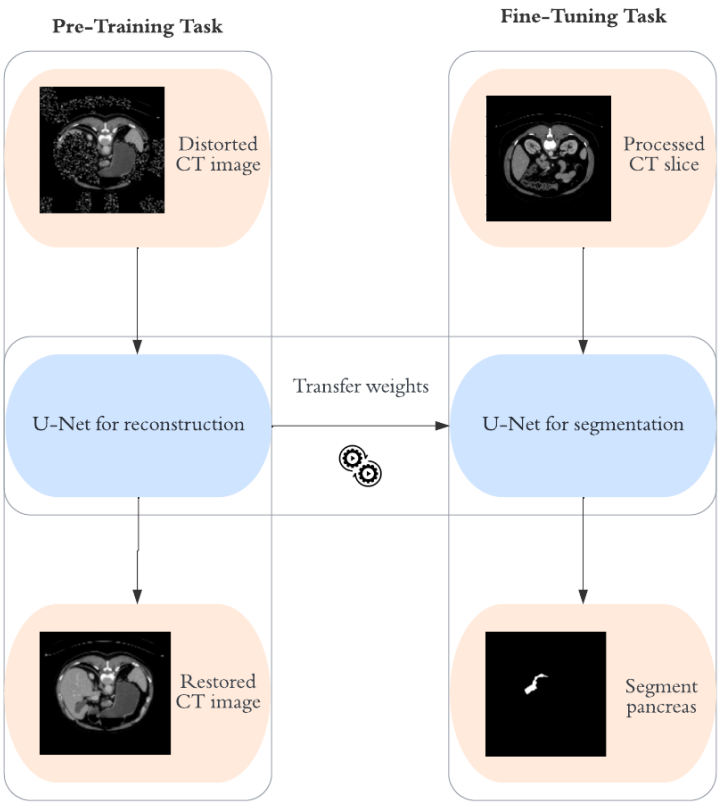

Fig. 1. Overview of the experimental set-up. The same U-Net architecture is applied in the pre-training and fine-tuning task. The pretraining task is trained to reconstruct a distorted image. The weights (from pretraining) are transferred to the fine-tuning, segmentation task. The output layer of the pretraining stage is a *ReLu* function whereas for a fine-tuning task, we use a sigmoid function.

4.1 Superpixel based context restoration

Superpixels are a connected regions of pixels in an image which share common characteristics such as pixel intensities or texture. The areas of the superpixel segments do not overlap and adhere well to object boundaries within the CT scans which can be utilized to segment parts of the image automatically. We leverage this characteristic to create a self-supervised learning task.

In the superpixel-based context-restoration method, a 2D U-Net is learned to approximate the function $g(x_d)$, where x_d is the distorted superpixel image, and $g(x_d)$ yields the original image x_o. This task is constructed as follows: first, each slice is segmented into N segments by using the SLIC algorithm (see Appendix A). After the image has been segmented, K segments are randomly chosen, and the intensity values are replaced with intensity values that are ran-

domly sampled from the image. The variable, 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 such that $K = R \cdot N$. Pixels values are randomly sampled from the original image to perserve the intensity distribution. This is important for the network to learn features belonging to a specific domain [3].

Algorithm 1 Image distortion from Superpixel Segmentation

Require: Image x_o

1: Transform image x_o into N superpixel segments. We denote the output as $S\epsilon[S_1, S_2...S_N]$, where S_i is a superpixel segment.
2: Randomly sample K superpixel segments into S', which yields $S'\epsilon[S_1, S_2 \dots S_K]$.
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.

4.2 U-Net Architecture

The model for both the pre-training tasks and pancreas segmentation is a 2D U-Net [19], which is used extensively in medical image segmentation tasks. The U-Net architecture is based upon the *fully convolutional network* [13], and 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. Unlike other encoder-decoder architectures, the contracting and expanding parts of the U-Net are not fully decoupled, due to skip connections. Feature maps of the convolutional layers in the contracting 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 [19].

The CT scan is grayscale. Hence, the input map is of size $208 \times 224 \times 1$, 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 size of 2×2. Batch Normalization is applied after each convolutional layer to make the network train faster and more stable [9]. 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. An overview of the architecture in shown in figure 2.

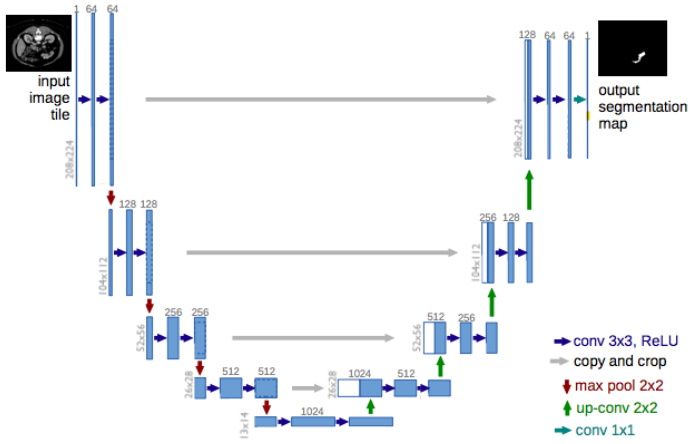

Fig. 2. Overview of the used U-Net architecture, figure adapted from [19]

The loss during the segmentation task is the Dice Loss, which is $\mathcal{L}_{Dice} = 1 - DSC$. Here, the Dice score coefficient (DSC) measures the overlap between the prediction and ground truth, which is given as:

$$DSC = \frac{2TP}{2TP + FP + TN} \tag{1}$$

The 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 [22]. Therefore, this is especially relevant for pancreas segmentation since the pancreas only makes up a small part of each CT scan [11]. 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 [31].

4.3 Loss functions Pretraining

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.

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 $L2_{loss}(x, y) = \sum_{i=1}^{N}(x_i - y_i)^2$.

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 [29]. 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 [30], which resembles how a human would evaluate the similarity between two images.

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.

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

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

$$C(x, y) = \frac{2\sigma_x\sigma_y + C_2}{2\sigma_x^2\sigma_y^2 + C_2} \tag{3}$$

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

$$S(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x^2\sigma_y^2 + C_3} \tag{4}$$

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

$$SSIM(x, y) = \alpha L(x, y) \cdot \beta C(x, y) \cdot \gamma S(x, y) \tag{5}$$

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

$$\mathcal{L}_{SSIM}(x, y) = 1 - SSIM(x, y) \tag{6}$$

4.4 Implementation Details

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.

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 [7]. 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. 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

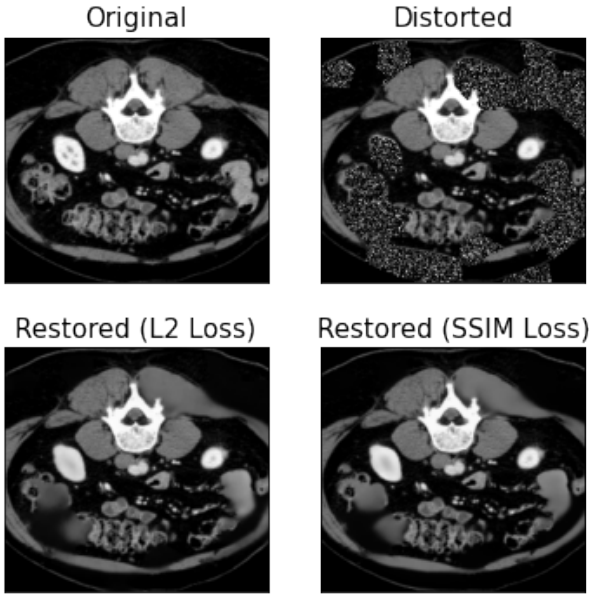

Fig. 3. Example of the restored images using superpixel-based pre-training.

Superpixels in our approach are generated by using the SLIC (Simple Linear Iterative Clustering) algorithm. The SLIC is initialized with 100 segments

and a compactness of 0.05, which is the trade-off for color-similarity and proximity.Figure 3 shows 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. Figure 4 shows a qualitative overview in which the predictions of both loss functions for superpixel-based pre-training for patient 70 are compared.

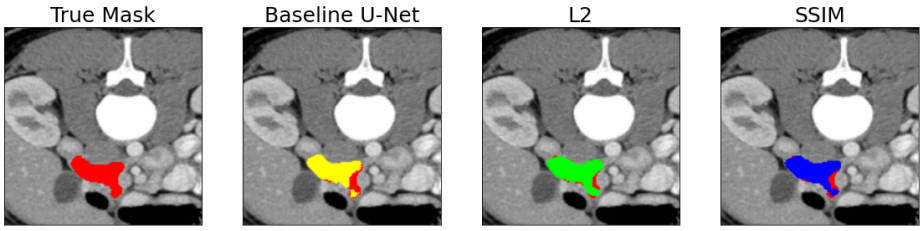

Fig. 4. Example of the predicted masks of each model using superpixel-based pretraining The ground truth is shown in red in all the images. The second, third and fourth column of figures show the U-Net segmentation solution with randomly initialized weights (yellow) and with weights transferred from our superpixel context restoration approach with L2 loss (green) and SSIM loss (blue) functions.

We provide quantitative results, comparing our approach with a U-Net pretrained with weights from [3] and the baseline of randomly initialized weights. Given these results, it is clear that pre-training with superpixels yields the best performance. Moreover, 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 1. Comparison of all three pre-training methods with different loss functions. PB refers to Patch Based pre-training by [3] and SP refers to our superpixels pretraining approach. As we used a four-fold cross validation, we provide the average, standard deviation, minimum and maximum Dice score and Jaccard index across the folds. Highest scores are shown in bold for each column are shown in bold. Results show that our Superpixel pretraining approach with the L2 loss function (marked with a *) outperformed the other methods.

This can also be seen when performing a qualitative assessment 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 5). It is clear that the the irregular and disconnected shapes of the pancreas are detected much better in comparison to other models.

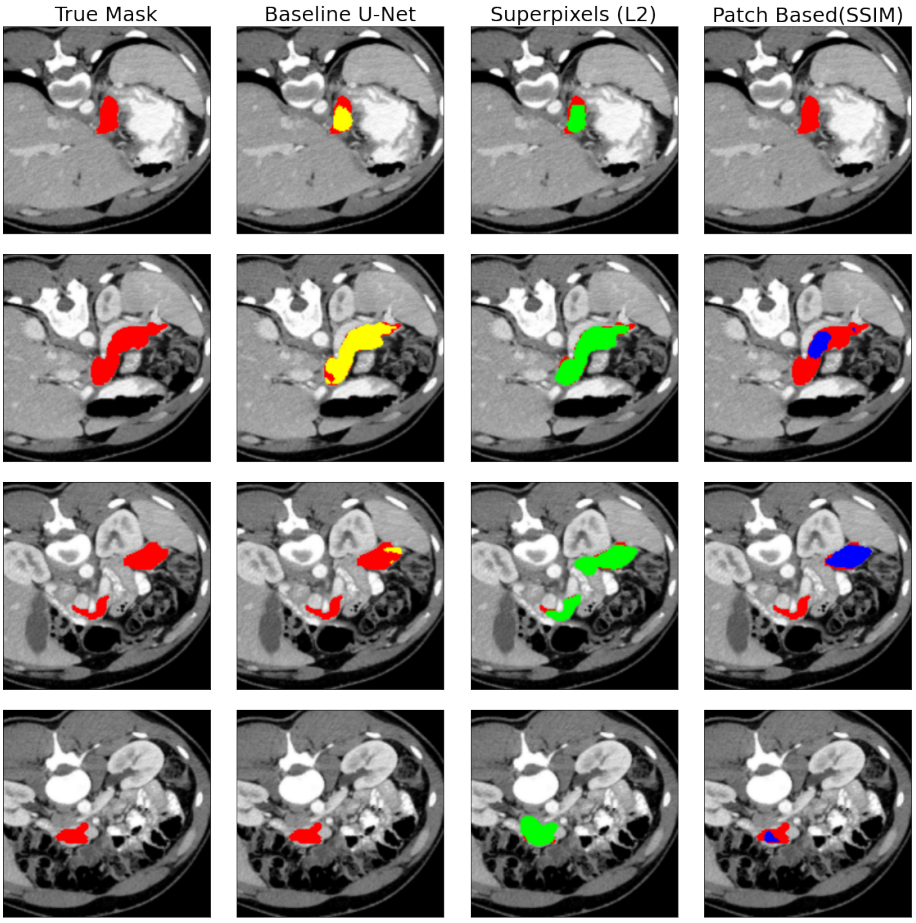

Fig. 5. The ground truth is shown in red in all the images. The second, third and fourth coloum of figures show the U-Net segmentation solution with randomly initialized weights (yellow), weights transfered from pretraining with superpixels context restoration (green) and weights transfered from the patch based context restoration (blue). Results show that the superpixels based approach has a higher overlap with the ground truth mask.

6 Discussion

The current research is the first to investigate the effect of superpixel-based context restoration in the context of 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 [10].

Future work includes comparing our approach to other SSL paradigms, such as methods from contrast- and context-based SSL. The effect of different hyperparameters should be further investigated as well. 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 [2]. 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 [31, 12]. 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.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2274–2282 (2012)
2. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., Norouzi, M.: Big self-supervised models advance medical image classification. In: Proceedings of the

IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3458–3468 (2021)
3. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D.: Self-supervised learning for medical image analysis using image context restoration. Medical Image Analysis **58**, 101539 (2019)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning (IMCL). pp. 22243—22255 (2020)
5. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1422–1430 (2015)
6. Giddwani, B., Tekchandani, H., Verma, S.: Deep dilated V-Net for 3D volume segmentation of pancreas in CT images. In: Proceedings of the 7th International Conference on Signal Processing and Integrated Networks (SPIN). pp. 591–596 (2020)
7. Huang, M., Huang, C., Yuan, J., Kong, D.: A semiautomated deep learning approach for pancreas segmentation. Journal of Healthcare Engineering **2021**, 1–10 (2021)
8. Imran, A.A.Z., Huang, C., Tang, H., Fan, W., Xiao, Y., Hao, D., Qian, Z., Terzopoulos, D.: Partly supervised multi-task learning. In: Proceedings of 19th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 769–774 (2020)
9. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning (IMCL). pp. 448–456 (2015)
10. Kayal, S., Chen, S., de Bruijne, M.: Region-of-interest guided supervoxel inpainting for self-supervision. In: Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 500–509 (2020)
11. Laoveeravat, P., Abhyankar, P.R., Brenner, A.R., Gabr, M.M., Habr, F.G., Atsawarungruangkit, A.: Artificial intelligence for pancreatic cancer detection: Recent development and future direction. Artificial Intelligence in Gastroenterology **2**(2), 56–68 (2021)
12. Li, M., Lian, F., Wang, C., Guo, S.: Accurate pancreas segmentation using multi-level pyramidal pooling residual U-Net with adversarial mechanism. BMC Medical Imaging **21**(1), 168 (2021)
13. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3431–3440 (2015)
14. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J., Yang, X.: AbdomenCT-1K: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2023)
15. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Proceedings of European Conference on Computer Vision, (ECCV). pp. 69–84 (2016)
16. Oda, M., Shimizu, N., Karasawa, K., Nimura, Y., Kitasaka, T., Misawa, K., Fujiwara, M., Rueckert, D., Mori, K.: Regression forest-based atlas localization and direction specific atlas generation for pancreas segmentation. In: Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 556–563 (2016)

17. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2536–2544 (2016)
18. Petit, O., Thome, N., Rambour, C., Soler, L.: U-Net transformer: Self and cross attention for medical image segmentation. arXiv (2021), https://arxiv.org/abs/2103.06104
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
20. Roth, H., Farag, A., Turkbey, E.B., Lu, L., Liu, J., Summers, R.M.: Data from Pancreas-CT. The Cancer Imaging Archive (2016), http://doi.org/10.7937/K9/TCIA.2016.tNB1kqBU
21. Shurrab, S., Duwairi, R.: Self-supervised learning methods and applications in medical imaging analysis: a survey. PeerJ Computer Science **8**(19), e1045 (Jul 2022)
22. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: 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 - Third International Workshop, DLMIA and 7th International Workshop, ML-CDS, held in Conjunction with MICCAI, pp. 240–248 (2017)
23. Tang, H., Liu, X., Han, K., Sun, S., Bai, N., Chen, X., Qian, H., Liu, Y., Xie, X.: Spatial context-aware self-attention model for multi-organ segmentation. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 938–948 (2021)
24. Xu, J.: 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 (2021)
25. Xue, J., He, K., Nie, D., Adeli, E., Shi, Z., Lee, S.W., Zheng, Y., Liu, X., Li, D., Shen, D.: Cascaded MultiTask 3-d fully convolutional networks for pancreas segmentation. IEEE Transactions on Cybernetics **51**(4), 2153–2165 (2021)
26. Yan, Y., Zhang, D.: Multi-scale U-like network with attention mechanism for automatic pancreas segmentation. PLOS ONE **16**(5), e0252287 (2021)
27. Yao, X., Song, Y., Liu, Z.: Advances on pancreas segmentation: a review. Multimedia Tools and Applications **79**(9-10), 6799–6821 (2019)
28. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S^4l: Self-supervised semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1476–1485 (2019)
29. Zhang, L., Zhang, L., Mou, X., Zhang, D.: A comprehensive evaluation of full reference image quality assessment algorithms. In: Proceedings of IEEE International Conference on Image Processing ICIP. pp. 1477–1480 (2012)
30. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Transactions on Computational Imaging **3**(1), 47–57 (2017)
31. Zhou, Y., Xie, L., Shen, W., Wang, Y., Fishman, E., Yuille, A.: A fixed-point model for pancreas segmentation in abdominal CT scans. In: Proceedings of the 20th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 693–701 (2017)

Appendix

Superpixel segemntation

In the current study, superpixels are generated by using the SLIC (Simple Linear Iterative Clustering) algorithm [1]. This algorithm generates superpixels by clustering pixels based on color similarity and closeness in the image plane. Since the CT images are grayscale, clustering is performed in three-dimensional [ixy] space, i is the intensity and $[xy]$ is the pixel position. The algorithm works through several steps explained in more detail below.

1. Denoting N as the number of pixels in the input image and K as the number of desired superpixel clusters, the first step consists of initializing K cluster centers at regular grid intervals S. Here $S = \sqrt{(N/K)}$. 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. Next, each pixel is assigned to the nearest cluster within the search area. 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 disjoint pixels.