Artificial Intelligence Review (2025) 58:220
https://doi.org/10.1007/s10462-024-11050-4

Deep learning for pancreas segmentation on computed tomography: a systematic review

Andrea Moglia[1] · **Matteo Cavicchioli**[1,2] · **Luca Mainardi**[1] · **Pietro Cerveri**[1,3]

Accepted: 29 November 2024 / Published online: 3 May 2025

Abstract
Pancreas segmentation has been traditionally challenging due to its small size in computed tomography abdominal volumes, high variability of shape and positions among patients, and blurred boundaries due to low contrast between the pancreas and surrounding organs. Many deep learning models for pancreas segmentation have been proposed in the past few years. We present a thorough systematic review based on the Preferred Reporting Items for Systematic Reviews and Meta-analyses statement. The literature search was conducted on PubMed, Web of Science, Scopus, and IEEE Xplore on original studies published in peer-reviewed journals from 2013 to 2023. Overall, 130 studies were retrieved. We initially provide an overview of the technical background of the most common network architectures and publicly available datasets. Then, the analysis of the studies combining visual presentation in tabular form and text description is reported. The tables group the studies specifying the application, dataset size, design (model architecture, learning strategy, loss function, and training protocol), results, and main contributions. We first analyze the studies focusing on parenchyma segmentation using datasets with only pancreas annotations, followed by those using datasets with multi-organ annotations. Then, we analyze the studies on the segmentation of tumors, cysts, and inflammation. The studies are clustered according to the different deep learning architectures. Finally, we discuss the main findings from the published literature, the challenges, and the directions for future research on the clinical need, deep learning and foundation models, datasets, and clinical translation.

Keywords Deep learning pancreas segmentation · Artificial intelligence pancreas segmentation · Pancreas segmentation · Pancreas tumor segmentation

1 Introduction

The pancreas is a small J-like-shaped glandular organ, located inside the deep part of the abdomen, and subdivided into three regions, namely head, body, and tail. A healthy pancreas generally occupies around 0.5% of a computer tomography (CT) abdominal volume (Zhou et al. 2023). In some patients, the healthy tissue may be affected by disorders such

Extended author information available on the last page of the article

as inflammation, e.g., pancreatitis, while in more severe cases it may be affected by cysts and tumors. The latter are particularly insidious because they generate few symptoms and are often diagnosed at an advanced stage. In addition, they are very aggressive and lethal. Pancreas tumors are the fourth leading cause of death among all cancer types in the United States for the male gender and the third one for the female gender Siegel et al. (2024). A five-year survival rate of 13% was reported in the United States in the period 2013–2019, which is the lowest one among all cancer types Siegel et al. (2024). Its early-stage diagnosis and treatment are notably challenging due to the generally ambiguous symptoms often occurring when the disease has already reached an advanced stage, namely III and IV Falconi et al. (2016). The diagnosis of a pancreatic mass involves clinical assessment, laboratory testing, and advanced imaging techniques. Patient history and physical examination are initially performed to identify symptoms and risk factors. Laboratory tests on blood samples are subsequently conducted to measure CA 19-9 marker, before imaging tests. Ultrasound is usually the preliminary imaging assessment tool, followed by CT or magnetic resonance imaging (MRI) to delineate the tumor size and location in more detail. In particular CT scans are pivotal for staging cancer, evaluating its resectability, and planning surgical interventions. In fact pancreas surgery requires accurate recognition of anatomical variations and the spatial relationships of the tumor location with the surrounding vessels and organs in order to determine the optimal location of the pancreas resection Miyamoto et al. (2024). MRI provides excellent soft tissue contrast, highlighting vascular and ductal details Şolea et al. (2024). The recent guidelines of the European Society for Medical Oncology recommended CT as the primary modality for detailing tumor characteristics and spread Conroy et al. (2023).

Given the rising demand for enhanced early detection of pancreatic diseases, precise segmentation from medical images has become imperative. In this regard, its segmentation from medical images is a prerequisite for accurate computer-assisted diagnosis, surgical planning, and disease progress monitoring (e.g., post-surgical follow-up, and radiotherapy).

1.1 Methods of abdominal organ segmentation

The methods of abdominal organ segmentation can be divided into human intervention-based, target-based, region-based, and learning-based. The first type can be further subdivided into manual, semi-automatic, and automatic methods. Traditionally, medical image segmentation, including pancreas segmentation, has relied heavily on manual delineation by expert radiologists. This poses critical challenges including inter- and intra-observer variability, time-consuming labor, and subjective interpretation. Limited availability of experts, human error, and scalability issues further complicate the process. Extensive training requirements and reproducibility concerns hinder the widespread adoption of manual segmentation methods (Chen et al. 2022c). Examples of semiautomatic approaches are statistical shape models and multi-atlas label fusion. The former involves the co-registration of images in a training dataset to derive anatomical correspondences, building a statistical model of the distribution of shapes and/or appearances of the corresponding anatomy in the training data, and fitting the resulting model to new images (Gibson et al. 2018). The latter registers images of a training set with new images and combines the reference segmentation to generate new segmentations (Gibson et al. 2018). However, they suffer from high variability in organ position, shape, and appearance among patients, in addition to soft tissue

deformation (Gibson et al. 2018). In automatic segmentation models, there is no human intervention. An example is the bottom-up approach proposed for pancreas segmentation based on simple linear iterative clustering to group similar pixels into superpixels, followed by classification of features of superpixels and patches (Farag et al. 2014). Target-based approaches are split into multi-organ and single-organ ones. Region-based segmentation methods can be divided into direct and two-stage (coarse-fine) methods (Chen et al. 2022d). The former approach directly uses labeled images to segment the organ. In contrast, two-stage methods are cascaded. They first train a localization network to obtain the pancreas region (coarse stage), and then use the location result to train a second model for segmentation (fine stage) (Chen et al. 2022d). For instance, localizing the pancreas CT scans before performing segmentation has two advantages. First, the peripheral regions with very similar intensity or textural properties to the pancreas can be easily removed. Second, specifying the location of the pancreas reduces the sizes of the original CT scans, with a benefit in terms of computational costs, especially for 3D CT scans (Qureshi et al. 2022).

Learning-based methods extract meaningful features from annotated CT scans to distinguish target organs (Ma et al. 2022b). They can be categorized into supervised learning methods if the datasets are labeled; partially supervised to learn a multi-organ segmentation model from different single-organ labeled datasets; semi-supervised learning if a small amount of labeled is combined with a large amount of unlabelled data to extract knowledge from the unlabelled data, e.g. generating pseudo annotations for unlabeled examples, which are used jointly with labeled data to train the model (pseudo-labeling); unsupervised learning when the model learns the underlying patterns or hidden data structures without labels; self-supervised learning, a form of unsupervised learning where input labels are generated from unlabeled data without external supervision (it can be realized with a pretext task to learn representation features for downstream tasks, or with contrastive learning to maximize similarity for positive pairs of data and minimize it for negative pairs); weakly supervised learning using weak annotations like scribbles; and continual learning to learn new tasks without forgetting the learned ones (Ma et al. 2022b; Chen et al. 2022c).

Deep learning (DL) approaches have been proposed to reduce inter-reader variation, and the resources in terms of time and costs related to the involvement of skilled clinicians (Shamshad et al. 2023). With the recent advancements in DL, convolutional neural networks (CNNs) were introduced as a learning-based method and applied to different tasks of medical imaging, e.g. classification, detection, and segmentation (Chen et al. 2022c). The CNNs for medical imaging segmentation can be categorized into 2D, 2.5D, and 3D models. In 2D networks, the data are sliced along one of the three image planes (axial, sagittal, and coronal). Then, the 2D slices are sent as the input to the DL model (Zhang et al. 2021d). They are computationally efficient but lack the spatial context to extract the interslice information embedded in volumetric CT data (Wang et al. 2021c). In contrast, 3D models use the entire CT volume as the input of the network, which can capture 3D spatial information of the CT volume. However, they are computationally expensive (Yan and Zhang 2021). In 2.5D models, three 2D models segment the input image separately in three image planes. Then, the segmentation is obtained by fusing the results of the three 2D models, for instance through voting (Zhang et al. 2021d). 2.5D models represent a compromise between 2D and 3D ones, by making up for the lack of spatial context information of 2D models, but at the same time reducing the computational cost of 3D models (Dai et al. 2023).

1.2 Challenges in deep learning for pancreas segmentation

From a medical imaging perspective, the segmentation of the pancreas is very challenging, but it is even more difficult in the case of tumors and inflammations since the conditions are exacerbated. Firstly, whereas the pancreas is very small, typically representing a small fraction of the CT volume, pancreatic tumors are even smaller, with most of them accounting for less than 0.1% of the entire CT abdominal volume. The tiny size of the pancreas and its lesions raise the issue of class imbalance where the positive class (parenchyma or tumors) occupies an extremely low number of positive pixels in contrast with the negative class (background). Secondly, the contrast between the pancreas and its surrounding organs in CT scans is weak, which is caused by the similar range of voxel intensities. As a consequence, the boundaries of the pancreas and tumors are blurred, and the contrast with surrounding tissues is low, especially at the head of the pancreas. As a result, it is difficult to distinguish not only between the pancreas and the duodenum but also between the tissue (parenchyma) and tumors of the pancreas Zhou et al. (2023); Dai et al. (2023). Likewise, the segmentation of an inflamed pancreas is more challenging than a normal one since it invades the surrounding organs causing blurry boundaries, and it has higher shape, size, and location variability than the normal pancreas Deng et al. (2023). As such, boundary errors remain critical in preoperative planning of the pancreas, such as tumor resections and organ transplantation. Thirdly, the pancreas exhibits an irregular shape and susceptibility to deformation, complicating accurate segmentation. Anatomical variations in size, shape, and tumor positioning among patients, particularly the diverse locations of pancreatic tumors, pose challenges in distinguishing parenchyma from cancerous masses Zhou et al. (2023); Dai et al. (2023). Lastly, differences in commercial CT scanners and CT phases can lead to significant variances in organ appearances Ma et al. (2022b).

From a DL perspective, the learning-based models for segmentation face several challenges. First, since supervised learning models need pixel-level annotations they rely on manual annotation from expert clinicians, thus sharing the same challenges as manual annotation. Second, a common problem for semi-supervised learning is the violation of the same statistic distribution for both labeled and unlabeled data (Chen et al. 2022c). Moreover, the performances can degrade due to incorrectly generated pseudo-labels (Chen et al. 2022c). Third, for self-supervised learning, designing pretext tasks for medical imaging can be challenging (Chen et al. 2022c). On the other hand, contrastive learning requires relatively large unlabeled datasets, which are difficult to obtain to maintain patient privacy. Lastly, weakly learning struggles to segment the organ boundaries accurately due to the limited information contained in weak annotations, e.g. scribbles and bounding boxes (Shi et al. 2023).

1.3 Work motivation

Progress in the past decade in DL has led to continuous improvements in medical imaging, including pancreas segmentation. An overview of pancreas segmentation based on DL is depicted in Fig. 1. Even though in the last years several reviews have delved into pancreas segmentation from CT scans using AI (Ghorpade et al. 2023; Kumar et al. 2019; Huang et al. 2022a; Yao et al. 2019; Aljabri and AlGhamdi 2022; Rehman and Khan 2020; Senkyire and Liu 2021), our preliminary literature search has unearthed a significant number of studies overlooked by them. These considerations underscore the necessity for an updated

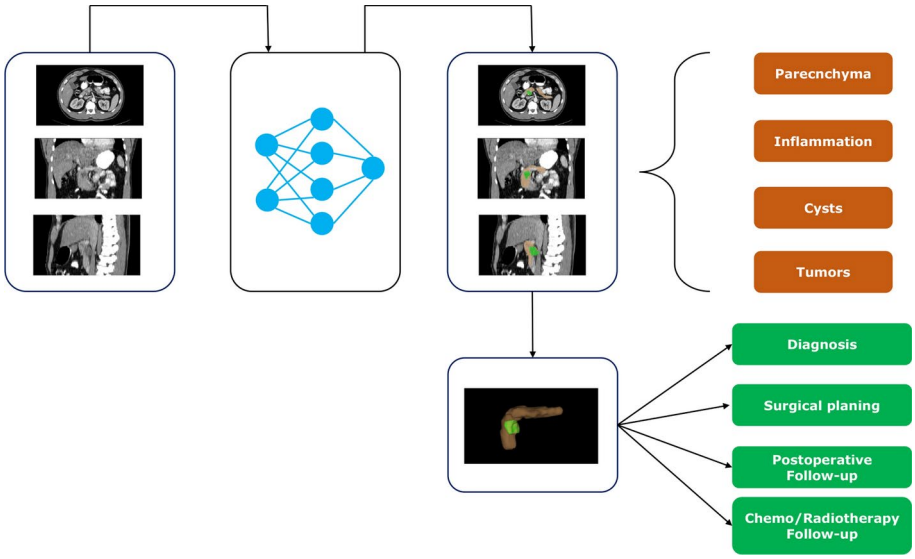

Fig. 1 Overview of pancreas segmentation based on DL. Radiological images are processed by neural network models outputting masks of the organ or lesions (e.g. cysts, and tumors). Applications include diagnosis, surgical planning, postoperative follow-up, and chemo/radiotherapy follow-up

systematic review to comprehensively cover the latest advancements in the field. Consequently, the goal of this review is to present systematically an in-depth analysis of DL for the segmentation of the parenchyma, tumors, cysts, and inflammation of the pancreas starting from CT scans.

1.4 Structure and contribution of the work

The review is structured as follows. In Sect. 2 we describe the method to perform the literature search and extract the included studies. We also report the limitations of the published reviews in the field. In Sect. 3 we illustrate the main DL architectures, the available public datasets, metrics, and loss functions for pancreas segmentation. In Sect. 4 we present the studies, by grouping them into those using datasets with only pancreas annotations, and those with datasets with multi-organ annotations. We further clustered them according to the DL architectures described in the previous section. Likewise, in Sect. 5 we present the studies on DL for the segmentation of pancreas tumors. In Sect. 6 and Sect. 7 we report the studies on cysts, and inflammation, respectively. In Sect. 8 we discuss the main findings of the review along with challenges and future directions. In Sect. 9 we present the conclusions. Our major contributions are the following:

- description of the main DL architectures used for pancreas segmentation;
- systematic and extensive review on the technical advancements of DL for pancreas segmentation (parenchyma, tumors, cysts, and inflammation);
- visual presentation of all retrieved studies in tabular form in terms of application, dataset size, DL architecture, learning strategy, loss functions, training protocol, results, and main contributions;

- comparison of the performances of the DL approaches for the various applications;
- discussion of challenges and future directions from the clinical and technical perspectives.

2 Methods

2.1 Research questions

The Sample, Phenomenon of interest, Design, Evaluation, and Research type (SPIDER) tool was used to formulate the research questions and to frame a thorough analysis of the published literature (Cooke et al. 2012).

RQ1: Which datasets (publicly available and/or private) were used for pancreas segmentation based on DL?
RQ2: What are the applications of DL for pancreas segmentation?
RQ3: Which DL models were specifically designed?
RQ4: What are the performances of these models?
RQ5: What are the main contributions of the studies?
RQ6: By considering the application, datasets, and DL models how the studies can be classified effectively?
RQ7: How the reviewed studies can be presented for a thorough analysis?
RQ8: What are the challenges and future directions on DL for pancreas segmentation?

2.2 Literature search

In October 2023, a literature search was conducted on PubMed, Web of Science, and Scopus following the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement (Page et al. 2021). The search was limited to articles in the English language with an abstract and published from January 1st, 2013 to October 31st, 2023. The following search terms were used: ("artificial intelligence" OR "deep learning" OR "convolutional neural network" OR "segmentation" OR "self-supervised learning" OR "supervised learning" OR "generative artificial intelligence" OR "encoder" OR "decoder") AND ("pancreas" OR "surgical planning pancreas" OR "preoperative planning pancreas"). Reviews, letters, non-peer-reviewed articles, conference abstracts, and proceedings were excluded from the analysis.

2.3 Data extraction

Identified articles were screened by title and abstract, followed by full-text review, data extraction, and review of references. Two reviewers (AM and MC) independently screened titles and abstracts for relevance. In case of insufficient information, the corresponding authors of the articles concerned were contacted for further details. References were checked to retrieve further studies.

2.4 Data analysis

For each group, a table was prepared to visually present the data of the studies. A customized SPIDER tool was applied to the studies of each group, reporting: the dataset size (Sample), the application (Phenomenon of Interest), the model architecture, the learning strategy, loss function, the training/validation/test split, the name of GPU, and the training time (Design), the results (Evaluation), and the main contributions of the study (Research).

2.5 Results of the literature search

The database search retrieved 2,851 results. After title and abstract screening, the full texts of 206 reported studies were analyzed, but only 140 were found eligible for inclusion. Twenty studies using imaging acquisition other than CT (magnetic resonance imaging (MRI), positron emission tomography (PET), and ultrasound) were excluded. The list of excluded articles and the reasons for exclusion are reported in Sect. 2.6. Ten additional studies were retrieved after a manual check of the references. A total of 130 studies were included for full-text analysis (Fig. 2). By considering the involved countries (Fig. 3, left panel), China led the ranking with a share of 54.4%, followed by the United States (17.1%), the United Kingdom (5.3%), Canada (3.5%), and Japan (3.5%). In the majority of studies, 3D neural networks (Fig. 3, central panel) were used (51.4%), followed by 2D models (42.7%), and 2.5D (5.8%). By considering the learning type, the vast majority concern studies on supervised learning (83.8%), followed by semi-supervised learning (9.5%), and unsupervised learning (4.4%). Other types of learning (reinforcement, weakly, and continual) are reported in 2.2% of the studies (Fig. 3, right panel). Overall, there is a positive trend in the number of published articles in peer-reviewed journals, included in the present review, even though the data for the year 2023 are available until October 31st (Fig. 4). Notably, there has been a surge in the number of studies on DL for the segmentation of pancreas tumors in 2023. The mean impact factor of the journals where the included studies were published is 5.4 according to the 2023 statistics by the Journal Citation Reports ^{TM}.

2.6 Excluded studies on MRI, PET, and ultrasound

The retrieval of the full-text articles included also 13 studies on MRI (Mazor et al. 2024; Yang et al. 2022a; Ding et al. 2022; Zhang et al. 2022; Kart et al. 2021; Chen et al. 2020a; Fu et al. 2018; Li et al. 2023g; Jiang et al. 2023; Liu et al. 2023; Li et al. 2023e, 2022b; Liang et al. 2020), one on PET (Zhang et al. 2023a), and six on ultrasound (Yao et al. 2021; Fleurentin et al. 2023; Iwasa et al. 2021; Tang et al. 2023b, a; Seo et al. 2022). After analysis, they were all excluded since they did not introduce technical advancements in terms of DL architectures, design of loss functions, semi-supervised, or unsupervised learning. In contrast, one study combining CT and MRI (Li et al. 2022c), and two combining CT and PET (Sundar et al. 2022; Wang et al. 2023) were included.

2.7 Limitations of published reviews

The published reviews are reported in Table 1. The most recent one was performed by (Ghorpade et al. 2023) and published in 2023. It is a narrative review of 44 studies (32 on

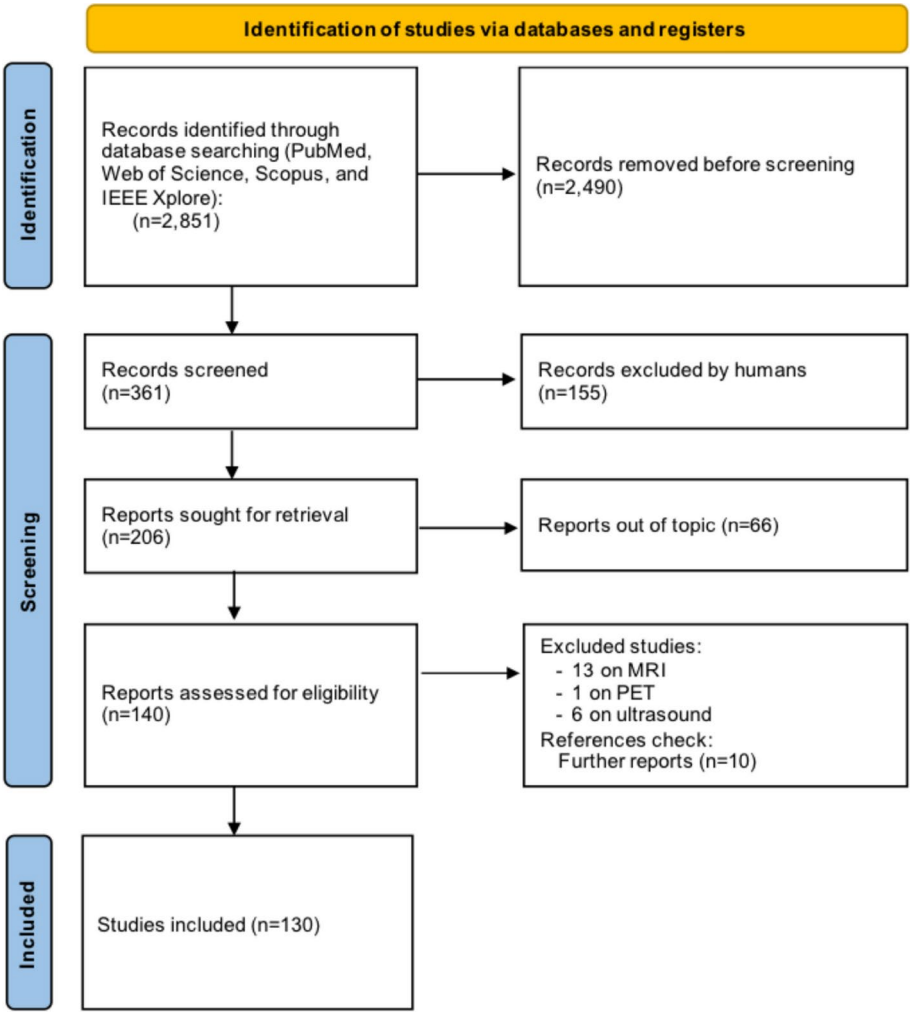

Fig. 2 Flow chart of the study selection process according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) 2020 statement (Page et al. 2021)

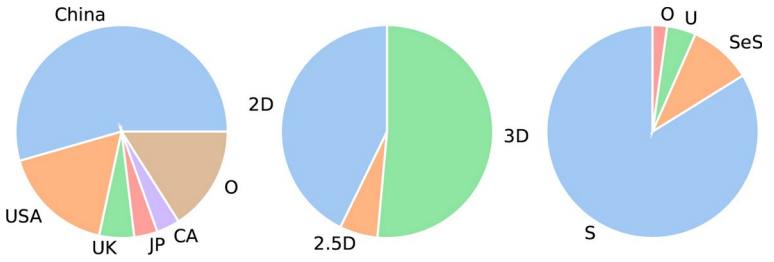

Fig. 3 Share of reviewed studies by country of affiliated institutions of authors (left), class of network (middle), and type of learning (right). *USA* United States of America, *UK* United Kingdom, *JP* Japan, *CA* Canada, *O* Other, *S* supervised, *SeS* semi-supervised, *U* unsupervised

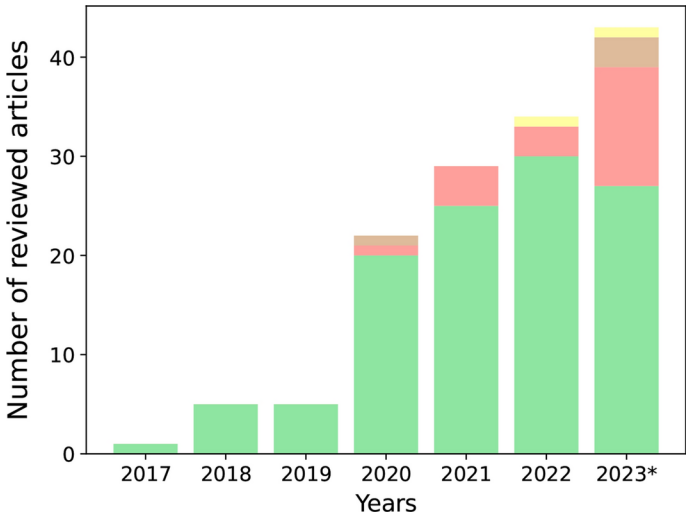

Fig. 4 Annual distribution of the 130 reviewed articles. Studies on parenchyma (in green), tumors (in pink), cysts (in brown), and inflammation (in yellow). Note: some studies concerned more than one application, e.g. parenchyma, and tumor. *Data for the year 2023 are available until October 31st. (Color figure online)

Table 1 Published reviews

Reference	Type of review	Databases	Covered years	Reviewed studies on CT	Pancreas specific
Aljabri and Al-Ghamdi (2022)	Systematic	Google Scholar	2014–2021	4 (parenchyma)	No
Ghorpade et al. (2023)	Narrative	PubMed and Web of Science	2013–2023	32 (parenchyma) 12 (tumors and cysts)	Yes
Huang et al. (2022a)	Narrative	PubMed, Embase, and Web of Science	Until 2022	7 (tumors)	Yes
Kumar et al. (2019)	Systematic	MEDLINE, Espacenet, Google Patents, and the United States Patent and Trademark Office Patent	Until 2018	16 (parenchyma)	Yes
Rehman and Khan (2020)	Narrative	–	Until 2019	8 (parenchyma)	No
Senkyire and Liu (2021)	Narrative	PubMed, Scopus, and Web of Science	Until 2020	13 (parenchyma)	No
Yao et al. (2019)	Narrative	Web of Science	2012–2018	12 (parenchyma)	Yes

parenchyma and 12 on tumors of the pancreas). The only systematic review on pancreas segmentation was performed by Kumar et al. (2019), which may be considered obsolete given the surge of published articles since 2020. It analyzed 19 studies (16 on CT and three on magnetic resonance). The review by Huang et al. (2022a) concerned artificial intelligence (AI) on pancreas cancer. Out of the included studies, only seven pertain to DL for pancreas segmentation. The review by Yao et al. (2019) discussed different approaches to

pancreas segmentation, with 12 studies on AI. The other reviews reported the published literature on DL on medical images of several anatomical structures (organs, and bones) in addition to the pancreas (Aljabri and AlGhamdi 2022; Rehman and Khan 2020; Senkyire and Liu 2021). As can be seen from Table 1 the number of the included studies on the published reviews on pancreas segmentation is considerably lower than the results of our literature search.

3 Technical background of deep learning techniques in pancreas segmentation

In this section the DL architectures specifically used for pancreas segmentation are illustrated. They are foundational to the interpretation of the results of the reviewed studies.

3.1 UNet and its variants

UNet is a U-shape fully connected network (FCN) with an encoder and decoder. The encoder extracts features through convolutions, while the decoder restores the initial resolution of the input image through deconvolutions. The key innovation of UNet is represented by skip connections between opposing convolutional and deconvolutional layers (Ronneberger et al. 2015). Skip connections successfully concatenate features learned at different levels to improve the segmentation performance, especially at the level of localization (Chen et al. 2022c). 3D UNet is the counterpart of UNet, where the 2D operations were replaced by the corresponding 3D implementation (Çiçek et al. 2016). In V-Net the forward convolutions were replaced by residual convolution units (Milletari et al. 2016). DenseVNet introduced a cascade of dense feature stacks. In dense blocks, the feature maps are concatenated enabling a streamlined gradient backpropagation (Gibson et al. 2018). A convolution is inserted into each skip connection to reduce the number of features. The maps generated in the decoding path are then concatenated and convolved. The result is added to a spatial prior, a low-resolution 3D map of trainable parameters bilinearly upsampled to the segmentation resolution, to generate the final result (Gibson et al. 2018). DRINet was developed by merging dense blocks, residual inception blocks, and unpooling blocks (Chen et al. 2018).

However, the optimal depth of an encoder-decoder in the traditional UNet architecture can vary from one application to another, depending on the task complexity. A solution would be to train models of different depths separately and then aggregate the resulting models at inference time. However, this approach is inefficient since the separate networks do not share a common encoder. Moreover, the design of skip connections requires the fusion of the same-scale encoder and decoder feature maps. UNet++ was designed to overcome these limitations. It is based on an ensemble of several UNet networks with different depths partially sharing the same encoder but retaining their specific decoder. Densely connected skip connections enable dense feature propagation along horizontal and vertical skip connections and more flexible feature fusion at the decoders (Zhou et al. 2020). The nnU-Net (Not a New UNet) framework is a cutting-edge DL framework for automating configuration across the segmentation pipeline, encompassing pre-processing, network architecture, training, and post-processing, adapting seamlessly to new datasets (Isensee et al. 2020). nnU-Net configuration method starts with the extraction of dataset properties

(image size, voxel spacing) and execution of heuristic rules. nnU-Net by default generates three different UNet configurations: a 2D UNet, a 3D UNet, and a 3D UNet cascaded in which the first UNet operates on down-sampled images, and the second is trained to refine the segmentation maps created by the former. After cross-validation nnU-Net empirically chooses the best performing configuration or ensemble. The outputs of nnU-Net are fully trained models that can be deployed to make predictions on unseen data. With open-source accessibility, nnU-Net stands as a pivotal tool, delivering state-of-the-art performance and driving advancements in automated medical image analysis (Isensee et al. 2020).

3.2 Attention and its variants

The concept of attention drew inspiration from human biological systems. For instance, the visual system focuses on some parts of an image rather than others (Chaudhari et al. 2021). Basically, attention in DL can be explained as a mechanism incorporating the concept of relevance to pay attention to only certain parts of an input (Chaudhari et al. 2021). The first use of attention in DL was presented by Bahdanau et al. (2014) for the encoder-decoder architecture for sequence-to-sequence tasks, like language translation. These models were based on recurrent neural networks for encoder and decoder, with the encoder compressing the input sequence into a single vector of fixed length at the last step of the encoding process, called hidden state. Unfortunately, in the case of long sequences, the compression may lead to loss of information (Chaudhari et al. 2021). To overcome this limitation the key idea of attention was to introduce a structure called context vector equivalent to a weighted sum of the hidden states of the decoder (one for each encoding step) and the corresponding attention weights. This enables the decoder to access the entire sequence of the encoder and focus on the relevant positions in the input sequence thanks to the attention weights (Chaudhari et al. 2021). Several types of attention were proposed for computer vision (Guo et al. 2022a). Attention gate was developed to learn to suppress irrelevant regions in an input image while highlighting salient features useful for a specific task (Oktay et al. 2018). Spatial attention can be performed by spatial transformers that are able to transform feature maps (Jaderberg et al. 2015). The spatial transformers include three submodules: a localization network with feature maps as input and the predicted transformation parameters by regression as output; a grid generator to use the regressed transformation parameters to create a sampling grid, consisting of a set of points where the input feature map should be sampled to produce the transformed output; and a sampler using the input feature map and sampling grid to create the output map (Jaderberg et al. 2015). Channel attention can be realized by squeeze and excitation (SE) block (Hu et al. 2017). SE was designed to perform feature recalibration. Essentially, SE adds a parameter to each channel of a CNN block to adjust the relevance of each feature map. In the first part, the squeeze operation performs global average pooling to reduce each feature map along width and height to a numeric value, obtaining a channel descriptor. In the excitation part, the numeric values are then fed to two fully connected layers with ReLU and sigmoid activation functions to obtain new numeric values which are used to weigh the original feature maps and assign each channel a specific relevance (Hu et al. 2017). The residual attention network is composed of a stack of several attention modules that generate attention-aware features (Wang et al. 2017). Each attention module is divided into a trunk branch and a mask branch. Each trunk branch has its mask branch to learn attention specialized for its features. The trunk branch performs feature extraction

and can be integrated into any network. The mask branch weighs output features from the truck branch (Wang et al. 2017). The attention mask serves as a feature selector during forward inference and as a gradient update filter during backpropagation. Moreover, the mask branches prevent wrong gradients from updating trunk parameters. Inside each attention module, both spatial and cross-channel dependencies are modeled (Wang et al. 2017). The convolutional block attention module (CBAM) was designed to emphasize meaningful features along channels and spatial axes in CNNs (Woo et al. 2018). The idea behind CBAM is that the channel attention module solves the problem of learning "what" since each channel of a feature map can be considered a feature detector, while the spatial attention module solves the problem of learning "where" since it is based on the inter-spatial relationship of features (Woo et al. 2018). Instead of computing the 3D attention map directly as in residual attention, CBAM decomposes the process of learning channel attention and spatial attention separately (Woo et al. 2018). In addition to global max-polling as in SE, CBAM uses also max-pooling (Woo et al. 2018). These two pooling methods are applied to an intermediate feature map. The results of both are forwarded to a shared network to produce a channel attention map. During the spatial attention process, average pooling and max-pooling are applied along the channel axis, and the results are concatenated. A convolution layer is then used to generate a spatial attention map. Channel and spatial attention can be arranged sequentially or parallelly, although the former provided better results (Woo et al. 2018).

3.3 Transformer and its variants

Although CNNs achieved impressive results, they generally suffered by modeling long-range sequences due to the locality of convolutional operations (Azad et al. 2024; Chen et al. 2022c). Transformers were introduced to address the challenge of processing long sequences. They were initially developed for natural language processing tasks. The original transformer consisted of an encoder and a decoder. The encoder converted an input sequence of tokens into a sequence of embedding vectors, called hidden state or context. The decoder used the encoder's hidden state to iteratively generate an output sequence of tokens, one token at a time (Vaswani et al. 2017). The encoder was a stack of modules each of which included multi-head self-attention (MSA), layer normalization, feedforward layers, and a second layer normalization. MSA refers to the fact that these weights are computed for all hidden states in the same sequence, e.g., all the hidden states of the encoder. Positional embedding is added to retain positional information (Vaswani et al. 2017). The decoder has several modules consisting of mask MSA and encoder-decoder attention blocks. The former ensures that the generated tokens are based on the past outputs and the current token being predicted, while the latter learns how to relate tokens from two different sequences, e.g. two different languages (Vaswani et al. 2017).

Inspired by the design of transformers for natural language processing, vision transformers (ViT) were proposed for imaging tasks (Dosovitskiy et al. 2020). In this architecture, the image is split into a sequence of flattened 2D patches which are projected to obtain the patch embeddings. Positional embeddings are added to the patch embeddings to retain positional information. The resulting sequence of embeddings is fed as input to the encoder consisting of a series of standard transformer blocks with normalization, MSA, and a second normalization. A multi-layer perceptron is then added for the classification task (Dosovitskiy et al. 2020). Since transformers lack translation equivariance and locality, they do not generalize

well when trained on insufficient amounts of data. For this reason, ViT was pre-trained on ImageNet-21k to obtain satisfying results (Dosovitskiy et al. 2020). In order to solve this issue data efficient image transformers (DeiT) were developed (Touvron et al. 2020). Another limitation of ViT is its unsuitability when the image resolution is high due to the quadratic computation complexity of MSA w.r.t image resolution (Liu et al. 2021). In fact, in standard transformers, MSA is obtained by computing globally the relationship between a token and the other tokens (Liu et al. 2021). To solve this issue Shifted Window (Swin) Transformer was proposed (Liu et al. 2021). This architecture builds hierarchical feature maps by starting from small-sized patches and gradually merging neighboring patches in deeper layers. The linear computational complexity is ensured by computing self-attention locally within non-overlapping windows that partition an image (Liu et al. 2021). Additionally, the window in a layer is shifted w.r.t. the previous layer, causing the self-attention computation in the new window to cross the boundaries of the previous window, thus providing connections among them (Liu et al. 2021).

In computer vision, transformers can be divided into pure and hybrid ones. In pure transformers, the MSA modules are used in both the encoder and decoder. Hybrid transformer architectures fuse the ViTs with convolution modules in the encoder, bottleneck, decoder, or skip connections to combine information about the global context and local details (Azad et al. 2024). Swin-UNet is a pure transformer with a UNet-like architecture (Fig. 5) employing the Swin transformer block in the encoder, bottleneck, and decoder (Cao et al. 2023a). CTUNet is a hybrid network (Fig. 6) for segmentation of the pancreas parenchyma with 3D channel transformer blocks inserted into the skip connection of a 3D UNet (Chen and Wan 2022). A pancreas attention module with a project and excite block was designed and added to each encoder to enhance the ability to extract context information, while cross attention was inserted between the output of each transformer and decoder to eliminate semantic inconsistency (Chen and Wan 2022).

Residual transformer UNet (RTUNet) is a UNet-like network for pancreas parenchyma segmentation with convolutional blocks consisting of residual blocks, residual transformers, and dual convolution down-sampling. The residual transformer block adds progressive up-sampling to the basic transformer (Qiu et al. 2023). UMRFormer-Net is a U-shaped encoder-decoder architecture (Fig. 7) with a hybrid CNN and transformer for segmentation of the pancreatic parenchyma and tumors (Fang et al. 2023). It has five 3D CNN layers and a double transformer module inserted into the bottleneck and skip connection of the fourth layer to encode the long-range dependencies semantic information in a global space (Fang et al. 2023).

Convolutional pyramid vision is a hybrid network of CNN and hierarchical transformers for tumor segmentation. It generates multi-scale features by incorporating multi-kernel convolutional patch embedding and local spatial reduction to reduce computational cost. In this way, the model is able to capture the local information of multi-scale tumors (Viriyasaranon et al. 2023).

3.4 Generative adversarial network and its variants

Generative Adversarial Networks (GANs) are generative models with a generator and a discriminator network which are trained to compete and overcome each other. In GANs there is a minimax two-player game, where the generator network tries to fool a discriminator

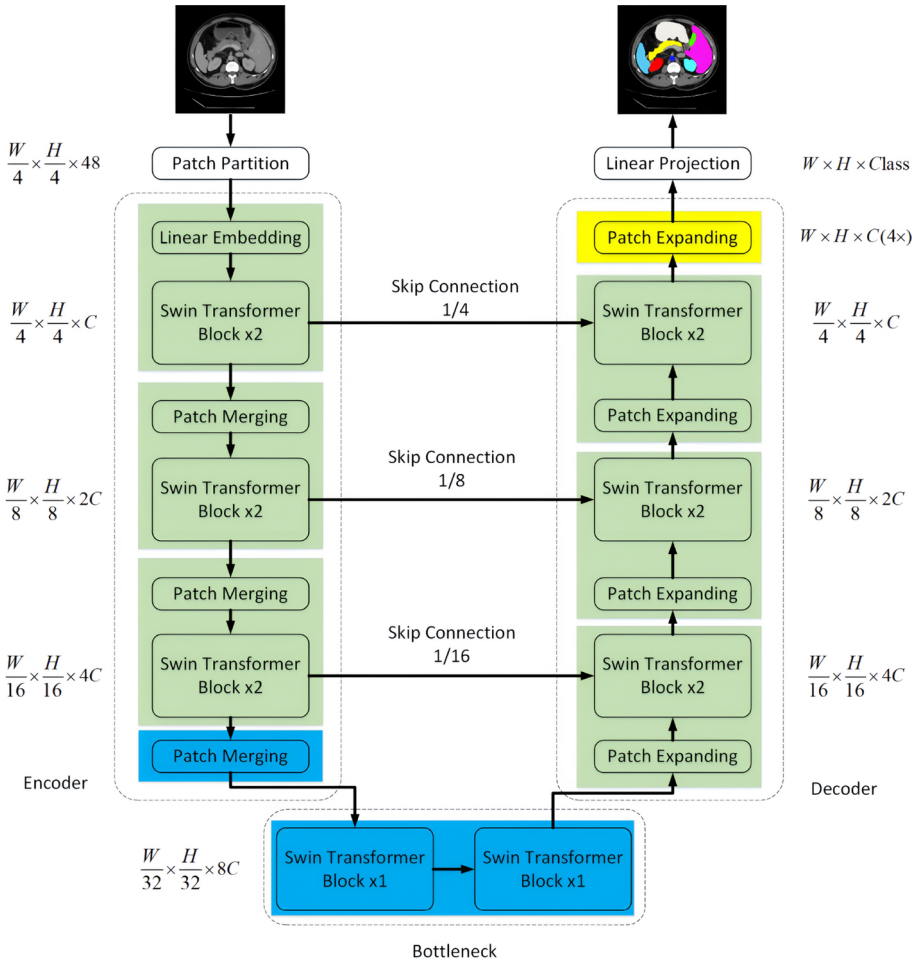

Fig. 5 Architecture of Swin-UNet from Cao et al. (2023a)

which has to distinguish between real images (coming from the training dataset) and false ones (generated by the discriminator starting from a random noise distribution (Goodfellow et al. 2014)). CycleGAN networks were proposed for the image-to-image translation task, converting an image from one domain to another one (Zhu et al. 2017). In contrast with previous approaches for image translation in computer vision with pair data between the two domains, in CycleGANs the the images are not paired (Zhu et al. 2017).

3.5 Dilated convolutions

The max-pooling and strides (downsampling) on CNNs layers result in feature maps with considerably reduced spatial resolution (Chen et al. 2016). Inspired by the efficient computation of the undecimated wavelet transform, known as "algorithm a trous", Chen et al. (2016) proposed atrous convolution, replacing the downsampling in the last max pooling

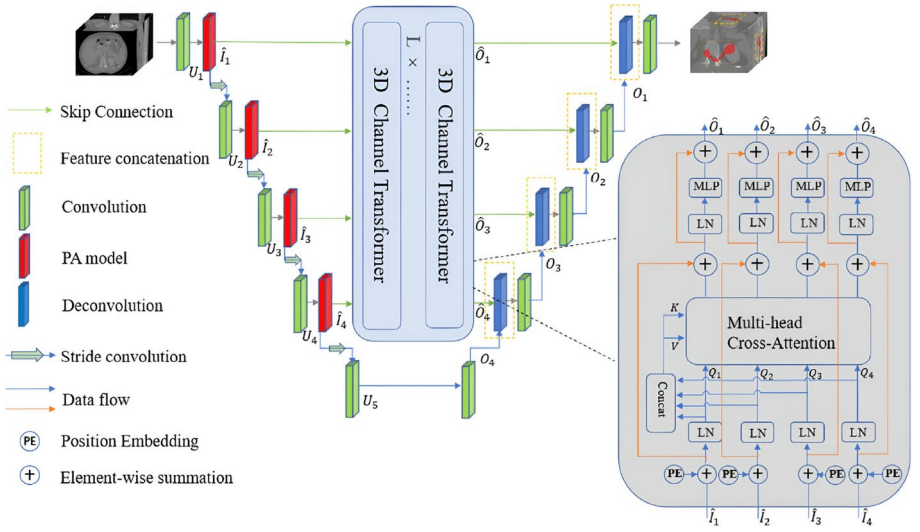

Fig. 6 Architecture of CTUNet from Chen and Wan (2022)

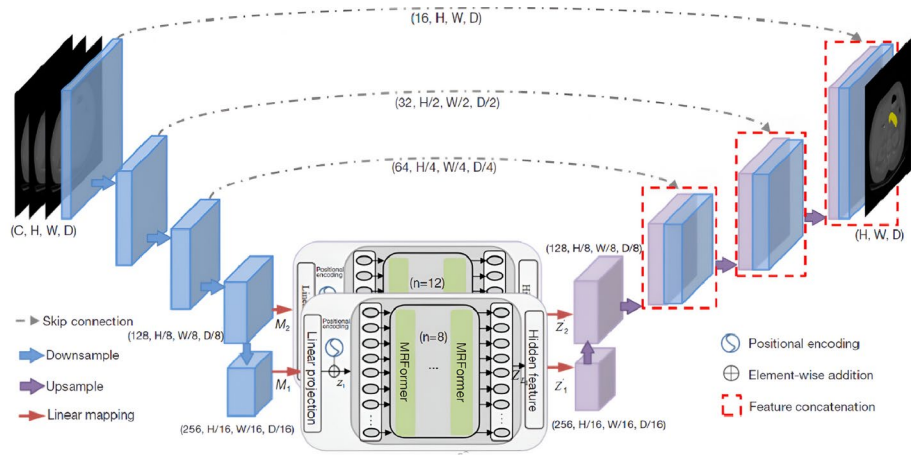

Fig. 7 Architecture of UMRFormer-Net from Fang et al. (2023)

layers of CNNs with upsampling the filters by inserting holes ("trous" in French) between nonzero filter values. As a result, the feature maps are computed at a higher sampling rate than in conventional CNNs. Atrous convolutions enable the enlargement of the field of view of filters without increasing the number of parameters or computational burden. Atrous convolution was later called "dilated convolution". By adopting multiple parallel atrous convolutional layers with different sampling rates it is possible to capture objects at different scales, in a way similar to spatial pyramid pooling. For this reason, this technique was named Atrous Spatial Pyramid Pooling (ASPP) (Chen et al. 2016).

3.6 Datasets

Ten open datasets for pancreas segmentation are available online (Table 2). Seven were largely adopted in the reviewed studies. The Cancer Image Archive (TCIA) from the National Institute of Health (NIH) is an online service[1] hosting medical imaging archives. The most investigated dataset for pancreas segmentation comes from this source and consists of 82 CTs. It is known as the NIH dataset. There are also published studies using 43 CTs from TCIA-NIH. From here onward it will be referenced as TCIA dataset. The NIH dataset includes only labeled images of the pancreas parenchyma, while the Medical Segmentation Decathlon (MSD) dataset also annotations of tumors (Fig. 8). The others incorporate the segmentation of multiple abdominal organs, namely 15 (AMOS-CT), 13 (Beyond the Cranial Vault (BTCV)), four (AbdomenCT-1k), and four for the Fast and Low GPU memory Abdominal oRgan sEgmentation (FLARE). Only AMOS-CT, AbdomenCT-1k, and FLARE are multi-vendor and multicenter, with data from two, 12, and 11 centers, respectively (Ji et al. (2022); Ma et al. (2022b)). In the NIH dataset, the pancreas was manually labeled by a medical student and then verified by an experienced radiologist (Roth et al. (2015); Ma et al. (2022b)). The images of the MSD dataset were provided by the Memorial Sloan Kettering Cancer Center (New York, NY, United States). The pancreatic parenchyma and pancreatic mass (cyst or tumor) were manually annotated by an expert radiologist (Simpson et al. (2019)). In the AMOS-CT dataset, 50 out of 500 CTs were initially annotated by humans. Then, one 3D UNet was trained using these 50 CTs to pre-label the remaining ones (coarse stage). Five junior radiologists refined the segmentation results. To further reduce errors, three senior radiologists with more than 10 years of experience checked and validated the results (fine stage). The process was iterated several times to reach a final consensus on the well-labeled annotations (Ji et al. (2022). For the AbdomenCT-1k dataset, 15 junior annotators (one to five years of experience) used the ITK-SNAP tool to manually segment the organs under the supervision of two board-certified radiologists. Then, one senior radiologist with more than 10 years of experience checked the annotations. After annotation, UNet models were trained to find the possible errors, which were double-checked by the senior radiologist (Ma et al. (2022b)). The dataset grouped the MSD Pancreas (420 cases), the NIH (80 cases), tumors of the liver (201 cases), tumors of the kidneys (300 cases), spleen (61 cases), and 50 CT scans from Nanjing University of patients with pancreas cancer (20 cases), colon cancer (20 cases), and liver cancer (10 cases) for a total of 1,112 CTs (Ma et al. (2022b)). The BTCV is a medical dataset for the MICCAI 2015 Multi-Atlas Abdomen Labelling Challenge. It consists of 50 CTs, manually labeled by two experienced undergraduate students, and verified by a radiologist. The annotations are multi-organ. The Synapse dataset includes 30 CT scans of BTCV (Landman et al. 2015).

A dataset of 90 CTs with annotations of eight abdominal organs was curated by combining 43 from TCIA and 47 from BTCV (Gibson et al. 2018). The FLARE dataset was curated for a medical imaging challenge. It consists of 511 CTs with annotations of the liver, kidneys, spleen, and pancreas (Ma et al. 2022a). Some reviewed studies used also the International Symposium on Image Computing and Digital Medicine (ISICDM) dataset, initially curated for the 2018 Pancreatic Segmentation Challenge. Concerning pancreas cancer, there are two open datasets with annotations of tumors. The first is called The Clinical Proteomic Tumor Analysis Consortium Pancreatic Ductal Adenocarcinoma Collection and contains

[1] https://www.cancerimagingarchive.net/

Table 2 Publicly available datasets of abdominal CT scans for the segmentation of the pancreas parenchyma and cancers

	Name	Country	Size	Phase	Application	Annotation
Roth et al. (2015)	National Institute of Health (NIH)	United States	82	Arterial	Parenchyma	Manual
Simpson et al. (2019)	Medical Segmentation Decathlon (MSD)	United States	420	Venous	Parenchyma tumors	Manual
https://www.synapse.org/ Synapse:syn3193805/ wiki/217789	Beyond the cranial vault	United States	50 from Vanderbilt University Medical Center)	Venous	Parenchyma (13 abdominal organs)	Manual
Gibson et al. (2018)	Multi-organ abdominal CT reference standard segmentation	United States	90 (43 from NIH 47 from beyond the cranial vault)	Arterial	Parenchyma tumors (8 abdominal organs)	Manual
Ji et al. (2022)	AMOS-CT	China	500	Venous arterial	Tumors (15 abdominal organs)	Semiautomatic
Ma et al. (2022a)	FLARE	China and other countries	511 based on MSD (liver, spleen, pancreas), NIH pancreas, 50 from Nanjing University)	Venous arterial	Parenchyma (4 abdominal organs)	Manual
Ma et al. (2022b)	AbdomenCT-1k	China	1,112 (MSD,NIH, 50 from Nanjing University)	Venous arterial	Parenchyma tumors (4 abdominal organs)	Manual
https://www.cancerimagingarchive.net/ collection/cptac-pda/	Clinical proteomic tumor analysis consortium pancreatic ductal adenocarcinoma	United States	71	–	Tumors	Manual
Chen et al. (2023)	CTpred-Sunitinib-panNET	United States	38	–	Tumors	Manual
Luo et al. (2022)	WORD	China	150	Venous arterial	Radiotherapy (16 abdominal organs)	Manual

data from CT, MRI, and US with 71 CTs on pancreatic ductal adenocarcinomas (PDACs), used in one study. The other is named CTpred-Sunitinib-panNET and includes) and includes 38 CTs on pancreatic neuroendocrine tumors (PNETs). The last dataset is WORD with 150 labeled CTs of 16 organs.

3.7 Metrics

This section presents a thorough mathematical formulation of the six distinct metrics identified in the systematic review for assessing model performance. These metrics are Dice Score Coefficient (DSC), Jaccard Index (JI), Hausdorff Distance (HD), 95th percentile Hausdorff Distance (HD95), Average Surface Distance (ASD), and Normalized Surface Dice (NSD). To formally define the metrics, let us consider that the medical images are represented by a collection of points $X = \{x_1, x_2, ..., x_n\}$, where each x_i corresponds to a voxel value within the image. The entire set X is organized within a three-dimensional grid, such that the total number of points (voxels) is given by $|X| = N$, where $N = w \times h \times d$. Here, w denotes the width, h the height, and d the depth of the grid, respectively. For each voxel $x \in X$, there are corresponding labels in the ground truth segmentation S_g and in the automatic segmentation predicted by the model S_p. We define the labeling function for the ground truth segmentation as $S_g : X \to \{0, 1\}$, where $S_g(x)$ denotes the label assigned to voxel x by S_g. Similarly, the labeling function for the predicted segmentation is defined as $S_p : X \to \{0, 1\}$, where $S_p(x)$ represents the label assigned to voxel x by S_p.

3.7.1 Region-based metrics

Building on this premise, this section first defines two metrics classified as overlap-based, namely DSC and JI. The first step is defining the four cardinalities that underlie these metrics, as delineated below:

$$TP = |\{x \in X : S_g(x) = 1 \text{ and } S_t(x) = 1\}| \tag{1}$$

$$FP = |\{x \in X : S_g(x) = 1 \text{ and } S_t(x) = 0\}| \tag{2}$$

$$FN = |\{x \in X : S_g(x) = 0 \text{ and } S_t(x) = 1\}| \tag{3}$$

$$TN = |\{x \in X : S_g(x) = 0 \text{ and } S_t(x) = 0\}| \tag{4}$$

where TP stands for true positive, FP for false positive, FN for false negative, and TN for true negative. The symbol $|\cdot|$ denotes the count of the set. The DSC, often called Dice or overlap index, is the predominant metric for validating medical volume segmentations. Beyond facilitating direct comparisons between automated and ground truth segmentations, the Dice metric is frequently employed to assess reproducibility and repeatability within these analyses (Kamnitsas et al. (2017); Ronneberger et al. (2015); Li et al. (2019)). A score of 0 indicates no overlap, while a score of 1 indicates perfect overlap, and its formulation is defined by:

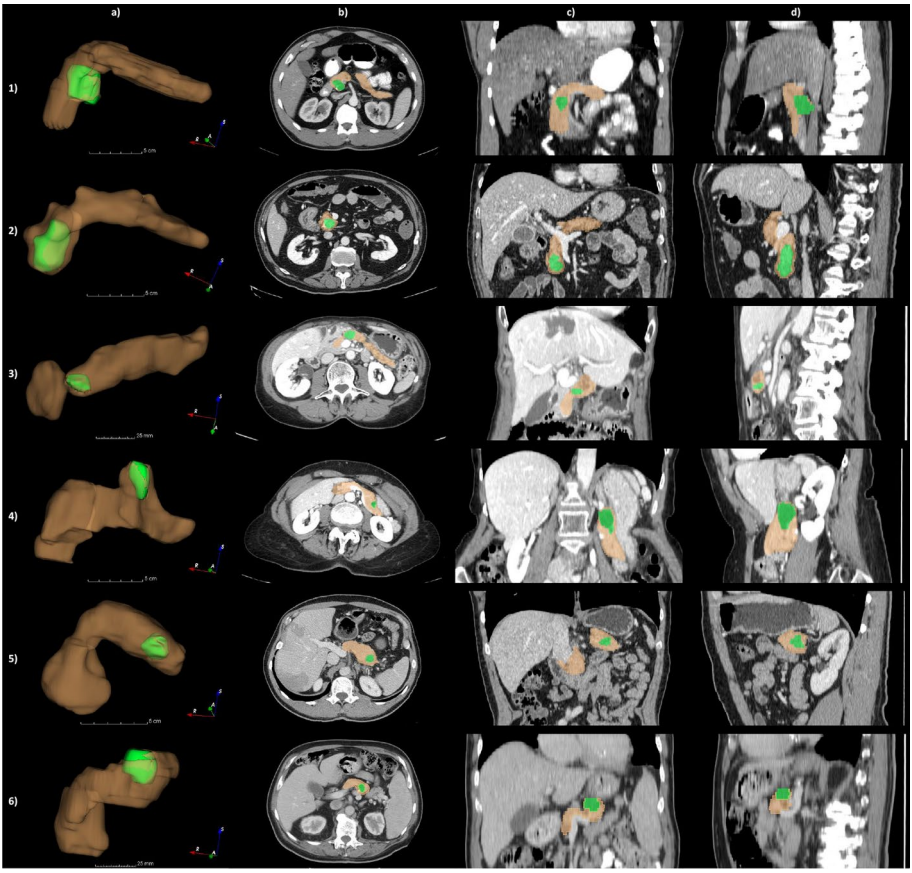

Fig. 8 Six cases of pancreas anatomy, along with a tumor, from the MSD dataset (rows 1–6) to show the large morphological variability (Simpson et al. (2019)). Column a: 3D model of parenchyma (in brown) and tumor (in green). Columns b,c,d: view on the axial, coronal, and sagittal plane. Case number of MSD dataset (from row 1 to row 6): #66, #64, #334, #126, #286, and #81. Pancreas subregions are grouped as follows: head (row 1 and row 2), body (row 3 and row 4), and tail (row 5 and row 6)

$$DSC = \frac{2|S_p \cap S_g|}{|S_p| + |S_g|} = \frac{2TP}{2TP + FN + FP} \tag{5}$$

Instead, the JI is determined by dividing the intersection of two sets by their union Jaccard (1912). This metric quantifies the similarity between the sets, represented mathematically as:

$$JI = \frac{|S_p \cap S_g|}{|S_p \cup S_g|} = \frac{TP}{TP + FN + FP} \tag{6}$$

DSC and JI range between 0 and 1, where 1 means perfect overlap and 0 means null intersection between S_p and S_g.

3.7.2 Distance-based metrics

The second part of this section defines the set of spatial distance-based metrics: HD, HD95, ASD, and NSD. These metrics represent a pivotal dissimilarity measure in evaluating image segmentation, especially when the task requires a proper edge delineation. HD was specifically designed to assess the shape similarity between two point sets within a given metric space Huttenlocher et al. (1993). HD's evaluation is independent of point correlations, focusing only on the pairwise distances between voxels. Nevertheless, it shows a significant vulnerability to outliers in the data set. It is defined as:

$$HD(S_g, S_p) = \max(h(S_g, S_p), h(S_p, S_g)) \tag{7}$$

where $h(S_g, S_p)$ is called the directed Hausdorff distance and is given by:

$$h(S_g, S_p) = \max_{x_g \in S_g} \min_{x_p \in S_p} \|x_g - x_p\| \tag{8}$$

where $\|x_g - x_p\|$ represents a norm such as euclidean distance. Nonetheless, the HD95 introduced by Huttenlocher et al. (1993) is the quantile approach to HD providing a method to reduce the influence of outliers by considering the q^{th} quantile of direct Hausdorrf distances instead of the maximum distance. The choice of q^{th} depends on the specific application and the characteristics of the point sets under analysis. Our systematic review focuses on the 95^{th} percentile HD95, widely used in literature. This metric is similar to the traditional HD but is defined as follows:

$$HD95(S_g, S_p) = \max(h_{95}(S_g, S_p), h_{95}(S_p, S_g)) \tag{9}$$

where h_{95} represents the 95^{th} ranked percentile of the set of minimum distances between points from one set to the nearest points in the other. Specifically, $h_{95}(S_g, S_p)$ is defined as:

$$h_{95}(S_g, S_p) = \underset{x_g \in Sg}{rank}^{95} \min_{x_p \in S_p} \|x_g - x_p\| \tag{10}$$

where $\|x_g - x_p\|$ denotes a norm such as euclidean distance. Another metric belonging to the distance-based class is the average ASD (ASSD). It is defined as the average of all the distances from points on the boundary of the ground truth segmentation to the boundary of the predicted segmentation, and vice-versa (Yeghiazaryan and Voiculescu (2018)). The ASSD is defined by:

$$ASD(S_g, S_p) = \frac{1}{|S_g| + |S_p|} \left(\sum_{x_{sg} \in S(S_g)} d(x_{sg}, S(S_p)) \quad + \sum_{x_{sp} \in S(S_p)} d(x_s p, S(S_g)) \right) \tag{11}$$

where $d(x_{sg}, S(S_p))$ is defined as

$$d(x_{sg}, S(S_p)) = \min_{s_{sp} \in S(S_p)} \|s_{sg} - s_{sp}\| \tag{12}$$

with $S(S_g)$ and $S(S_p)$ represent the surfaces (boundary) of S_g and S_p respectively. HD, HD95, and ASD are initially expressed in units of voxels and then converted into millimeters (mm) based on the voxel spacing of the medical images. Lastly, the NSD, introduced by Nikolov et al. (2021), quantifies the accuracy of segmentation boundaries by measuring the proportion that meets a specified deviation threshold, τ. This threshold represents the maximum clinically acceptable error in pixels, offering a precise metric for evaluating how closely a predicted segmentation aligns with the actual boundary within a tolerable margin of error. The NSD is defined as

$$NSD = \frac{|D_g| + |D_p|}{|D'_g| + |D'_p|} \tag{13}$$

where D_g and D_p are the nearest neighbour distances computed respectively from the surface $S(S_p)$ to the surface $S(S_g)$ and vice-versa, while D'_g and D'_p are respectively the subset of distances in D_g and D_p that are smaller or equal to acceptable deviation τ as defined by:

$$D'_g = \{d_g \in D_g \mid d_g \leq \tau\} \tag{14}$$

$$D'_p = \{d_p \in D_p \mid d_p \leq \tau\} \tag{15}$$

The NSD ranges between 0 and 1 Seidlitz et al. (2022). A score of 0 signifies either complete inaccuracy, with all measured distances exceeding the predefined acceptable deviation threshold τ, or the image's absence of the predicted class. Conversely, a score of 1 means no corrections to the segmentation boundary are needed, as all deviations from the reference boundary fall within the acceptable threshold τ.

3.8 Loss functions

This section presents a thorough mathematical formulation of the three most commonly used loss functions identified in the systematic review. Following the conventions outlined in Sect. 3.7, the mathematical formulations of Binary Cross Entropy loss (L_{BCE}), Focal loss (L_{Focal}), and Dice loss (L_{Dice}) will be presented below. Binary Cross Entropy loss function belongs to the class of distribution-based losses, designed with the purpose of minimizing discrepancies between two probability distributions. The formulation of Binary Cross Entropy loss is given by:

$$L_{BCE} = -\frac{1}{N} \sum_{x \in X} [S_g(x) \log(S_p(x)) + (1 - S_g(x)) \log(1 - S_p(x))] \tag{16}$$

Focal loss function also belongs to the class of distribution-based losses. This loss modifies the conventional cross entropy by emphasizing misclassified pixels or voxels. It reduces the significance of the loss in well-classified samples, allowing it to effectively address imbalances between foreground and background classes. The formula below is an adaptation of the multiclass Focal loss of Lin et al. (2017) for binary classification, defined as:

$$L_{Focal} = -\frac{1}{N}\sum_{x\in X}\Big[(1-S_p(x))^{\gamma}S_g(x)\log(S_p(x)) + (1-(1-S_p(x)))^{\gamma}(1-S_g(x))\log(1-S_p(x))\Big] \tag{17}$$

Dice loss function belongs to the class of overlap-based losses. This function aims to quantify the degree of overlap between the ground truth segmentation S_g and the predicted segmentation S_p Isensee et al. (2019). It directly optimizes the DSC defined in section 3.7, and its formula is given by:

$$L_{Dice} = 1 - \frac{\sum_{x\in X} S_g(x)S_p(x)}{\sum_{x\in X} S_g(x) + \sum_{x\in X} S_p(x)} \tag{18}$$

4 Segmentation of the parenchyma

This section starts by showing the variability of the pancreas parenchyma in terms of size and location (Sect. 4.1). Then, the different approaches to the segmentation of pancreas parenchyma are analyzed. Overall, a total of 104 out of the 130 reviewed studies fall under this topic, with 81 using only pancreas annotations (Sect. 4.2), while the remaining 23 used multi-organ annotations (Sect. 4.3). Both sections present a comparison of the performances of the different DL models on the publicly available datasets, described in Sect. 3.6, and on the private/internal ones.

4.1 Variability of parenchyma size and location

In order to provide an example of the variability of the pancreas parenchyma in terms of size and location, a registration was performed on 281 CTs of the MSD dataset using Elastix software (Klein et al. 2010), adapting inter-subject registration parameters from the study by Qiao et al. (2016) to the CT domain. Subject #29 of MSD was considered a reference image by virtue of its high-quality image and centrality within the range of variations observed in the dataset. A Hounsfield unit (HU) from 100 to 500 was used for all the images to improve the registration process, enhancing bones and brighter abdominal structures. The results are illustrated in Fig. 9. A histogram with the frequency distribution is shown in Fig. 10. It was created by measuring the volumetric distances from the centroid of the pancreas in subject #29 of MSD to the centroids of the pancreas from all other subjects after performing the registration.

4.2 Studies on datasets with only pancreas annotations

The 81 studies were clustered according to the network architecture: CNNs (n=6, Table 3), UNet and variants (n=48, Table 4), attention applied to CNNs and UNet (n=15, Table 5), transformers and hybrid transformers (n=4, Table 6), and GAN (n=8, Table 7). The NIH dataset was by far the most used dataset, recurring in 74 out of 81 studies (91.3%), as follows: in 51 studies it was the only one adopted, while in 24 it was coupled with others (in

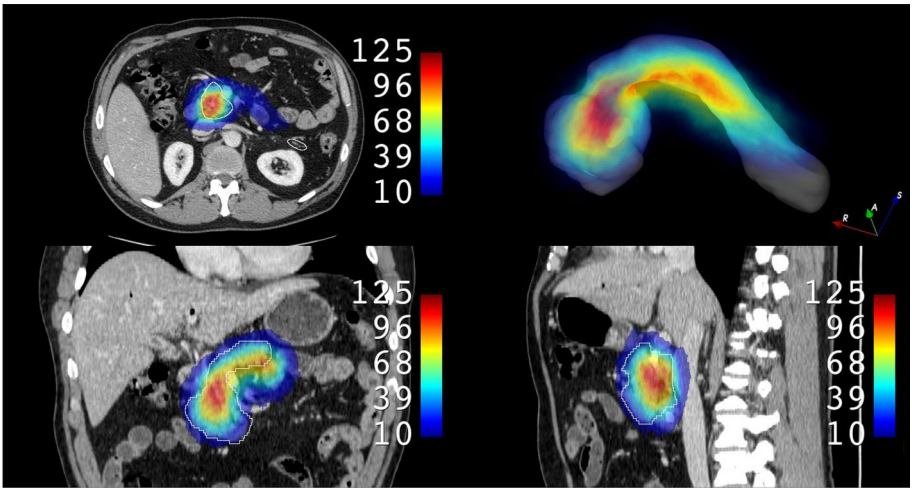

Fig. 9 Spatial distribution and frequency of pancreas within the MSD dataset (Simpson et al. (2019)) (281 cases with case #29 as a reference in the image): most frequent pancreases in the dataset in red, least frequent ones in blue. Boundary of case #29 in white

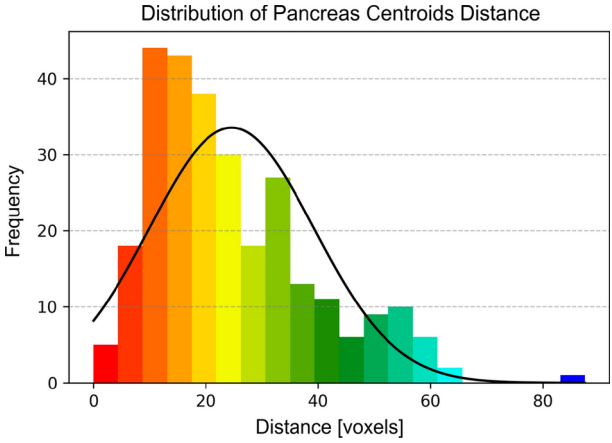

Fig. 10 Frequency distribution of centroids distance of the pancreas in the MSD dataset with 281 cases. Case #29 was used as a reference to compute the distance in voxels

14 cases with MSD dataset). The MSD was used in 18 studies (22.2%). Although one study employed the AbdomenCT-1k dataset, which includes multi-organ annotations, the authors kept only those of the pancreas after stripping out those of the other organs (Tian et al. 2023). For this reason, it was included in this section. Pancreas segmentation was based on a two-stage (coarse-fine) approach in 28 works, described in Section 1.1 of the Appendix. One study designed a three-stage method (Zhang et al. 2021d). In addition to localization and segmentation phases, a 3D level-set was applied for better boundary delineation of the pancreas (Zhang et al. 2021d). Six studies designed specific loss functions, described in Section 1.4 of the Appendix. The vast majority of the studies $(n = 70)$ used supervised learning. Seven works were based on semi-supervised, while four were based on self-supervised

Table 3 Reviewed studies on the segmentation of pancreas parenchyma with CNNs architectures

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Pradip and Kakarwal (2023)	Segmentation of pancreas	80 (NIH)	CNN	Supervised	Cross entropy Dice loss	–	–	–	88.68% (DSC) 98.82% (Jaccard) 68.22% (Precision) 98.66% (Recall)	Design of a multi-scale features and reorganization module. Data augmentation to reduce class imbalance
Hu et al. (2021b)	Segmentation of pancreas	82 (NIH) 70 (CT-Zheyi dataset)	DenseNet161 for dense atrous spatial pyramid pooling (Localization) DenseNet161 for distance-based saliency (Segmentation)	Supervised	Binary cross entropy	4-fold cross-validation	Nvidia GTX 2080 Ti	13.5 h	NIH: 85.49% (DSC) CT-Zheyi: 85.48% (DSC)	Dense atrous spatial pyramid Pooling to cover larger receptive fields. Saliency map is computed through geodesic distance based saliency transformation. Both localization and saliency information are used to aid segmentation
Bagheri et al. (2020)	Segmentation of pancreas	82 (NIH)	Superpixels and random forest classifier (Localization) Holistcally nested neural networks (Segmentation)	Supervised	–	4-fold cross-validation	–	–	78.00% (DSC)	Superpixels to get bounding boxes. Fusing holistically nested networks to generate interior and boundary
Mo et al. (2020)	Segmentation of pancreas	82 (NIH)	VGGNet with extraction of hierarchical features at different levels	Supervised	Dice loss	–	Nvidia GTX Titan	–	82.47% (DSC)	3D residual network to extract and aggregate hierarchical features at different levels. Concatenation of the result with features at each level to choose more discriminative features. This process is iterated.

Table 3 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Roth et al. (2018a)	Segmentation of pancreas	82 (NIH)	Holistically-nested networks for: Localization (fusing the three orthogonal axes) + Segmentation (boundaries and interior cues to produce superpixels aggregated by random forests)	Supervised	Cross entropy loss	4-fold cross-validation	Nvidia Titan X	9–12 hours	81.27% (DSC) 68.87% (Jaccard) 17.71 mm (HD) 0.42 mm (Average distance)	Segmentation incorporates deeply learned organ interior and boundary mid-level cues with subsequent spatial aggregation
Farag et al. (2017)	Segmentation of pancreas	80 (NIH)	Oversegmentation for superpixels (Middle-level representations) + Random forest classifier + AlexNet	Supervised	–	6-fold cross-validation	–	9 h	70.70% (DSC) 57.90% (Jaccard) 71.60% (Precision) 74.40% (Recall)	Bottom-up approach for image segmentation, consisting of: superpixels (from oversegmentation of slices), patch labeling by random forests or deep learning, and cascaded random forests classifiers based on previous patch labeling

Table 4 Reviewed studies on the segmentation of pancreas parenchyma with UNet architecture

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Jain et al. (2023)	Segmentation of pancreas	82 (NIH)	K-mean and Gaussian mixture model (Unsupervised) (Localization) UNet, holistically-nested edge detection, and Dense-Res-InceptionNet (Segmentation)	Unsupervised + supervised	Dice loss	4-fold cross-validation	Nvidia GeForce RTX 2060	–	81.75% (DSC) 83.03% (Precision) 81.70% (Recall)	Unsupervised localization of pancreas after segmenting liver and spleen using K-means and Gaussian mixture models
Li et al. (2023a)	Segmentation of pancreas	82 (NIH) 281 (MSD) 104 (Private)	UNet with: Meta-learning (Localization) Latent-space feature flow generation (Segmentation)	Supervised	Design of adaptive loss with: recall loss, cross entropy and dice loss	4-fold cross-validation	Nvidia GeForce GTX 3090	–	NIH (trained on MSD and private): 80.24% (DSC) 1.92mm (ASD) MSD (trained on NIH and private): 81.09% (DSC) 1.99 mm (ASD) Private (trained on NIH and MSD): 84.77% (DSC) 1.28 mm (ASD)	First generalization model for pancreas segmentation. Model-agnostic meta-learning to improve generalization of the coarse stage. Appearance-style feature flow generation in the fine stage to generate a sequence of intermediate representations between different latent spaces for simulating large variations of appearance-style features

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Li et al. (2023c)	Segmentation of pancreas	82 (NIH) 281 (MSD) 104 (Renji Hospital Shanghai Private dataset)	UNet with meta-learning (Localization) 3D UNet: Global feature contrastive learning 3D UNet: Local image restoration (Segmentation)	Self- supervised	Binary cross entropy loss Dice loss Squared error loss Adversarial loss	4-fold cross-validation	Nvidia GeForce GTX 3090	–	Training on NIH generalization on MSD: 66.73% (DSC) generalization on private: 73.85% (DSC) training on MSD generalization on NIH: 76.71% (DSC) Generalization on private: 83.50% (DSC) training on private generalization on NIH: 65.03% (DSC) generalization on MSD: 70.08% (DSC)	Dual self-supervised generalization model to enhance characterization of high-uncertain regions. Global-feature self-supervised contrastive learning reducing the influence of extra-pancreatic tissues. Local image restoration self-supervised module to exploit anatomical context to enhance characterization of high-uncertain regions

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Tian et al. (2023)	Segmentation of pancreas	200 (from AbdomenCT-1k) 82 (NIH) 281 (MSD) 50 (Jiangsu Province Hospital) (Generalization)	nnU-Net (Localization) + Hybrid variational model to capture weak boundaries (Segmentation)	Supervised	Cross entropy dice loss	5-fold cross-validation	Nvidia Titan V100	–	AbdomenCT-1k: 89.61% (DSC) NIH: 87.67% (DSC) MSD: 87.13% (DSC) Generalization: 90.72% (DSC)	First stage: 3D CNN for coarse segmentation Second stage: a new hybrid variational model to capture the pancreas weak boundary
Zheng and Luo (2023)	Segmentation of pancreas	80 (NIH)	UNet-like for both Localization and Segmentation	Supervised	Weighted binary cross entropy loss	4-fold cross-validation	Nvidia GeForce GTX 3090	–	85.58% (DSC) 74.99% (Jaccard) 86.59% (Precision) 85.11% (Recall)	Extension-contraction transformation network with a shared encoder for feature extraction and two decoders for the prediction of the segmentation masks and the inter-slice extension and contraction transformation masks

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Zhu et al. (2023)	Segmentation of pancreas	82 (NIH) 70 (Zheyi): Zhejiang University Hospital	Adversarial network + 3D ResUNet + Attention (Squeeze-Excitation)	Supervised + Unsupervised	Dice loss cross entropy loss	5-fold cross-validation	Nvidia Titan V		NIH (supervised): 85.45% (DSC) Zheyi (unsupervised): 75.43% (DSC)	Training with 3D ResUNet and attention module using pairs of labeled images fro one center and unlabeled ones from a different center to generate multi-scale feature maps. Labeled and unlabeled data are then trained by a discriminator for domain identification
Huang et al. (2022c)	Segmentation of pancreas, liver, and brain	82 (NIH) datasets of liver and brain	ResNet-34 and 3D-ResUNet	Supervised	Binary cross entropy Auxiliary loss to upsample feature maps to the same spatial resolution as the inputs	4-fold cross-validation	–	–	2D: 83.67% (DSC) 85.60% (Precision) 82.58% (Recall) 3D: 86.32% (DSC) 85.52% Precision) 84.51% (Recall)	Two sample balancing methods were proposed: positive–negative subset selection and hard-easy subset selection

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Huang and Wu (2022)	Segmentation of pancreas	82 (NIH)	MobileNet-UNet: UNet + MobileNet-V2	Supervised	Weighted cross entropy	4-fold cross-validation	–	–	82.87% (DSC) 70.97% (Jaccard) 90.54% (AUC) 89.29% (Precision) 77.37% (Recall) 99.95% (Specificity)	Implementation of MobileNet-UNet. Compared to original UNet, it adds zero padding, depthwise convolution, batch normalization, and ReLU
Khasawneh et al. (2022)	Segmentation of pancreas	294 (from 1,917 of Mayo Clinic)	UNet-like for: Localization and Segmentation	Supervised	–	–	–	–	88.00% (DSC) 79.00% (Jaccard)	Comparison of manual segmentation by experts using 3D Slicer and automatic segmentation by CNN
Lim et al. (2022)	Segmentation of pancreas	1,006 (Gil Medical Center Gachon University College of Medicine) 82 (NIH) (generalization)	UNet-like configurations	Supervised	Dice loss	4-fold cross-validation	Nvidia Tesla V100		Internal: 86.9% (Precision) 84.2% (Recall) 84.2% (DSC) Generalization: 77.90% (Precision) 74.90% (Recall) 73.50% (DSC)	Comparison of four 3D architectures based on UNet on a large dataset (1,006 CT)

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Liu et al. (2022a)	Segmentation of pancreas	82 (NIH)	Graph-enhanced nnU-Net	Semi- supervised	Cross entropy Dice loss	4-fold cross-validation	Nvidia Titan XP	–	84.22% (DSC) 73.10% (Jaccard) 6.63 voxel(HD95) 1.86 voxel (ASD	A graph CNN was added to nnU-Net to distinguish the low contrast edges of a pancreas. Pseudo labels are refined using an uncertainty iterative strategy
Liu et al. (2022d)	Segmentation of pancreas	82 (NIH)	Region of interest based on surrounding organs + VGG-UNet	Supervised	Dice loss	–	–	–	85.40% (DSC) 73.20% (Jaccard) 18.26 mm % (HD)	Dynamic extraction of region of interest of the pancreas based on the central point of surrounding organs (liver, kidney, and spleen). Initially pre-trained on ImageNet

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Qiu et al. (2022a)	Segmentation of pancreas	82 (NIH) 281 (MSD)	UNet3+ + Multi-scale feature calibration in both Localization and Segmentation	Supervised	Dice loss	4-fold cross-validation	–	–	NIH: 86.30% (DSC) 76.26% (Jaccard) 85.91% (Precision) 86.85% (Recall) MSD (Generalization): 85.41% (DSC)	Dual enhancement module to multiply the coarse segmentation probability map with the input image to coarse stage. Cropping of the output by the localizatio model. The cropped images are sent as input to fine stage. Multi-scale feature calibration module in both stages to calibrate features vertically to preserve boundary details and avoid feature redundancy
Qiu et al. (2022b)	Segmentation of pancreas	82 (NIH)	UNet-like with: Spiking neural P systems (Localization) + (Segmentation)	Supervised	Cross entropy	–	Nvidia Titan XP	–	81.94% (DSC)	Deep dynamic spiking neural P systems are integrated into UNet to solve memory limitation of 3D CNNs

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Qureshi et al. (2022)	Segmentation of pancreas	82 (NIH)	VGG-19 (Localization) + UNet (Segmentation)	Supervised	Mean Dice	4-fold cross-validation	Nvidia GeForce 2080 Ti	8 h	88.53% (DSC)	A morphology prior (a 3D volume template), defining the general shape and size of the pancreas, was integrated with the soft label from the second stage to improve segmentation
Shi et al. (2022)	Segmentation of pancreas endocardium, right and left ventricle, and myocardium	82 (NIH) two datasets of cardio structures	UNet V-Net ResNet-18	Semi- supervised	Cross entropy loss for conservative and radical model (labeled data) Cross entropy based for certain regions + Consistency loss for uncertain regions (Unlabeled data)	80% training, 20% test	Nvidia GeForce 2080 Ti	–	UNet: 67.01% (DSC) V-Net: 79.67% (DSC) 66.69% (Jaccard) 1.89 voxels (ASD) 7.59 voxels (HD) ResNet-18: 80.58 %(DSC) 67.91% (Jaccard) 2.27 voxels (ASD) 8.34 voxels (HD)	A conservative-radical module to automatically identify uncertain regions. A training strategy to separately segment certain and uncertain regions. Mean teacher model for uncertain region segmentation

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Yang et al. (2022b)	Segmentation of pancreas	82 (NIH) 281 (MSD)	AX-UNet with Atrous spatial pyramid pooling	Supervised	Focal loss Dice loss	4-fold cross-validation	Nvidia Tesla V100	–	NIH: 87.70% (DSC) 78.20% (Jaccard) 92.90% (Precision) 90.90% (Recall) MSD: 85.90% (DSC) 77.90% (Jaccard) 93.10% (Precision) 86.30% (Recall)	Design of a loss function to address the blurry boundary issue. Code available at: https://github.com/zhangyuhong02/AX-Unet.git
You et al. (2022)	Segmentation of pancreas, and left atrium	82 (NIH) datasets of other organs	V-Net for knowledge distillation	Semi- supervised	Cross entropy Dice loss Mean squared error (Supervised) Design of: Boundary-aware contrastive, Pair-wise distillation, and Consistency losses	80% training, 20% test	Nvidia 1080Ti	–	89.03% (DSC)	Contrastive distillation model with multi-task learning (segmentation map and signed distance map from boundary). Structured distillation in the latent feature space followed by contrasting the boundary-aware features in the prediction space for better representations

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Zeng et al. (2022)	Segmentation of pancreas, and left atrium	82 (NIH) datasets of other organs	V-Net	Semi- supervised	Cross entropy	80% training, 20% test	Nvidia Tesla V100	–	84.77% (DSC) 73.71% (Jaccard) 6.24 voxel (HD95) 1.58 voxel (ASD	Teacher-student trained in parallel: the student learns from pseudo labels generated by the teacher learning in turn from the performances of student on the labeled images
Dogan et al. (2021)	Segmentation of pancreas	82 (NIH)	Mask R-CNN (Localization) + UNet (Segmentation)	Supervised	Binary cross entropy	4-fold cross-validation	Nvidia GeForce RTX 2060	–	86.15% (DSC) 75.93% (Jaccard) 86.23% (Precision) 86.27% (Recall) 99.95% (Accuracy)	Less powerful GPUs are required
Huang et al. (2021b)	Segmentation of pancreas	82 (NIH)	UNet + Deformable convolutional module	Supervised	Design of Focal generalized Dice loss	4-fold cross-validation	Nvidia Tesla P40		87.25% (DSC) 88.98% (Precision) 89.97% (Recall)	Deformable convolutions for adaptive receptive fields. Focal generalized Dice loss to balance the size of foreground and background
Knolle et al. (2021)	Segmentation of pancreas, and brain	281 (MSD) Generalization: 85 (Internal dataset) dataset of brain	UNet-like with Dilated convolutions	Supervised	Dice loss	70% training, 30% test	–	–	78.00% (DSC) 1.78 mm (HD) Generalization: 70.00% (DSC)	Small network with dilated convolutions designed to run on low-end hardware within federated learning

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Li et al. (2021a)	Segmentation of pancreas	82 (NIH) 281 (MSD)	Bi-directional recurrent UNet	Supervised	Dice similarity coefficient loss	4-fold cross-validation	Nvidia GeForce RTX 2080Ti	–	NIH: 85.35% (DSC) 1.10 mm (ASD) 3.68 mm (HD) MSD: 85.65% (DSC)	Combination of a 2D slice with probabilistic map of two adjacent slice for local 3D context. The result is propagated through a 2.5D UNet. A bi-directional (forward-backward) recurrent scheme is applied to the primary segmentation to optimize the local 3D information.
Long et al. (2021)	Segmentation of pancreas	82 (NIH)	Encoder with channel attention to enhance semantics + Feature fusion pooling attention module + Decoder	Supervised	–	4-fold cross-validation	Nvidia GTX 1080Ti	–	86.62% (DSC) 86.07% (Precision) 87.37% (Recall)	Parallel module in the encoder to extract semantic and spatial features. Channel attention module to enhance acquisition of semantic information. Both modules sent as input to feature fusion pooling attention to fuse semantic and spatial information

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Ma et al. (2021a)	Segmentation of pancreas, lung, and cell contour	82 (NIH) datasets of other organs	UNet + Multi-scale convolutional block + Down-sampling + Context module	Supervised	Binary cross-entropy (lung) Dice loss (pancreas)	4-fold cross-validation	Nvidia GTX 1080Ti	–	NIH: 88.48% (DSC)	Customized UNet with: convolutional modules concatenating features from three branches; a hybrid pooling consisting of max-pooling, average pooling, and convolutions; skip connections with atrous convolutions (context module)
Ma et al. (2021b)	Segmentation of pancreas, liver tumor, kidney tumor, and left atrium	Combination of: 82 (NIH) 281 (MSD) dataset of another organ	VNet with Global active contour	Supervised	Dice loss L1 loss Geodesic active contour loss	5-fold cross-validation	Nvidia Tiatn V100	15–40 h	NIH + MSD: 83.60% (DSC) 18.50 mm (HD) 1.93 mm (ASSD)	First application integrating geodesic active contour into CNN to reduce boundary errors. Geodesic active contour loss can consider more global information than dice loss or cross entropy loss because it is built on the level set function-based representation

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Panda et al. (2021)	Segmentation of pancreas	1,917 (Mayo Clinic) 41 (TCIA) 80 (NIH)	UNet for two stages: Localization + Segmentation	Supervised	Tversky loss Asymmetric dice loss	72% training, 13% validation, 15% test	Nvidia Tesla V100	–	Internal dataset: 91.00% (DSC) TCIA (Generalization): 96.00% (DSC) NIH (Generalization): 89.00% (DSC)	Evaluation of dataset size on model performances: in the second stage 3D UNet was evaluated on 200; 500; 800; 1,000; 1,200; and 1,500 CTs (Internal dataset). Generalization on two datasets
Petit et al. (2021)	Segmentation of pancreas	82 (NIH)	UNet	Supervised Semi- supervised	–	5-fold cross-validation	Nvidia RTX 2080Ti	–	77.53% (DSC)	Fusion of a FCN probability prediction volume with 3D spatial prior representing the probability of organ presence
Shan and Yan (2021)	Segmentation of pancreas, skin lesions, and thyroid	82 (NIH) datasets of skin lesions and thyroid	UNet: Encoder with residual blocks Decoder with Spatial attention + Channel attention	Supervised	Soft dice loss	70% training, 10% validation, 20% test	Nvidia Tesla P100	–	91.37% (DSC) 85.30% (Jaccard) 30.79 mm (ASSD)	Spatial attention to focus on the target spatial regions and to ignore irrelevant background. Channel attention to highlight the relevant channels and reduce the irrelevant ones

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Tian et al. (2021)	Segmentation of pancreas	82 (NIH)	Markov chain Monte Carlo + UNet	Supervised	Binary cross entropy	88% training, 12% test	Nvidia GTX Titan X	–	87.49% (DSC) 84.12% (Precision) 93.81% (Recall)	Markov Chain Monte Carlo applied to 3D UNet for patch selection during localization and segmentation. This method solved theissue of memory limit, class imbalance, and data scarcity in 3D segmentation
Wang et al. (2021a)	Segmentation of pancreas	82 (NIH)	Dual input + v-mesh UNet + Attention + Spatial + Transformation and fusion	Supervised	Binary cross entropy	5-fold cross-validation	Nvidia Tesla P100	–	87.40% (DSC) 89.50% (PPV) 87.70% (Sensitivity)	Dual input FCN: original CT and images processed by graph-based visual saliency with specific intensity features to grasp more information on the boundary. Horizontal and vertical connections with attention mechanism. Spatial transformation and fusion for deformable convolutions

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Wang et al. (2021c)	Segmentation of pancreas	82 (NIH)	UNet (Localization) View adaptive Unet (Segmentation)	Supervised	Dice loss Weighted focal loss	4-fold cross-validation	Nvidia GTX 1080Ti	28 h for localization 31 h for segmentation	86.19% (DSC)	Data augmentation on three axes. Axial, coronal, and sagittal volumes are fed simultaneously to the network
Xue et al. (2021)	Segmentation of pancreas	82 (NIH) 59 (Fujian Medical University)	UNet for both: Localization and Segmentation	Supervised	Cross entropy Regression loss	4-fold cross-validation	Nvidia Titan XP	–	NIH: 85.90% (DSC) 75.70% (Jaccard) 87.60% (Precision) 85.20% (Recall) Fujian: 86.90% (DSC) 77.30% (Jaccard) 91.00% (Precision) 83.50% (Recall)	Multi-task second stage. Regression (task 1) of object skeletons as descriptor of the shape of the pancreas to guide subsequent segmentation (task 2). Conditional random fields to remove small false segments
Zhang et al. (2021a)	Segmentation of pancreas	82 (NIH) 281 (MSD)	CNN (Localization) Encoder-decoder (Segmentation) Prior propagation module (both stages) Scale-transferrable feature fusion module (second stage)	Supervised	Dice loss	4-fold cross-validation	Nvidia GTX 1080Ti	100 h	NIH: 84.90% (DSC) MSD: 85.56% (DSC)	Scale-transferrable feature fusion module to learn rich fusion features with lightweight architecture. Prior propagation module to explore informative and dynamic spatial priors to infer accurate and fine-level masks

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Zhang et al. (2021d)	Segmentation of pancreas	36 (ISICDM) 82 (NIH) 281 (MSD)	Multi-atlas registration (Localization) 3D patch-based and 2.5D slice-based UNet (Segmentation) 3D level set to refine the probability map (Refine stage)	Supervised	Cross entropy Dice coefficient loss	4-fold cross-validation	–	–	NIH: 84.47% (DSC) MSD: 82.47% (DSC)	Coarse stage for localization. Fine stage for segmentation: 3D patch-based and 2.5D slice-based CNN to extract local and global features. Refine stage to improve segmentation: 3D level-set for better boundary delineation.
Boers et al. (2020)	Segmentation of pancreas	82 (NIH)	UNet	Supervised	DSC-based loss weighted by voxel-specific map + Loss for volume difference	5-fold cross-validation	Nvidia RTX 2070	12 h	78.10% (DSC)	Scribbles are drawn to refine initial segmentation
Chen et al. (2020b)	Segmentation of pancreas	82 (NIH)	UNet, ResNet, 3D-DSN Encoder: Squeeze and excitation Decoder: Hierarchical fusion	Supervised	Weighted cross entropy	4-fold cross-validation	–	–	UNet: 87.04% (DSC) ResNet: 87.26% (DSC) DSN: 82.53% (DSC)	Hierarchical fusion model to retain boundary information

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Gong et al. (2020)	Segmentation of pancreas	40 (ISICDM)	UNet	Supervised	–	50% training, 50% validation	–	–	83.00% (DSC) 85.00% (Recall)	Fractional differentiation to increase the pancreas contrast. Level set (regularization term, intensity constraint term and length term to increase accuracy at contours)
Karimi and Salcudean (2020)	Segmentation of pancreas, liver, and prostate	282 images datasets of other organs	UNet	Supervised	Losses based on: distance transform Morphological operations Convolutions with circular/spherical kernels	4-fold cross-validation	Nvidia GeForce GTX Titan X	–	78.40% (DSC) 21.3 mm (HD) 1.84 mm (ASD)	Three different methods to reduce HD: distance-transform, morphological erosion, and convolutions with circular/spherical kernels. Three losses based on these methods for stable training

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Li et al. (2020a)	Segmentation of pancreas	82 (NIH)	UNet + Multi-scale convolution + Residual blocks	Supervised	Dice loss	4-fold cross-validation	Nvidia GeForce GTX Titan X	8 h	87.57% (DSC) 78.77% (Jaccard)	Three strategies to solve over-segmentation, under-segmentation, and shape inconsistency: skip network (adding residuals between encoder and decoder directly), residual network (adding residuals to the continuous convolution blocks of the encoder and decoder separately) multi-scale residual network (with multi-scale convolution block between high-resolution endocer and decoder)

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Li et al. (2020c)	Segmentation of pancreas	82 (NIH) 281 (MSD)	Multi-scale Attention dense Residual UNet	Supervised	Binary cross entropy Dice loss	4-fold cross-validation	Nvidia GeForce RTX 2080Ti	–	NIH: 86.10% (DSC) 75.55% (Jaccard) 86.43% (Sensitivity) 84.97% (Specificity) 4.40mm (HD) 1.27 mm (ASD) MSD: 88.52% (DSC) 79.42% (Jaccard) 91.86% (Sensitivity) 89.66% (Specificity) 3.78 mm (HD) 0.95 mm (ASD	Multi-scale convolution and channel attention to solve interclass indistinction. Dense residual blocks to solve intraclass inconsistency. Code available at: https://github.com/Mrqins/ pancreas-segmentation
Nishio et al. (2020)	Segmentation of pancreas	82 (NIH)	deepUNet	Supervised	Dice loss	4-fold cross-validation	–	–	78.90% (DSC) 65.80% (Jaccard) 76.20% (Recall)	Use of three data augmentation methods: conventional ones, mixup, and random image cropping and patching

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Zheng et al. (2020)	Segmentation of pancreas	82 (NIH)	3D VNet (Localization) 2.5D Encoder-decoder (Segmentation)	Self supervised	Square root Dice loss	4-fold cross-validation	4 Nvidia Titan XP	–	78.10% (DSC)	Square Root Dice loss to deal with the trade-off between sensitivity and specificity. Slice shuffle for pre-training before input to the network which learns to reorder and understand organ shape. Capturing of non-local information through attention, pooling, and convolutional layers. Ensemble learning and recurrent refinement to improve accuracy
Zhu et al. (2020)	Segmentation of pancreas, liver, and prostate	82 (NIH) datasets of other organs	UNet + Residual blocks + Attention focused modules	Supervised	–	76% training, 24% test	–	–	83.90% (DSC)	Attention modules into skip connections to focus on segmented regions and reduce influence of background. Dense connected residual blocks in down-sampling and up-sampling to reduce computational load and network parameters. Code available at: https://github.com/ahukui/ SIPNet

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Liu et al. (2020)	Segmentation of pancreas	82 (NIH)	ResNet (Localization) Ensemble UNet (Segmentation)	Supervised	Dice loss Focal loss Jaccard distance loss Class balanced cross entropy Binary cross entropy	4-fold cross-validation	Nvidia Titan X	–	84.10% (DSC) 72.86% (Jaccard) 84.35% (Precision) 85.33% (Recall)	Superpixes generated by oversegmentation. Classification of superpixels by ResNet. candidate regions obtained by ensemble of classification results of three different scale of superpixels. Segmentation by ensemble of multiple network with different loss functions
Man et al. (2019)	Segmentation of pancreas	82 (NIH)	Localization agent (Localization) + Deformable UNet (Segmentation)	Reinforcement (Localization) Supervised (Segmentation)	Dice loss	4-fold cross-validation	–	–	86.93% (DSC)	First application of Deep Q Learning to medical image segmentation. Localization agent to adjust localization, by learning a localization error correction policy based on deep Q network. Deformable convolution for learnable receptive fields, instead of fix ones

Table 4 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Zeng and Zheng (2019)	Segmentation of pancreas, hip, and lumbar intravertebral discs	82 (NIH) datasets of other organs	UNet + Holistic Decomposition Convolution + Dense Upsampling Convolution	Supervised	Cross entropy loss Dice loss	4-fold cross-validation	Nvidia GTX 1080 Ti	–	83.00% (DSC)	Network agnostic segmentation approach. Holistic decomposition convolution to reduce size of data for subsequent processing: periodic down-shuffling to input to get low resolution channels, followed by convolutions on these channels. Periodic dense upsampling convolutions to recover full resolution: low resolution convolutions with periodic up-shuffling
Heinrich et al. (2018)	Segmentation of pancreas	82 (NIH)	UNet + Ternary weights + Ternary activations	Supervised	Weighted cross entropy	5-fold cross-validation	Nvidia Titan XP	15 min	71.00% (DSC)	Implementation of TernaryNet with ternary weigths and ternary hyberbolic tangent to reduce computational load.

Table 5 Reviewed studies on the segmentation of pancreas parenchyma with attention mechanism applied to UNet and CNNs architectures

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Shen et al. (2023)	Segmentation pancreas, duodenum, gallbladder, liver, and stomach	42 (TCIA)	UNet with: Spatial attention (location and size of organs) + Dilated convolution + Multi-scale attention	Supervised	Dice loss Cross entropy loss	4-fold cross-validation	Nvidia GTX 1080 Ti	–	75.42% (DSC) 61.84% (Jaccard) 19.99 mm (HD)	Spatial attention to highlight location and sizes of target organs (pancreas, duodenum, gallbladder, liver, and stomach). Deformable convolutional blocks to deal with variations in shapes and sizes. Skip connections with multi-scale attention to eliminate interference of complex background
Wu et al. (2023)	Segmentation of pancreas, left ventricle, myocardium, right ventricle, and colon	80 (NIH) datasets of colon and cardiac structures	V-Net + Attention	Semi- supervised	Cross entropy Dice loss	4-fold cross-validation	Nvidia Titan RTX	–	74.03% (DSC) 59.70% (Jaccard) 2.12 voxel (ASD) 9.10 voxel (HD95)	Instead of using model predictions as pseudo labels, high-quality pseudo labels are generated by comparing multiple confidence maps produced by different networks to select the most confident one (a compete-to-win strategy. A boundary-aware enhancement module was integrated to enhance boundary discriminative features. Code available at: https://github.com/ Huiimin5/comwin
Xia et al. (2023)	Segmentation of pancreas	82 (NIH)	VNet + Multi-dimensional Feature attention	Semi- supervised	Cross entropy Dice loss Mean square error	4-fold cross-validation	Nvidia GeForce RTX 3060	–	79.55% (DSC) 66.87% (Jaccard) 7.67 mm (HD95) 1.65 mm (MSD)	Multi-dimensional feature attention and improved cross pseudo supervision to effectively use unlabeled data reducing the need of labeled data

Table 5 (continued)

Author	Application	Data-set size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Chen et al. (2022a)	Segmentation of pancreas	82 (NIH) 281 (MSD)	Encoder-Decoder Attention feature fusion (Localization) Encoder-Decoder Attention feature fusion Coordinate Multi-scale Attention (Segmentation)	Supervised	Dice loss Binary cross entropy	4-fold cross-validation	Nvidia GeForce RTX 3090		NIH: 85.41% (DSC) 74.80% (Jaccard) 85.60% (Precision) 85.90% (Recall) MSD: 70.00−80.00% (DSC) 60.00% (Jaccard) 80.00−90.00% (Precision) 60.00−70.00% (Recall)	Attention feature fusion on low and high level features to keep context. Multi-scale attention to aggregate long-range dependencies, positional information, and exploit multi-scale spatial information
Chen et al. (2022b)	Segmentation of pancreas	82 (NIH) 281 (MSD)	VGG-16 with Attention gate (Localization) VGG-16 with Residual multi-scale dilated attention (Segmentation)	Supervised	Dice loss Binary cross entropy	4-fold cross-validation	Nvidia GeForce GTX 1080 Ti	11 h	NIH: 85.19% (DSC) 74.19% (Jaccard) 86.09% (Precision) 84.58% (Recall) MSD (generalization): 76.60% (DSC) 62.60% (Jaccard) 87.70% (Precision) 69.20% (Recall)	Attention gate used in the localization stage to suppress irrelevant background regions. Weight conversion module to transform segmentation map of the first stage into spatial weights to refine input of the second stage. Residual multi-scale dilated attention to exploit inter-channel relationships and extract multi-scale spatial information. Code available at: https://github.com/meiguiyulu/TVM

Table 5 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Chen et al. (2022d)	Segmentation of pancreas	82 (NIH) 281 (MSD)	UNet (Localization) Unet with: Fuzzy skip connection + Target attention in the decoder (Segmentation)	Supervised	Dice loss	4-fold cross-validation	2 Nvidia GeForce RTX 2080 Ti		NIH: 87.91% (DSC) 78.52% (Jaccard) 90.43% (Precision) 85.77% (Recall) MSD: 84.40% (DSC)	Fuzzy skip connections to reduce the redundant information of non-target regions. Attention to make the decoder more sensitive to target features
Cui et al. (2022)	Segmentation of pancreas, and liver	82 (NIH) dataset of liver	UNet++ with: Channel attention (Squeeze and excitation) + Skip connection with sharpening filter	Supervised	Dice loss Cross entropy loss	70% training, 10% validation, 20% test	Nvidia GeForce RTX 3090		83.64% (DSC) 83.97% (Precision) 84.20% (Recall)	Laplacian sharpening filter integrated into skip connections to reduce bad artifacts. Squeeze and excitation to help the model to focus on the features beneficial for segmentation
Liu et al. (2022b)	Segmentation of pancreas	82 (NIH) 72 (ISICDM) challenge dataset of spleen	ResNet18 + Atrous spatial pyramid pooling for multi-scale feature extraction for both Localization and Segmentation + Saliency module for fusion	Supervised	Dice loss (Region and boundary level) Binary cross entropy (Pixel level)	4-fold cross-validation	Nvidia Titan RTX	–	NIH: 88.01% (DSC) ISICDM: 87.63% (DSC)	Segmentation network with three branches to extract pixel, boundary, and region features, fused by a saliency module. Design of a loss function integrating information from pixel-level classification, edge-level localization, and region-level segmentation

Table 5 (continued)

Author	Application	Data-set size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Qu et al. (2022)	Segmentation of pancreas	224 (Internal) Peking Union Medical College Hospital 66 (External) Hehan Cancer Hospital	M3Net (3D Encoder- 2D Decoder): Multi-scale Multi-view Multi-phase + Attention	Supervised	Binary cross entropy Mean square error of distance field between foreground and background	62% training, 9% validation, 29% test	Nvidia Titan X	–	Internal: 90.29% (DSC) External: 86.34% (DSC)	Dual path segmentation models for arterial and venous phase. Each model is constituted by an encoder composed of 3D convolutions and a decoder composed of 2D convolutions. Inter-phase contextual information is explored via cross-phase non-local attention between the two models. Replication for axial, coronal, and sagittal views. Ensemble of the three views. Fusion of high and half resolution to capture local and global features
Zhu et al. (2022)	Segmentation of pancreas	82 (NIH) 281 (MSD) 70 (Zheyi)	Residual blocks + Squeeze-Excitation Attention + UNet	Domain adaptation: Supervised learning (source) Unsupervised learning (target)	Cross entropy Dice loss	5-fold cross-validation	Nvidia Titan V	–	NIH adapted to Zheyi 72.73% (DSC) MSD adapted to Zheyi 71.17% (DSC)	Adversarial multiscale domain adaption (from source) to generalize to external datasets (target domain)
Yan and Zhang (2021)	Segmentation of pancreas	82 (NIH)	UNet + Spatial attention + Channel attention (Localization and Segmentation)	Supervised	Dice loss	4-fold cross-validation	Nvidia GeForce GTX 1060	–	86.61% (DSC)	2.5D UNet with spatial and channel attention integrated into skip connections.

Table 5 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Zhang et al. (2021c)	Segmentation of pancreas and brain	82 (NIH) dataset of other organs	Shared encoder and two decoders. Second decoder: Context residual Mapping + Context residual Attention	Supervised	Binary cross entropy Dice loss	4-fold cross-validation	Nvidia GeForce RTX 2080 Ti	–	86.06% (DSC)	The context residual decoder takes the residual feature maps of adjacent slices produced by the decoder as its input, and provides feedback to the segmentation decoder as a kind of attention guidance
Lu et al. (2019)	Segmentation of pancreas	82 (NIH)	UNet + Channel attention + Spatial attention + Ring residual module	Supervised	Design of Complex-coefficient loss	10-fold cross-validation	Nvidia GeForce RTX 2080	6 h	88.32% (DSC)	Ringed residual module, consisting of forward and backward residual propagation to address the boundary blur issue of pancreas. Convolutional block attention module with spatial and channel attention to improve accuracy. Complex-coefficient loss to focus not only on the ratio of the coincident area to the total area, but also on the shape similarity between the real result and the predicted result
Schlemper et al. (2019)	Segmentation of pancreas	82 (NIH)	UNet + Attention gate	Supervised	Dice loss	4-fold cross-validation	–	–	83.10% (DSC) 82.50% (Jaccard) 84.10% (Recall)	Attention gate in integrated into skip connections of UNet to highlight salient features and suppress irrelevant regions Code available at: https://github.com/ozan-oktay/Attention-Gated-Networks

Table 5 (continued)

Author	Application	Data-set size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Wang et al. (2019a)	Segmentation of pancreas liver colon hippocampus brain tumor	281 (MSD) datasets of other organs	UNet with: Residual blocks nested with dilations + Squeeze and excitation	Supervised	Focal loss	70% training, 10% validation, 20% test	–	–	84.76% (DSC)	Residual blocks nested with dilations added in the first few layers to help network adapt to targets of any size. Squeeze and excitation to boost essential features for each task

Table 6 Reviewed studies on the segmentation of pancreas parenchyma with transformers and hybrid transformers architectures

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Dai et al. (2023)	Segmentation of pancreas	82 (NIH) 281 (MSD)	UNet (Localization) Deformable convolution Vision Transformer (Segmentation)	Supervised	Binary cross entropy Dice loss	4-fold cross-validation	Nvidia GeForce RTX 3090	4 h	NIH: 89.89% (DSC) 89.59% (Precision) 91.13% (Recall) MSD: 91.22% (DSC) 93.22% (Precision) 91.35% (Recall)	Skip connections integrating: vision transformer, deformable convolutions, and scale interactive fusion (combining global and local features, and merging feature maps of different scales). Two-dimensional wavelet decomposition to solve the issue of blurred boundaries
Chen and Wan (2022)	Segmentation of pancreas	82 (NIH)	UNet with Transformers in skip connections	Supervised	Dice loss Focal loss	4-fold cross-validation	Nvidia GeForce RTX 3080	–	86.80% (DSC) 76.90% (Jaccard) 87.60% (Precision) 88.00% (Recall)	3D channel transformer in the skip connections of UNet. Attention module in each encoder level with project and excite block to enhance extraction of context information. Cross attention between output of each transformer and decoder to eliminate semantic inconsistency
Fang et al. (2023)	Segmentation of pancreas	281 (MSD) 91 (Clinical Proteomic Tumor Analysis Consortium Pancreatic Ductal Adenocarcinoma)	U-shaped Encoder-Decoder CNN + Transformer	Supervised	Dice loss Cross entropy loss	70% training, 10% validation, 20% test	Nvidia GeForce RTX 3090	–	MSD: 77.36% (DSC) 8.34 (95HD) Tumor dataset: 85.54% (DSC) 4.05 mm (HD95)	A transformer module with multi-head self-attention and residual convolutional block was designed to capture both local and global features. Code available at: https://github.com/ supersunshinefk/ UMRFormer-Net

Table 6 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Qiu et al. (2023)	Segmentation of pancreas	82 (NIH)	DeepUNet (Localization) Residual transformer UNet (Segmentation)	Supervised	Dice loss with Hausdorff distance term	4-fold cross-validation	–	48 h	86.25% (DSC)	UNet like network with each convolutional block consisting of residuals blocks, residual transformers, and dual convolution down-sampling (for translational equivariance)

Table 7 Reviewed studies on the segmentation of pancreas parenchyma with GAN architectures

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Ge et al. (2023)	Segmentation of pancreas	45 (Nanjing Drum Tower Hospital) (Reconstruction) 15 (Nanjing General PLA Hospital) for generalization 90 (liver tumor) for generalization	Average Super Resolution GAN with: 3D CNN (Reconstruction) + 3D UNet for both Localization and Segmentation	Supervised	Mean squared error loss Dice loss Cross entropy loss	80% training and validation, 20% test	2 Nvidia Titan XP	20 h	Generalization (pancreas): 84.20% (DSC) 0.54 mm (ASD)	GAN: Super resolution network to reduce anisotropy resolution. A generator reconstructs thin slices in z axis The discriminator optimizes the output of generator. The optimized generated images are sent to a dual-stage network for segmentation. Predictions on high-resolution are down-sampled to restore initial resolution
Li et al. (2022d)	Segmentation of pancreas	82 (NIH)	Attention-guided Dual adversarial UNet (ADAU-Net)	Supervised	Basic loss of conventional segmentation Adversarial loss	4-fold cross-validation	Nvidia GeForce GTX 1080 Ti	–	83.76% (DSC) 72.38% (Jaccard) 1.07mm (ASD) 2.17 (RMSE)	First dual adversarial network with an attention mechanism for pancreas segmentation
Li et al. (2021b)	Segmentation of pancreas	82 (NIH)	Dual GAN with UNet (Generators) and CNNs (Discriminators) + Pyramidal pooling	Supervised	Adversarial loss	4-fold cross-validation	Nvidia GeForce GTX 1080 Ti	–	83.31% (DSC) 71.76% (Jaccard) 84.09% (Precision) 83.30% (Recall)	First GAN to preserve spatial information. Second GAN increases the preservation of spatial information and leads to more realistic segmentation results. Pyramidal pooling to replace original pooling layers in UNet

Table 7 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Li et al. (2021c)	Segmentation of pancreas	82 (NIH)	Multi-scale selection Adversarial Multi-channel fusion UNet	Supervised	Basic loss of conventional segmentation Adversarial loss	4-fold cross-validation	Nvidia GeForce RTX 1080 Ti	–	84.10% (Precision) 82.50% (DSC)	GAN with a generator integrating: Multi-scale field selection to grasp global spatial features; Multi-channel fusion integrating information from different locations to obtain comprehensive details
Li et al. (2021d)	Segmentation of pancreas	82 (NIH)	Multi-level pyramidal pooling Adversarial UNet	Supervised	Adversarial loss	4-fold cross-validation	Nvidia GeForce RTX 1080 Ti	–	83.03% (DSC) 84.60% Recall)	Generator consisting of: UNet with residual blocks, and multi-level pyramidal pooling to gather contextual information
Li et al. (2021e)	Segmentation of pacnreas	82 (NIH)	Dual GAN + Unet (DAUnet) + CNN for multilevel cue	Supervised	Adversarial loss	4-fold cross-validation	Nvidia GeForce RTX 1080 Ti	–	83.08% (DSC) 71.39% (Jaccard) 82.19% (Recall) 2.22 mm (RMSE)	Two GANs with generators based on UNet. The second one fuses features from different convolutional layers to obtain additional details for segmentation.
Li et al. (2020b)	Segmentation of pancreas	82 (NIH)	Two stages GAN based on UNet	Supervised	Conventional segmentation loss Adversarial loss	4-fold cross-validation	Nvidia GeForce RTX 1080 Ti	–	83.06% (DSC) 71.41% (Jaccard)	Adversarial training on a model already trained with GAN

Table 7 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Ning et al. (2020)	Segmentation of pancreas	82 (NIH)	GAN Generator with: Autoencoder with Dilated convolution + LSTM	Supervised	Adversarial loss	4-fold cross-validation	Nvidia Titan X	–	89.87% (DSC) 95.85% (Accuracy)	First application integrating dilated convolutions, GAN, and LSTM. Generator: dilated convolution autoencoder with dilated convolution layers in the encoder, and an LSTM boosting the pancreas boundary segmentation by modeling the contextual spatial correlation between neighbouring CT scan patches

learning. They are described in Sections 1.2 and 1.3 of the Appendix, respectively. The code was publicly available for seven studies.

Despite the promise of transformers, either alone or in combination with CNNs to form hybrid models, and the different architectures proposed for two-stage approaches, a UNet configuration with residual blocks in the encoder and a decoder with spatial and channel attention obtained the highest DSC score (91.37%) on the NIH dataset (Shan and Yan 2021). This was followed by a two-stage hybrid transformer, with a UNet for localization and ViT for segmentation, reporting a DSC of 89.89% on NIH (Dai et al. 2023). In comparison, the highest DSC using CNNs and GAN was 88.68% (Pradip and Kakarwal 2023) and 89.87% (Ning et al. 2020), respectively. Notably, several two-stage approaches reached almost the same value of DSC (slightly above 86.0%) using UNet for localization and residual transformer with UNet for segmentation (Qiu et al. 2023), UNet3+ with multi-scale feature calibration in both stages (Qiu et al. 2022b), mask R-CNN for localization and UNet for segmentation (Dogan et al. 2021), UNet in both stages (Wang et al. 2021c), VGG with attention gate for localization) and VGG-16 with residual multi-scale dilated attention for segmentation (Chen et al. 2022b). Concerning the MSD dataset, the two-stage hybrid transformer (TD-Net) proposed by Dai et al. (2023) achieved the highest DSC (91.22%), while it was ranked second for the NIH (see above). It was followed by a multiscale attention-dense residual UNet (MAD-UNet) reaching 88.52%. One study combined both NIH and MSD into a dataset of 283 CT reporting a DSC of 83.60% with VNet (Ma et al. 2021b).

The majority of the studies on NIH and MSD datasets reported performances using region-based metrics like DSC and Jaccard, neglecting the importance of boundary-based metrics (Ma et al. 2022b).

Overall, the methods using supervised learning achieved higher DSC scores on the NIH dataset than those based on semi-supervised and unsupervised learning. VNet models obtained the highest DSC score on semi-supervised (89.03%) and unsupervised learning (78.10%), respectively (You et al. 2022; Zheng et al. 2020).

A typical limitation of DL models is the lack of a demonstration of how they perform on external datasets. Therefore, some methods were proposed to address this limitation for pancreas segmentation. They were described in Section 1.5 of the Appendix. Of all the studies on NIH few were tested for generalization, in all cases on the MSD dataset, reaching a DSC of 85.41% Qiu et al. (2022b), 76.60% Chen et al. (2022b), and 81.09% Li et al. (2023a) using supervised learning. The latter included also an internal dataset of 104 CTs for training in addition to NIH. Concerning self-supervised learning for model generalization from NIH to MSD, a UNet for both localization and segmentation achieved a DSC of 66.73% (Li et al. 2023c).

Some large private datasets were internally curated. For instance, a two-stage model based on UNet for both localization and segmentation was applied to a dataset of 1,917 CTs from Mayo Clinic (United States). This model reached 91.00% of DSC on 41 cases of TCIA. When generalized to the NIH dataset, the DSC score was 89.00% (Panda et al. 2021). A subset of this dataset (294 cases) reported a slightly lower DSC (88.00%) using a similar two-stage architecture on UNet (Khasawneh et al. 2022).

4.3 Studies on datasets with multi-organ annotations

The 23 studies were grouped as follows: CNNs (n= 4, Table 8), UNet and variants (n= 11, Table 9), attention applied to CNNs and UNet (n= 5, Table 10), transformers (n= 2, 11), and GAN (n= 1, Table 12). They are described in Section 2 of the Appendix. The BTCV dataset was the most used dataset, recurring in 11 out of 23 studies (47.8%), followed by TCIA in five (21.7%), Synapse in four (17.4%), and AbdomenCT-1k in three (13.0%). Pancreas segmentation was based on a two-stage (coarse-fine) approach in four works (Li et al. 2022a, 2023d; Roth et al. 2018b; Tong et al. 2023). Four studies designed specific loss functions, described in Section 2.1 of the Appendix. Almost all the studies were based on supervised learning. Two used partially supervised learning (Shi et al. 2021; Zhang et al. 2021b), one used both semi-supervised and unsupervised (Xia et al. 2020), one partially supervised for abdominal organs and semi-supervised learning for other organs (Liu and Zheng 2023), while another study employed supervised, semi-supervised, weakly, and continual learning (Ma et al. 2022b). The code was publicly available for six studies.

Being a multi-organ dataset, AbdomenCT-1k (Sect. 3.6) was used in studies needing an annotated dataset for the pancreas and in those evaluating DL models for segmentation on four labeled organs (pancreas, spleen, kidneys, and liver). The most comprehensive analysis on AbdomenCT-1k was performed by Ma et al. (2022b). In addition to DSC, NSD was used for the assessment of segmentation results at boundary level (Ma et al. 2022b). When using the MSD subset of AbdomenCT-1k for training, nnU-Net reached a DSC of 86.10% (with only annotations of the pancreas), while the metric value rose to 90.10% if nnU-Net was trained on MSD (with annotations of the pancreas, liver, spleen, and kidneys) and tested on the liver tumor part of AbdomenCT-1k (Ma et al. 2022b). If trained with supervised learning with MSD plus 40 cases of liver tumors, and 40 of kidney tumors the score of DSC dropped to 78.10% when tested in 50 challenging and 50 random cases. When tested on the 50 CT scans of Nanjing University of the AbdomenCT-1k dataset (Sect. 3.6) with cancers of the colon, pancreas, and liver, but keeping the same training strategy, nnU-Net reached 82.50% and 82.30% for supervised and semi-supervised learning, respectively (Ma et al. 2022b). Evidence showed that DSC can vary substantially when choosing a random subset of AbdomenCT-1k. A UNet for localization and another UNet for segmentation, with multi-branch feature attention in the encoder and feature attention aggregation in the decoder, obtained a DSC of 86.20% on 500 random cases AbdomenCT-1k (Li et al. 2023d). In contrast, in the same work the DSC on 240 random scans of AMOS-CT dropped to 78.40% using the same network (Li et al. 2023d). TCIA and BTCV datasets were used alone or in combination, typically with 43 scans of the former and 47 of the latter. In all cases the proposed methods were based on UNet variants or encoder-decoder, reaching a maximum DSC of 84.00% for a two-stage model with an encoder-decoder for localization and a 2.5D network for segmentation, improving the results of the first study (78.00%) using the DenseVNet model (Li et al. 2022a; Gibson et al. 2018). When training combined BTCV, NIH, and datasets for other organs, a DSC of 88.00% was reported on a test set consisting of NIH and BTCV (Zhao et al. 2022a). The dataset with the lowest DSC scores was the Synapse one. In this case, the highest DSC values (65.67%) were reported by hybrid transformers (Table 11), while a UNet-like model with attention obtained 62.77% (Yuan et al. 2023). Another large private dataset of 1,150 CTs was curated at John Hopkins (United States) with annotations

Table 8 Reviewed studies on the segmentation of pancreas parenchyma with CNNs on multi-organ annotations

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Li et al. (2022a)	Segmentation of spleen, kidney, gallbladder, esophagus, liver, stomach, pancreas, and duodenum	90: 43 (TCIA) 47 (BTCV) for training 511 (FLARE) for generalization	3D Encoder-Decoder (Localization) 2.5D network (Segmentation)	Supervised	Design of parameter loss to remove the false positive of dice loss	4-fold cross-validation	Nvidia Tesla V100	–	TCIA+BTCV: 84.00% (DSC) 5.67 mm (HD95) FLARE: 83.00% (DSC)	Circular inference (a sort of micro-attention mechanism) and parameter Dice loss in the first stag to reduce uncertain probabilities of blurred boundaries.
Park et al. (2020)	Segmentation of pancreas and other 16 anatomical structures	1,150 (John Hopkins)	CNN-based Two-stage Organ attention network	Supervised	–	4-fold cross-validation	–	–	87.80% (DSC)	Annotation of 22 structures. Use of two-stage organ attention network. The first used reverse connections to get more semantic information. The results became attention-organ module to guide the second network. This architecture was applied to each view. The outputs from axial, coronal, and sagittal views were then fused

Table 8 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Xia et al. (2020)	Segmentation of pancreas	82 (NIH) for training 90: 43 (TCIA) 47 (BTCV) for validation 281 (MSD) MSD liver for domain adapation	Encoder-Decoder based on ResNet18 for Multi-view Co-training and Domain-adaptation	Semi- supervised Unsupervised	Combination of conventional segmentation loss (labeled) and computational function based on uncertainty-weighted label fusion (unlabeled)	76% training, 24% test	Nvidia Titan RTX	24 h	NIH (semi-supervised): 78.77%–81.18% (DSC) TCIA+BTCV (semi-supervised): 73.86%–77.91% (DSC) MSD (unsupervised): 74.38% (DSC)	Co-training to maximize the similarity of the predictions among different views, generated by rotation or permutation transformations. Uncertainty weighted label fusion module for accurate pseudo labels generation for each view. Adaptation from multi-organ to pancreas dataset without source domain data
Chen et al. (2018)	Segmentation of liver, spleen, pancreas, and kidkeys	150 (Internal) 133 (Stroke) 285 (Brain tumor)	DRINet	Supervised	Cross entropy loss	50% training, 50% test	Nvidia Titan XP	–	83.42% (DSC) 87.95% (Precision) 80.29% (Recall)	Implementation of DRINet consisting of dense connection blocks, residual inception, and unpooling blocks

Table 9 Reviewed studies on the segmentation of pancreas parenchyma with UNet on multi-organ annotations

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Li et al. (2023d)	Segmentation of liver, kidney, spleen, and pancreas	500 (AbdomenCT-1k) for train and test 240 (AMOS-CT) for train and test	3D UNet (Localization) UNet with: Multi-branches feature attention (Encoder) + Feature attention aggregation (Decoder) (Segmentation)	Supervised	Dice loss	80% training, 20% validation	Nvidia RTX A6000	–	AbdomenCT-1k: 86.20% (DSC) AMOS-CT: 78.40% (DSC)	Network with self-adjustable attention and receptive field size to segment liver, kidney, spleen, and pancreas. Different kernel sizes to capture different scale features of different organs using: multi-branch feature attention with four branches, and feature attention aggregation with two branches
Liu and Zheng (2023)	Segmentation of liver, spleen, pancreas, and kidneys	30 (BTCV) 281 (MSD) liver MSD spleen MSD kidney tumor for training and test dataset of heart, pulmonary artery, and aorta for training and test	nnU-Net	Partially supervised (abdominal organs) Semi- supervised (other organs)	Cross entropy loss Dice loss (labeled data) Context-aware voxel-wise contrastive learning loss (unlabeled data)	60% training, 20% validation, 20% test	Nvidia Tesla V100	–	BTCV: 80.60% (DSC) 3.56 mm (HD95) MSD: 83.60% (DSC) 4.30 mm (HD95)	Exploiting unlabeled information in partially labeled datasets. Context-aware voxel-wise contrastive learning inserted into the bottleneck layer of a 3D nnU-Net to increase context awareness in patch-based strategy

Table 9 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Pan et al. (2022)	Segmentation of spleen, kidneys, gallbladder, esophagus, liver, stomach, aorta, vena cava, splenic veins, adrenal glands, and pancreas	59 (Institutional dataset with: heart, kidneys, liver, lungs, spinal cord, and stomach) 30 (Synapse)	VNet with Multi-layer perceptron Mixer replacing CNN	Supervised	Cross entropy Dice loss	5-fold cross-validation	Nvidia RTX 6000	–	BTCV: 79.00% (DSC) 5.73 mm (HD)	Multi-layer perceptron mixer was integrated into VNet to linearize the computational complexity of transformers
Sundar et al. (2022)	Segmentation of adrenal glands, aorta, bladder, brain, heart, kidneys, liver, pancreas, spleen	50 (Internal)	nnU-Net	Supervised	–	80% training, 20% test	–	48 h	85.00% (DSC)	Development of Multiple-organ objective segmentation (MOOSE) framework. Code available at: https://github.com/QIMP-Team/ MOOSE

Table 9 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Zhao et al. (2022a)	Segmentation of spleen, kidneys, gallbladder, esophagus, liver, stomach, aorta, veins, pancreas, duodenum, colon, lung spinal cord, and heart	82 (NIH) 47 (BTCV) liver tumor, MSD colon, datasets of other organs for training	UNet	Supervised	Mean square error loss	80% training, 5% validation, 15% test	Nvidia GeForce RTX 1650		NIH+BTCV: 88.00% (DSC)	3D UNet used for contour interpolation
Isensee et al. (2020)	Segmentation of heart, atrium, ventricles, myocardium, aorta, trachea, lung, hyppocampus, esophagus, liver, kidneys, pancreas, spleen, colon, gallbladder, and stomach	281 (MSD) BTCV datasets of other organs	nnU-Net	Supervised	Cross entropy loss Dice loss Weighted binary cross entropy loss	5-fold cross-validation			MSD: 2D UNet: 77.38% (DSC) 3D UNet Full resolution: 82.17% (DSC) 3D UNet low resolution: 81.18% (DSC)	Original paper on the implementation of nnU-Net. nnU-Net has three configurations: 2D UNet, 3D UNet with full resolution, and 3D UNet with low resolution. Code available at: https://github.com/MIC-DKFZ/nnUNet ?tab=readme-ov-file

Table 9 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Ma et al. (2022b)	Segmentation of liver, kidney, spleen, and pancreas	1,112 (AbdomenCT-1k)	3D nnU-Net (Supervised and semi-supervised) 2D nnU-Net + Conditional random fields (Weakly supervised) nnU-Net (Continual)	Supervised Semi-supervised Weakly supervised Continual	Dice loss Cross entropy loss	5-fold cross-validation	Nvidia Titan V100 or Nvidia GeForce RTX 2080Ti	–	Training on MSD and test on the others: 80.00%−90.10% (DSC) 63.10%−82.30% (NSD) Training on LiTS and test on the others: 81.10%−86.10% (DSC) 61.40%−78.80% (NSD) Training on KiTS and test on the others: 80.50%−87.40% (DSC) 61.50%−76.70% (NSD) Training on Spleen and test on the others: 80.20%−88.80% (DSC) 60.70%−79.90% (NSD) Benchmarks with test on Nanjing University (50) Supervised (training on MSD and NIH, or liver (40), and kidney (40)): 76.30%−82.50% (DSC) 88.30%−93.60% (NSD) Semi-supervised: 72.50%−82.30% (DSC) 57.90%-70%−70.70% (NSD) Weakly supervised: 16.80%−30.60% (DSC) 14.80%−21.40% (NSD) Continual: 59.10% (DSC) 85.40% (NSD)	Presentation of a large dataset with the addition of multi-organ (liver, kidney, spleen, and pancreas) annotations to original datasets. Definition of benchmark and baseline for supervised, semi-supervised, weakly supervised, and continual learning. Code available at: https://github.com/JunMa11/AbdomenCT-1K

Table 9 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Shi et al. (2021)	Segmentation of liver, spleen, pancreas, and kidney	30 (BTCV) 281 (MSD) MSD liver MSD spleen kidney tumor	nnU-Net	Partially supervised	Marginal loss Exclusive loss	80% training, 20% test	–	–	Training on BTCV: BTCV (test): 80.20% (DSC) 6.31 mm (HD) MSD (test): 69.50% (DSC) 21.37 mm (HD) Training on MSD: BTCV (test): 68.70% (DSC) 18.36 mm (HD) MSD (test): 72.80% (DSC) 5.46 mm (HD) Training on BTCV+MSD: BTCV (test): 82.30% (DSC) 8.16 mm (HD) MSD (test): 75.30% (DSC) 8.56 mm (HD) Training on all datasets: BTCV (test): 83.60% (DSC) 3.24 mm (HD) MSD (test): 80.80% (DSC) 3.96 mm (HD)	Implementation of marginal loss (for background) label and exclusion loss (different organs are mutually exclusive)

Table 9 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Zhang et al. (2021b)	Segmentation of liver, pancreas, spleen, and kidney	For training: 281 (MSD) liver MSD, spleen MSD, kidney tumor For test: 30 (BTCV) 82 (NIH) dataset of liver	nnU-Net + Auxiliary information into decoder	Partially supervised	Dice loss Focal loss	4-fold cross-validation	–	–	Training on MSD and test on liver and NIH: 77.33% (DSC) 64.01% (Jaccard) Training on four datasets and test on liver and NIH: 54.12%−79.94% (DSC) 40.36%−66.89% (Jaccard) Training on MSD and test on BTCV: 81.37% (DSC) 68.88% (Jaccard) Training on four datasets and test on BTCV: 69.51%−82.93% (DSC) 55.32%−70.01% (Jaccard)	Four datasets with annotations of different organs (liver, pancreas, spleen, and kidney). An auxiliary conditional tensor is concatenated into the decoder to select the specific organ to segment
Gibson et al. (2018)	Segmentation of spleen, kidney, gallbladder, esophagus, liver, stomach, pancreas, and duodenum	90: 43 (TCIA) 47 (BTCV)	DenseVNet	Supervised	L2 regularization loss Dice loss	9-fold cross-validation	Nvidia Titan X Pascal	6 h	78.00% (DSC) 5.9 mm (HD95)	Implementation of DenseVNet with: cascaded dense feature stacks, V-network with downsampling and upsampling, dilated convolutions, map concatenation, and a spatial prior. Application to eight abdominal organs
Roth et al. (2018b)	Segmentation of artery, vein, liver, spleen, stomach, gallbladder, and pancreas	331 (Internal for training) 150 (external for testing)	3D UNet for both: Localization and Segmentation	Supervised	Weighted cross entropy loss	4-fold cross-validation	Nvidia GeForce GTX Titan X	2–3 days	Internal dataset: 63.10% (DSC) External dataset: 82.20% (DSC)	Application of cascaded networks for localization (coarse stage) and segmentation (fine stage)

of 22 anatomical structures. A two-stage organ attention network reached a DSC of 87.80% (Wang et al. 2019b; Park et al. 2020).

Generalization was investigated in a few studies. Li et al. (2022a) reported a DSC of 83.00% on the FLARE dataset, after training on the TCIA and BTCV. Shi et al. (2021) performed different generalizations, from BTCV to MSD, and from MSD to BTCV with a DSC of 69.50% and 68.70%, respectively. Likewise, Zhang et al. (2021b) assessed generalization from MSD to NIH, and from MSD to BTCV with a DSC of 77.33% and 81.37%, respectively. Roth et al. (2018b) evaluated generalization among private datasets, reporting a DSC of 82.20%.

5 Segmentation of pancreas tumors

The most common type of pancreatic cancer forms from exocrine cells, which make and move digestive enzymes. Pancreatic adenocarcinomas originate from exocrine cells that line the tube-like ducts of the pancreas and are also called PDACs (Ducreux et al. 2015; McGuigan et al. 2018). Pancreatic duct dilation is associated with a high risk of PDAC and other pancreatic cancers. A healthy pancreatic duct is normally almost invisible on abdominal CT scans due to its slender shape and small size. For this reason, its visibility in CTs could be a sign of the presence of pancreas cancers (Zou et al. 2023). Less often, pancreatic cancer forms from endocrine cells responsible for hormone production and are called PNETs (Ducreux et al. 2015; Rawla et al. 2019; Burns and Edil 2012). Pancreas cancer can originate from cystic lesions, which are fluid-filled sacs and are increasingly common incidental findings on abdominal imaging tests (Duh et al. 2023). Pancreatic cysts can be nonneoplastic and neoplastic. The latter include benign lesions, such as serous cystadenomas, mucinous cystic neoplasms, intraductal papillary mucinous neoplasm, and cystadenocarcinoma, with many subclasses existing due to their different characteristics (Beger et al. 2009; Adsay 2008; Khan et al. 2011; Duh et al. 2023). The head of the pancreas has a higher anatomical variability than the other two subregions, namely the body and tail. It is also the subregion where most (60%-70%) pancreas cancers occur (Sureka et al. 2021; Vareedayah et al. 2018).

Therefore, the accurate segmentation of pancreas tumors is essential for the clinical integration with quantitative imaging biomarkers which have shown promising results of early detection of pancreas tumors, and for precise 3D modeling for surgical and radiotherapy planning (Mukherjee et al. 2023).

5.1 Variability of tumors size and location

Following a similar approach described for parenchyma segmentation (cfr. Section 4), a registration was performed on 281 CTs of the MSD dataset, with case #29 as reference. The result is depicted in Fig. 11, showing the broad spatial distribution and frequency of pancreas tumors within the MSD dataset.

Table 10 Reviewed studies on the segmentation of pancreas parenchyma with attention applied to CNN and UNet on multi-organ annotations

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Irshad et al. (2023)	Segmentation of pancreas	43 (TCIA) 47 (BTCV)	UNet UNet++ Attn-UNet Two topologies: sharing encoder and decoder, but task-specific final layers; or sharing endoder with task-specific decoders	Supervised	Dice loss Binary cross entropy	67% training, 10% validation, 23% test	Nvidia Tesla P100	–	TCIA: UNet: 72.30% (DSC) UNet++: 62.10% (DSC) Attn-UNet: 72.40% (DSC) UNet: 0.9111 mm (HD) UNet++: 1.718 mm (HD) Attn-UNet: 0.928 mm (HD) BTCV: UNet: 68.20% (DSC) UNet++: 62.50% (DSC) Attn-UNet: 67.40% (DSC) UNet: 1.741 mm (HD) UNet++: 1.842 mm (HD) Attn-UNet: 1.377 mm (HD)	Multi-task network for: region segmentation and edge prediction. Comparison among UNet, UNet++, and Attn-UNet. Code available at: https://github.com/samra-irshad/ 3d-boundary-constrained-networks
Tong et al. (2023)	Segmentation of liver, kidney, spleen, and pancreas	511 (FLARE)	Encoder-Decoder (Localization) ResUNet and Multi-scale Attention (Segmentation)	Supervised	Dice loss (Localization) Dice loss Mean square error (Segmentation)	70% training, 10% validation, 20% test	Nvidia GeForce GTX 1080 Ti	50 h	59.10% (DSC) 42.20% (NSD)	Coarse stage for localization. Fine stage with multi-scale attention to segment pancreas, liver, spleen, and kidney.

Table 10 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Yuan et al. (2023)	Segmentation of aorta, gallbladder, kidneys, liver, pancreas, spleen, and pancreas	30 (Synapse)	UNet-like with: Gated recurrent units for skip connections + Gated-dual attention (Multi-scale weighted channel attention + Transformer self attention)	Supervised	–	60% training, 40% test	Nvidia GeForce RTX 3080 Ti	–	62.77% (DSC)	Gate recurrent units integrated into skip connections to reduce the semantic gap between low and high-level features. Gated-dual attention to capture information on small organs and global context. Code available at: https://github.com/DAgalaxy/ MGB-Net
Mx et al. (2022b)	Segmentation of pancreas	82 (NIH) 47 (BTCV)	DenseNet121 with: Improved Refinement Residual Block and Channel Attention Block (smooth network) + Shared bottleneck attention module + Improved Refinement Residual Block (Border network)	Supervised	Dice loss Focal loss	4-fold cross-validation	Nvidia Tesla P100	–	NIH: 82.82% (DSC) 71.13% (Jaccard) 83.16% (Precision) 83.54% (Recall) BTCV: 79.34% (DSC) 66.02% (Jaccard) 1.15 mm (ASD)	Discriminative feature attention module to address intra-class inconsistency and inter-class indistinction. Attention to avoid a module for localization. Improved refinement residual block to highlight spatial positions and aggregate contextual information
Tong et al. (2020)	Multi-organ Segmentation	90: 43 (TCIA) 47 (BTCV)	Encoder-Decoder with dual attention: Squeeze and Excitation (Channel attention) Convolutional layer (Spatial attention)	Supervised	–	9-fold cross-validation	Nvidia Tesla	–	79.24% (DSC) 1.82 mm (ASD)	A self-paced learning strategy for the multi-organ segmentation to adaptively adjust the weight of each class

Table 11 Reviewed studies on the segmentation of pancreas parenchyma with hybrid transformers on multi-organ annotations

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Huang et al. (2023)	Segmentation of aorta, gallbladder, spleen, kidneys, liver, pancreas, spleen, stomach, heart, and retina	30 (Synapse) 100 MRI (Automated cardiac diagnosis challenge) 40 (Digital Retinal Images for Vessel Extraction)	Encoder-decoder with: transformer blocks in all encoding and decoding steps + Transformer context bridge between encoder and decoder (fusion of multi-scale information)	Supervised	–	5-fold cross-validation	Nvidia GeForce RTX 3090	–	Synapse: 65.67% (DSC)	Hierarchical encoder-decoder with ReMix-FFN module in each transformer block with a convolution and a skip connection between the two fully connected layers to capture local information in addition to global dependencies. Features of different scale as output of each encoder step are concatenated, and sent to ReMixed transformer context bridge with self-attention to capture global dependencies. The output features are split into different scale feature maps and sent to ReMix-FFN modules of the decoder to mix global dependencies with local context. Code available at: https://github.com/ZhifangDeng/ MISSFormer
Zhao et al. (2022b)	Segmentation of aorta, gallbladder, kidneys, liver, pancreas, spleen, stomach, and hearth	30 (Synapse) 100 MRI (Automated cardiac diagnosis challenge)	UNet-like with: Encoder: ResNet-50 + Progressive sampling module + Vision Transformer (Hybrid CNN-Transfomer)	Supervised	Cross entropy loss Dice loss	60% training, 40% test	Nvidia Titan RTX	–	Synapse: 59.84% (DSC)	A progressive sampling module to ensure that highly relevant regions of the organ are in the same patch

5.2 Studies on tumors

A total of 21 out of the 130 reviewed studies pertain to segmentation of pancreas tumors, with eight based on UNet architectures (Table 13), nine on attention applied to CNNs and UNet (Table 14), two on hybrid transformers (Table 15), and two on GANs (Table 16).

All the reviewed studies on tumors were based on supervised learning, except two that concerned semi-supervised learning (Section 3.3 of the Appendix). Three studies proposed a two-stage (coarse-fine) approach for segmentation, while two a three-stage method. They are described in Section 3.1 of the Appendix. Three works proposed the design of loss functions (Section 3.4 of the Appendix.)

MSD was the most adopted dataset (13 out of 21 reviewed works or 61.9%) since it includes annotation of pancreas tumors (Sect. 3.6). Some studies employed only the annotated 281 CTs of the MSD dataset, while others the whole 420 scans of MSD, including 139 unlabeled CTs. For this reason, the performances of the reviewed studies on MSD varied greatly. In contrast with the reviewed studies on parenchyma, there was more use of private (internal and external) datasets as alternative sources of data for the segmentation of cancer. As a consequence, a comparison among these studies was not possible. Furthermore, there was a heterogeneity in terms of the type of tumors. In fact, six works concerned segmentation of both parenchyma and tumors (without further information on the type of cancer), four did not provide information on the type, three were on PDAC, one on pancreas neuroendocrine tumor, one on five types of cancer (including PDAC), one on parenchyma and 11 types of cancer, one on parenchyma and two types of cancer, one on tumor and surrounding vessels, one on dilated pancreas duct and five types of cancer. The code was publicly available for three out of 21 studies.

UNet-based models reported the highest DSC score on the portion of 281 annotated CTs of the MSD dataset. UNet coupled with channel attention and multi-scale convolutions achieved 80.12% of DSC. The multi-scale convolutions were embedded in each layer of the encoder to extract semantic information at different scales to localize small or very small tumors, and inserted also in the decoder layers (Du et al. 2023). This configuration outperformed by a large margin a two-stage model, with UNet for both localization and segmentation, reaching a DSC of 63.36% (Ju et al. 2023). This model integrated a spatial visual cue fusion module, based on the conditional random field to learn the global context, and an active localization offset module to adjust dynamically the localization results during the coarse stage (Ju et al. 2023). When considering all 420 CTs of the MSD dataset, the highest DSC score (51.83%) was achieved by a hybrid transformer. This model consisted of a Swin-transformer as an encoder with two modules, the first as an auxiliary block for boundary extraction to obtain rich and discriminative feature representation, and the second to preserve the pancreas boundary. The decoder was a CNN (He and Xu 2023). When used for generalization on MSD, nnU-Net reported a DSC of 82.00% on a portion of 152 cases of MSD after training on the dataset curated at Mayo Clinic of 921 CTs, described by Panda et al. (2021) (Mukherjee et al. 2023). Concerning semi-supervised learning, one study based on GAN achieved a DSC of 52.90% on 282 cases of MSD (Chaitanya et al. 2021). Overall, the DSC score for tumor segmentation on the MSD dataset dropped if compared to the results for parenchyma on the same dataset, underlining the further complexity due to the particularly small size of pancreas tumors (cfr. Section 1, Sect. 4.2, and Sect. 4.3).

Table 12 Reviewed studies on the segmentation of pancreas parenchyma with GAN on multi-organ annotations

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Francis et al. (2023)	Segmentation of liver, kidneys, spleen, and pancreas	1,112 (AbdomenCT-1k)	Conditional GAN: Dilated UNet + Attention gate (Generator) Fully Convolutional Network (Discriminator)	Supervised	Adversarial loss	70% training, 10% validation, 20% test	4 Nvidia A100	–	86.10% (DSC) 6.65 mm (HD95) 86.80% (Precision) 86.60% (Recall)	Residual dilated convolution block and spatial pyramid pooling replacing convolutions and max pooling in UNet. Attention gate inserted into skip connections

One study assessed segmentation on pancreas subregions combining a Bayes model and UNet on two datasets of precancerous and cancerous lesions, each consisting of 15 CTs. The model achieved a DSC of 97.0% and 89.3% for the head, 95.0% and 90.1% for the body, 94.3% and 90.6% for the tail on the datasets with precancerous and cancerous annotation, respectively (Javed et al. 2022).

6 Segmentation of pancreas cysts

Three studies addressed the segmentation of pancreas cysts. They all used internal private datasets. One of them assessed also performances on parenchyma using the NIH dataset (Table 17, and Section 4 of the Appendix). The code was publicly available for one study. UNet with ASPP and spatial pyramid pooling (Li et al. 2023b) reported a DSC of 84.53% on a dataset of 107 cases of cysts, while VGGNet a DSC of 83.31% on 131 CTs for pancreatitis (Xie et al. 2020).

7 Segmentation of pancreas inflammations

Acute pancreatitis, an inflammation of the pancreas, is the leading cause of hospital admission for gastrointestinal disorders in the United States and several other countries (Deng et al. 2023). The segmentation of an inflamed pancreas is more challenging than the normal pancreas since it invades the surrounding organs causing blurry boundaries, and it has higher shape, size, and location variability than the normal pancreas (Deng et al. 2023).

Two of the reviewed studies concerned the segmentation of pancreatitis (Table 18). Deng et al. (2023) performed the first study on the segmentation of acute pancreatitis on an internal dataset of 89 CTs. An FCN with a region proposal was used for the detection of pancreatitis region. The detected region was cropped and sent to the 2D U-Net for segmentation (Deng et al. 2023). Guo et al. (2022b) adopted UNet++ to segment chronic inflammation of the common bile duct in pediatric patients. A ResUNet network was then used to classify the degree of severity of inflammation.

8 Discussion

8.1 Main findings

In this systematic review, we analyzed the published literature, consisting of 130 original studies, on DL for the segmentation of parenchyma, tumors, cysts, and inflammation of the pancreas. By looking at the geographical origin of the reviewed studies, China is leading the ranking with more than half of the published articles in peer-reviewed journals, ahead of the United States, UK, Japan, and Canada. Unexpectedly, there are countries with an established tradition in pancreatic surgery, like Italy, not present in this ranking, underlying a research gap from a technical point of view with the others (Abu Hilal et al. 2023). We observe that few studies of out the reviewed ones were led by clinicians, with seven works published in journals in the medical field (Bagheri et al. 2020; Guo et al. 2022b; Li et al.

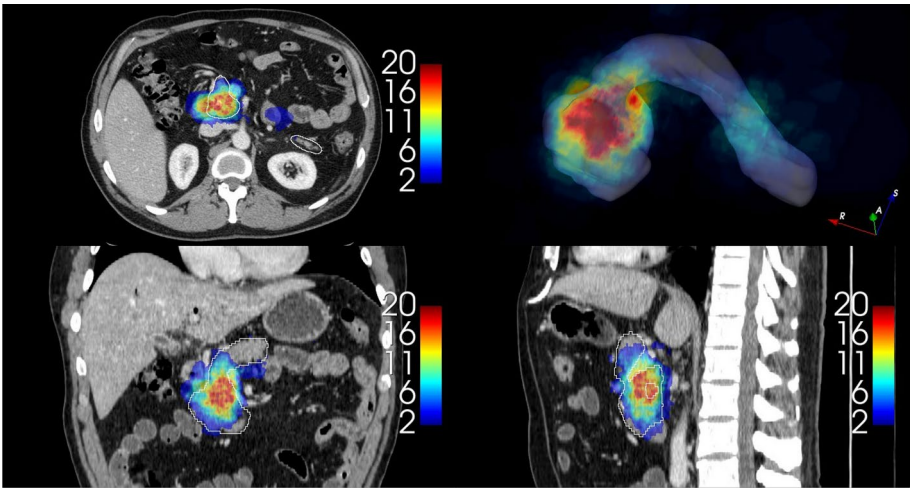

Fig. 11 Spatial distribution and frequency of pancreas tumors within the MSD dataset with 281 cases and case #29 as a reference in the image (Simpson et al. (2019)): most frequent pancreases in the dataset in red, least frequent ones in blue. Boundary of case #29 in white. (Color figure online)

2023b; Mukherjee et al. 2023; Park et al. 2020; Si et al. 2021; Sundar et al. 2022). In contrast, the rest of the studies were published in technical journals or cross-disciplinary ones at the boundary between medicine and computer science. The trend in the number of published studies is constantly growing, thus reflecting an increase in interest in the community. There are fewer studies on tumor segmentation as they are more challenging to segment than parenchyma. DL segmentation on other tiny structures like the dilated pancreatic duct and surrounding vessels has been only recently proposed, with initial studies published in 2022 (Mahmoudi et al. 2022; Shen et al. 2022; Zou et al. 2023). Only one study concerned segmentation of pancreas subregions on tumors (Javed et al. 2022).

Our review highlights an enormous variety of DL architectures specifically designed for pancreas segmentation from standard UNet to transformers up to hybrid transformers. Likewise, many attention blocks have been designed from attention gate to SE up to reverse attention (Oktay et al. 2018; Zhou et al. 2023).

Almost all the studies used region-based metrics (e.g., DSC, and Jaccard index). Only 20 out of 104 on parenchyma, and five out of 26 on tumor and cysts segmentation used a boundary-based metric like HD. Two works used NSD as a boundary-based metric (Ma et al. 2022b; Tong et al. 2023). Dice loss function was used in most studies to mitigate the class imbalance issue with the background as the prominent class, followed by parenchyma of the pancreas, and tumor as the least present. As reported in Section 1.4 and Section 3.4 of the Appendix, several studies proposed the design of new loss functions to improve metrics results.

The present review has shown that a UNet configuration with residual blocks in the encoder and a decoder with spatial and channel attention obtained the highest DSC score (91.37%) on the NIH dataset for pancreas segmentation, outperforming transformers-based models and other novel architectures (Shan and Yan 2021). In contrast, a two-stage transformer, with a UNet for localization and ViT for segmentation, reported the highest DSC (91.22%) on the MSD dataset (Dai et al. 2023). The performances on datasets with multi-

Table 13 Reviewed studies on the segmentation of pancreas tumors with UNet architectures

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Cao and Li (2024)	Parenchyma and tumors	82 (NIH) 281 (MSD)	UNet with: High resolution spatial information recovery + Multi-scale high resolution pre-segmented feature fusion + Pyramid multi-scale feature perception and fusion	Supervised	Difficulty-guided adaptive boundary-aware loss	4-fold cross-validation	Nvidia GeForce RTX 3070	–	Parenchyma (NIH): 88.96% (DSC) 89.27% (Precision) 89.98% (Recall) Parenchyma (MSD): 89.52% (DSC) 93.19% (Precision) 88.71% (Recall) Tumors (MSD): 54.38% (DSC) 69.58% (Precision) 53.17% (Recall)	High-resolution spatial information recovery module: encoder and decoder features of the same layer are sent to high resolution spatial information filtering module to extract high-resolution pre-segmented images, which are then fused. Multi-scale high-resolution pre-segmented feature fusion module: features of the encoder and decoder finely processed into a high-resolution pre-segmented feature map. Pyramid multi-scale feature perception and fusion module uses the extracted pre-segmented images to guide the network to focus on the dimensional changes of the segmented targets. Design of Difficulty-guided adaptive boundary-aware loss function to address the class imbalance and improve segmentation of uncertain boundaries

Table 13 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Ju et al. (2023)	Parenchyma and tumors	82 (NIH) 281 (MSD)	UNet: Spatial visual cue fusion + Active localization offset (Localization) UNet (Segmentation)	Supervised	Dice loss Binary cross entropy loss	4-fold cross-validation	Nvidia GeForce RTX 3080	–	Parenchyma (NIH): 85.15% (DSC) Tumor (MSD): 63.36% (DSC)	Spatial visual cue fusion, based on conditional random field, learns global spatial context. It combines the correlations between all pixels in the image to optimize the rough and uncertain pixel prediction during the coarse stage. Active localization offset adjusts dynamically the localization results during the coarse stage. Code available at https://github.com/PinkGhost0812/SANet
Mukherjee et al. (2023)	Pancreas ductal adenocarcinoma	1,151: Mayo Clinic for training 152 (MSD) for test 41 (TCIA) for test	3D nnU-Net	Supervised	Dice loss Cross entropy loss	80% training, 20% test	Nvidia A100	80 h	Overall: 84.00% (DSC) 4.6 mm (HD) Generalization on MSD: 82.00% (DSC) 2.6 mm (HD) Generalization on TCIA: 84.00% (DSC) 4.30 mm (HD)	Bounding boxes by cropping the CT images to a 3D bounding box centered around the tumor mask. nnU-Net applied to bounding boxes
Ni et al. (2023)	Recurrence of pancreas ductal adenocarcinoma after surgery	205 (Internal) 64 (For recurrence prediction with radiomics)	AX-UNet with Atrous spatial pyramid pooling	Supervised	–	4-fold cross-validation	Nvidia Tesla V100	–	85.90% (DSC) 74.20% (Jaccard) 89.70% (Precision) 87.60% (Recall)	AX-UNet combining UNet and atrous spatial pyramid pooling. Code available at: github.com/ zhangyuhong02/ AX-Unet

Table 13 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Wang et al. (2023)	Tumors	93 (Shanghai Changhai Hospital) dataset of head and neck	3D UNet-like: Encoder: Multi-modal fusion downsampling block Decoder: Multi-modal mutual calibration block using attention	Supervised	Dice loss	3-fold cross-validation	Nvidia GeForce RTX 2080 Ti	–	76.20% (DSC) 63.08% (Jaccard) 6.84 mm (HD) 75.96% (Precision) 84.26% (Recall)	Multi-modal fusion downsampling block to fuse semantic information from PET and CT, and to preserve unique features of different modal images. Multi-modal mutual calibration block to calibrate different scale semantics of one modal images guided by attention maps from the other modal images
Javed et al. (2022)	Segmentation of pancreas	82 (NIH) 15 (pre-cancer) 15 (cancer)	Bayes model + UNet	Supervised	Dice loss Focal loss	4-fold cross-validation	Nvidia GeForce RTX 2080 Ti	8 h	NIH: Head: 96.1% Body: 93.8% Tail: 92.9% Pre-cancer: Head: 97.0% Body: 95.0% Tail: 94.3% Cancer: Head: 89.3% Body: 90.1% Tail: 90.6%)	First study on segmentation of pancreas subregions (head, body, and tail). The probability map of UNet is updated with a probability map of a Bayes model indicating the three subregions.
Huang et al. (2021a)	Pancreatic neuroendocrine neoplasms	98: First Affiliated Hospital of Sun Yat-Sen University Cancer Center of Sun Yat-Sen University) 72 for recurrence: from both above centers	UNet	Supervised	Cross entropy loss	10-fold cross-validation	Nvidia GeForce RTX 1080 Ti	10 h	First dataset: 81.80% (DSC) 83.60% (Precision) 81.40% (Recall) Second dataset: 74.80% (DSC) 87.20% (Precision) 68.60% (Recall)	Semi-automatic segmentation. A radiologists identified tumors by drawing bounding boxes to delineate region of interest sent as input to UNet. Radiomic analysis to predict pathohistologic grading

Table 13 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Si et al. (2021)	Pancreatic ductal adenocarcinoma and other four types of cancers	319 (Second Affiliated Hospital Shanghai) for training 347 (First and Second Affiliated Hospital Shanghai) for generalization	ResNet18 (Localization) UNet32 (Segmentation)	Supervised	Cross entropy loss	See previous table cell on datasets for the study	Nvidia GeForce GTX 1050	–	Generalization: 83.70% (DSC)	Three different networks used for pancreas location, segmentation, and diagnosis (presence of tumors)

Table 14 Reviewed studies on the segmentation of pancreas tumors with attention mechanism applied to UNet and CNNs architectures

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Cao et al. (2023b)	Parenchyma and tumors	82 (NIH) 420 (MSD)	UNet with three attention mechanisms on skip connections: Spatial + Channel + Multi-dimensional features	Supervised	Weighted cross entropy loss	96% training, 4% test	Nvidia GeForce GTX 1060	2 h	Parenchyma (NIH): 83.04% (DSC) 81.71% (Precision) 84.42% (Recall) Parenchyma (MSD): 83.39% (DSC) 85.51% (Precision) 81.37% (Recall) Tumors (MSD): 40.15% (DSC) 52.32% (Precision) 35.29% (Recall)	Design of a loss function to capture edge details of pancreas and tumors. Multi-dimensional attention gate integrated into skip connections for small target feature localization in multiple dimensions of space and channels, and for filtering redundant information in shallow feature maps, thus enhancing the feature representation of the pancreas and pancreatic tumor

Table 14 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Du et al. (2023)	Tumors	55 (Qingdao University Hospital) 281 (MSD)	UNet with multi-scale channel attention	Supervised	Binary cross entropy	5-fold cross-validation	Nvidia Tesla T4	–	Qingdao: 68.03% (DSC) 59.31% (Jaccard) 12.04 mm (HD) MSD: 80.12% (DSC) 74.17 (Jaccard) 2.26 mm (HD)	Integration of multi-scale convolutions and channel attention into each encoder and decoder block
Li et al. (2023f)	Parenchyma and tumors	281 (MSD)	nnU-Net with attention + Balance temperature loss + Rigid temperature optimizer + Soft temperature indicator	Supervised	Balance temperature loss	5-fold cross-validation	Nvidia Tesla P100	–	Parenchyma: 85.06% (DSC) Tumors: 59.16% (DSC)	Segmentation of both pancreas and tumors. Balance temperature loss to dynamically adjust weights between tumors and the pancreas. Rigid temperature optimizer to avoid local optima. Soft temperature indicator to optimize the learning rate

Table 14 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Zhou et al. (2023)	Parenchyma and two types of pancreas cancers	116 (Shanghai Changhai Hospital) 42 normal pancreas (Internal)	Dual branch encoder-decoder (Pancreas segmentation) Encoder-decoder: contrast enhancement block + reverse attention block (Tumor segmentation)	Supervised	Dice loss	2-fold cross-validation	Nvidia Tesla K40	36 h	Abnormal: 78.72% (Jaccard) 89.07% (Precision) 87.42% (Recall) Normal: 87.74% (Jaccard) 91.47% (Precision) 95.50% (Recall)	Dual branch encoder combining semantic information extraction and detailed information extraction. Aggregation of feature maps of the two branches. Decoder to segment pancreas. Enhancement encoder-decoder network to improve segmentation accuracy of pancreatic tumors. Contrast enhancement block after each encoding step to extract the edge detail information. Reverse attention block inverting the decoder feature to guide the extraction of effective information in the encoder to generate an accurate prediction map

Table 14 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Zou et al. (2023)	Dilated pancreatic duct and five types of pancreas cancers	150 (Nanjing Drum Tower Hospital, (Internal)) 40 (Jiangsu Province Hospital of Chinese Medicine) for generalization	3D nnU-Net for: (Localization) Terminal anatomy attention module (Segmentation) Terminal distraction attention module (Refine stage)	Supervised	Terminal Dice loss	5-fold cross-validation	–	–	Internal: 84.17% (DSC) 11.11 mm (HD) Generalization: 82.58% (DSC)	First work on errors on terminal regions of the dilated pancreatic duct. Terminal anatomy attention module to learn the local intensity from the terminal CT images, feature cues from the coarse predictions, and global anatomy information. Terminal distraction attention module to reduce false positive and false negative cases. Design of terminal Dice loss for segmentation of tubular structures

Table 14 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Mahmoudi et al. (2022)	Tumors and surrounding vessels	138 (MSD) 19 (Internal, with vessellabels, (for fine-tuning))	3D local binary pattern (Localization) Ensemble of: Attention gate + Texture Attention block (Scale invariant feature transform and local binary pattern) (Segmentation)	Supervised	Generalized Dice loss Weighted Pixelwise Cross entropy loss Boundary loss	–	Nvidia GeForce GTX 1080 Ti	–	Tumor: 60.60% (DSC) 3.73 mm (HD95) 57.80% (Precision) 78.00% (Recall) Superior mesenteric artery: 81.0% (DSC) 2.89 mm (HD95) 76.00% (Precision) 87.00% (Recall) Superior mesenteric vein: 73.00% (DSC) 3.45 mm (HD95) 68.00% (Precision) 81.00% (Recall)	Design of texture attention block with scale invariant feature transform or local binary pattern to provide a comprehensive representation of pathological tissue. Integration of attention gate and texture attention gate into skip connections of texture attention UNet. Use of a 3D CNN as an ensemble of attention UNet and texture attention UNet. Design of Generalized Dice loss, Weighted Pixelwise Cross entropy loss, and Boundary loss to address unbalanced data, and boundary between pancreas and tumors

Table 14 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Shen et al. (2022)	Dilated pancreatic duct	82 (NIH) for localization 30 (Internal) for segmentation	3D UNet (Localization) 3D UNet + Squeeze and excitation (Segmentation)	Supervised	Dice loss Focal loss	NIH: 58% training, 20% validation, 22% test Internal: 5-fold cross-validation	Nvidia Quadro P6000		NIH: 75.9% (DSC) 72.4% (Recall) Internal: 49.90% (DSC) 51.90% (Recall)	First study on automated 3D segmentation of dilated pancreatic duct. Generation of an annotated dataset on dilated pancreatic duct. Attention block with squeeze and excitation inserted into the bottleneck of a 3D UNet
Wang et al. (2021b)	Pancreatic ductal adenocarcinoma	800 (John Hopkins) 281 (MSD) for generalization	UNet with Inductive attention guidance	Semi- supervised	Cross entropy loss	4-fold cross-validation	Nvidia Titan RTX	–	John Hopkins: 60.29% (DSC) 99.75% (Recall) MSD: 32.49% (DSC)	Attention guided framework for classification and segmentation with partially labeled data (few annotated images for segmentation). Training using multiple instance learning with cancer and background regions as bags instead of per-voxel pseudo labels as in typical semi-supervised learning

Table 14 (continued)

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Turečková et al. (2020)	Parenchyma and tumors	420 (MSD) datasets of kidney tumors, and liver tumors	UNet and VNet with Attention gate in skip connections	Supervised	Dice loss Cross entropy loss	5-fold cross-validation	–	–	Parenchyma (UNet): 81.81% (DSC) 81.21% (Precision) 84.51% (Recall) Tumors (UNet): 52.68%(DSC) 62.98% (Precision) 55.84% (Recall) Parenchyma (VNet): 81.22% (DSC) 80.61% (Precision) 84.10% (Recall) Tumors (VNet): 52.99%(DSC) 64.62% (Precision) 54.39% (Recall)	Attention gate integrated into skip connections for segmentation of pancreatic tumors

organ annotations were lower compared with those with only annotations of the pancreas, due to the weak boundaries among the organs on CT scans (Wang et al. 2019b). The score on segmentation of smaller lesions, like tumors, on specific datasets like MSD is much lower. Overall, we could not decree which models are the most suitable for pancreas parenchyma or tumor segmentation, since the data split for training, validation, and test sets were different among the studies on the same dataset.

8.2 Challenges and future directions

8.2.1 Clinical need perspective

Any surgical intervention involving the pancreas is complex and requires a deep understanding of individual anatomy due to its position and the high level of interaction with surrounding vessels and structures (Russell and Aroori 2022). Surgical resection of the area affected by the tumor is the only effective treatment for pancreatic neoplasms such as PDAC (Wei and Hackert 2021) and PNET (Johnston et al. 2020), and its outcomes are doubtful due to intraoperative or postoperative complications such as recurrences or infections. For these reasons, the National Comprehensive Cancer Network® (NCCN®) recommended following a specific pancreatic CT protocol for preoperative planning to avoid and minimize adverse surgical outcomes. By carefully assessing the patient's medical history, imaging, and laboratory results, surgeons can better characterize the tumor, making preoperative planning more effective. Three-dimensional visualization of the pancreas parenchyma, the tumor lesion, and the vessels is of fundamental importance for several reasons. The virtual inspection of accurate 3D reconstructions allows the surgeon to make critical decisions on the intervention feasibility and the choice among different surgical procedures (pancreatoduodenectomy, distal pancreatectomy, central pancreatectomy) better than using the flattened 2D visualization. Furthermore, a tridimensional view of the regions of interest eases the comprehension of anatomical variants and vascular involvement, possibly leading to a different and more challenging surgical procedure for the same pathology. The high risk associated with pancreatic operations makes them subject to significant morbidity and mortality, and proper pre-operative design helps identify risk, reduce complications, and prepare for potential intraoperative challenges such as bleeding Zhang et al. (2023b). Pancreatic fistula, delayed gastric emptying, diabetes, or recurrence are the undesired outcomes of pancreatic surgery. Therefore, segmentation has been proposed to improve the visualization of the pancreas and its lesions to support clinicians during diagnosis, preoperative planning, and disease progress monitoring (e.g. during radiotherapy treatment). Unfortunately, the pancreas has been traditionally regarded as one of the toughest abdominal organs for the segmentation task due to its small volume compared with a full CT scan, blurred boundaries, and large variations among patients in terms of shape and position. DL methods are no exception, as highlighted by the present systematic review. Moreover, DL-based segmentation of pancreas cancer is a relatively new research topic since the first published studies date back to 2020 (Xie et al. 2020; Turečková et al. 2020).

Future directions: Future research efforts should be pushed towards the segmentation of pancreas tumors, especially at the multi-class level, given their clinical challenges from diagnosis to surgical treatment. Research on precancerous lesions, e.g., pancreatic intraepithelial neoplasms, is also encouraged. Another area of investigation concerns the assess-

Table 15 Reviewed studies on the segmentation of pancreas tumors with hybrid transformers architectures

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
He and Xu (2023)	Parenchyma and tumors	420 (MSD) spleen MSD dataset of knee	Hybrid CNN-Transformer Encoder: (3D Swin-Transformer + boundary extracting module) + Boundary preserving module + Decoder: CNN	Supervised	Dice loss Cross entropy loss	67% training and validation, 33% test	Nvidia GeForce RTX 3090 Ti	–	Parenchyma: 81.47% (DSC) 1.77 mm (ASSD) Tumor: 51.83% (DSC) 17.13 mm (ASSD)	Application of boundary awareness into 3D CNN and transformers. Swin-transformer as encoder and auxiliary boundary extracting module to obtain rich and discriminative feature representations. Boundary preserving module to fuse boundary map and features from the encoder
Qu et al. (2023)	Parenchyma and 11 types of pancreas cancers	313 (Peking Union Medical College Hospital) for training and test 53 (Guandong General Hospital) for generalization 50 (Jingling Hospital) for generalization MSD (420) for training and test	Swin Transformer and 3D CNN (Based on M3NET) Feature alignment: Transformer guided fusion + Cross-network attention (Decoder)	Supervised	Weighted cross entropy loss	First dataset: 63% training, 9% validation, 28% test. MSD: 67% training and validation, 33% test			Pancreas: Peking: 92.51% (DSC) Guangdong: 89.56% (DSC) Jingling: 88.07% (DSC) MSD: 85.71% (DSC) Tumors: Peking: 80.51% (DSC) Guangdong: 67.17% (DSC) Jingling: 69.25% (DSC) MSD: 43.86% (DSC)	CNN and transformer branches perform separate feature extraction in the encoder. Progressive fusion between CNN and transformer in the decoder. Transformer guidance flow to address the inconsistency of the feature resolution and channel numbers between the CNN and transformer branches. Cross network attention into CNN decoder to enhance fusion capability with the transformer

ment of DL models on the pancreas subregions, especially the head, which is the subregion where most pancreas tumors occur. Our review included only one study on this application (Javed et al. 2022).

8.2.2 DL models perspective

CNNs became popular in the 2010s for computer vision tasks (Liu et al. 2022c). nnU-Net was proposed at the end of that decade and was tested on 23 public datasets (including MSD for the pancreas). Remarkably, nnU-Net set new benchmarks in several applications by virtue of automatic configuration capable of optimizing preprocessing, network architecture, training, and post-processing. nnU-Net highlighted the importance of method configuration over architectural variations, showing that details in configuration have more impact on performance than DL models (Isensee et al. 2020). In a recent study, the original nnU-Net architecture outperformed Auto3DSeg, a novel framework part of the MONAI[2] ecosystem with autoconfiguration, in the segmentation task on 15 organs of the AMOS dataset (Isensee et al. 2024).

Some of the reviewed studies in the present work exploited the autoconfiguration adaptability of nnU-Net to AdbomenCT-1k, BTCV, and MSD datasets. Unfortunately, the studies on the other architectures did not provide details on configuration. The proposed models in the reviewed studies were assessed on a few datasets, mostly one or two, in contrast with the original nnU-Net. In several cases, the models were assessed on small test sets that introduce result instability and question the significance of minor performance gains (Isensee et al. 2024). In this regard, 5-fold cross-validation could improve reliability, thus representing a pragmatic solution (Isensee et al. 2024).

In 2020, the advent of ViT altered the landscape of DL (Liu et al. 2022c). Later, hybrid transformers were introduced in computer vision tasks to combine the strengths of CNNs and transformers, which are inductive bias and the capability to process long-range sequences, respectively (Liu et al. 2022c). It has been largely believed that the effectiveness of hybrid transformers was related to the superiority of transformers, rather than the inherent inductive biases of convolution (Liu et al. 2022c).

To retain the inherent inductive bias of convolutions while taking advantage of transformers, the ConvNeXt architecture was recently proposed by integrating into CNNs several key design components of transformers, e.g., inverted bottleneck and large kernel sizes (Liu et al. 2022c). More recently, MedNeXt, a 3D UNet-like model based on ConvNeXt, was proposed as the first ConvNeXt architecture for medical image segmentation (Roy et al. 2023). It was tested on the 15 organs of the AMOS dataset reporting a mean DSC of 91.77%, outperforming the original nnU-Net, transformers, and Auto3DSeg (Isensee et al. 2024).

The metrics like DSC and HD, used by the vast majority of the reviewed studies, should be carefully considered for the assessment of small structures like the pancreas and its lesions for several reasons. First, a few pixel differences between two predictions can have a large impact on DSC (Reinke et al. 2011). Second, by considering the same prediction, HD is better for low-resolution images while high-resolution is required to delineate accurately small anatomical structures like the pancreas and its lesions (Reinke et al. 2011). Third, oversegmentation leads to higher DSC than undersegmentation. Therefore, a model aiming

[2] https://monai.io/

Table 16 Reviewed studies on the segmentation of pancreas tumors with GAN architectures

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Li et al. (2022c)	Tumors	163 (Shanghai Jiao Tong University) 468 MRI (for style transfer) 281 (MSD) for generalization	CycleGAN-like for: Synthetic data from MRI (Style transfer) ResNet: Extraction of knowledge from MRI (Meta-learning I) + Integration with salient knowledge from CT (Meta-learning II)	Supervised	Adversarial loss Cycle consistency loss Dice loss	4-fold cross-validation	Nvidia GeForce RTX 2080 Ti	–	Shanghai Jiao Tong University: 64.12% (DSC) MSD: 57.62% (DSC)	First study on meta-learning from one to a different modality. Random style transfer on MRI: generation of synthetic images with continuously intermediate styles between MRI and CT to simulate domain shift. First meta-learning: the model learns the common knowledge of synthetic data, and provides pancreatic cancer-related prior knowledge for the target segmentation task. Second meta-learning: the model learns the salient knowledge of the CT data to enhance segmentation
Chaitanya et al. (2021)	Tumors	282 (MSD) datasets of other organs	GAN + UNet	Semi- supervised	Adversarial loss	variable percentage for training, validation, and test	–	–	52.90% (DSC)	Semi-supervised learning for data augmentation. Adversarial term to help two generators synthesize diverse set of shape and intensity variations present in the population, even in scenarios where the number of labeled examples are extremely low. Code available at: https://github.com/krishnabits001/task_driven_data_augmentation

to maximize DSC may tend to oversegment, increasing the risk of resection of healthy tissue (Ansari et al. 2022).

The reviewed studies did not unveil details on the size of their models. Therefore, claims of superiority may not be objective if comparing larger to smaller models. A recent study on segmentation on the AMOS dataset has demonstrated a significant boost in performance with larger models (Isensee et al. 2024). Although this research reported results on the mean of the 15 organs and not specifically on each single anatomical structure, we may assume that this rule holds true also for pancreas datasets (Isensee et al. 2024).

An increasing number of researchers opt for releasing publicly their code (Chen et al. 2022c). This is certainly beneficial for the community since it encourages researchers and clinicians to become more familiar with them. Overall, we found 19 studies with public code on GitHub. However, to facilitate reproducibility readers of publications should check carefully how samples from the datasets were selected (Chen et al. 2022c).

Future directions: Future developments should concern models with a described configuration, and compared against a well-configured baseline. In addition, those model claiming their superiority should not be limited to one single dataset. In particular, separating development datasets and independent test datasets for cross-validation against baselines would offer a more reliable assessment of performance (Isensee et al. 2024). Moreover, new models should be designed to improve the segmentation of tumors and to increase the adaptability of the DL models to changes in the size and shapes of the lesions over time. More effective metrics for surface and boundary overlapping should also be developed (Ansari et al. 2022). Finally, there is the need to evaluate emerging DL models like MedNeXt on segmentation of the pancreas and its lesions (Roy et al. 2023).

8.2.3 Foundation models perspective

Although this systematic review concerns DL, we cannot ignore the foray of foundation models in healthcare. Therefore it is worthwhile to report the latest progress. Thanks to the unsupervised pretraining on massive data and prompt techniques, foundation models can generalize to tasks different than those used for training (Kirillov et al. 2023). After the initial demonstration of the prowess of large language models to pass medical examinations, requested by national boards in different countries to obtain the license for clinical practice, multi-modal large language models combining text and images were proposed (Moglia et al. 2024). More recently, Segment Anything Model (SAM) was released as an open-source foundation model for image segmentation as the result of an initiative from a tech giant like Meta Corporation (Menlo Park, CA, United States) (Kirillov et al. 2023). It consists of a ViT as an image encoder, a prompt encoder, and a decoder fusing the outputs of the encoders (Kirillov et al. 2023). SAM generates segmentation masks after users select a point, draw a bounding box in the image, or use a text prompt (Kirillov et al. 2023). Although it was originally trained on a curated dataset of one billion of segmented objects in natural images, called SA-1B, it has quickly spawned interest in the medical imaging community. The zero-shot capabilities of SAM in medical imaging were first tested on 19 datasets, achieving an IoU around 70.0% for pancreas segmentation on the MSD dataset, when using bounding boxes as prompt (Mazurowski et al. 2023). However, models like SAM have shown limitations in segmenting medical images with weak boundaries and low contrast (Ma et al. 2024). For this reason, fine-tuned versions of SAM were proposed. SAMed was

Table 17 Reviewed studies on the segmentation of pancreas cysts

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Xie et al. (2020)	Parenchyma and pancreatic cysts	82 (NIH) 200 (John Hopkins: 11 abdominal organs and five blood vessels) 131 (John Hopkins: pancreatic cysts)	VGGNet with Hierarchical recurrent saliency transformation network between Localization and Segmentation	Supervised	Dice loss	4-fold cross-validation	Nvidia Tital X Pascal	–	NIH: 84.53% (DSC) Renal donors: 87.74% (DSC) Pancreatic cysts: 83.31% (DSC)	Saliency transformation module beteween first and second stage to transforms the segmentation probability map as spatial weights, iteratively, from the previous to the current iteration. Hierarchical version to segment first the pancreas and then the internal cysts. Code available at: https://github.com/198808xc/OrganSegRSTN
Li et al. (2023b)	Pancreatic cysts	107 (Internal)	UNet with: Atrous pyramid attention module + Spatial pyramid pooling module	Supervised	Dice loss Binary cross entropy loss	5-fold cross-validation	–	–	84.53% (DSC) 75.81% (Jaccard)	Atrous pyramid attention module and spatial pyramid pooling module inserted into bottleneck layer to extract features at different scales, and contextual spatial information, respectively
Duh et al. (2023)	Pancreatic cysts	136 (Internal)	UNet with Attention gate in skip connections	Supervised	Dice loss	68% training, 32% test	Nvidia Tesla T4	–	93.10% (Recall)	Attention gate integrated into skip connections for segmentation of pancreatic cysts

implemented by fine-tuning SAM on the Synapse dataset, with the image encoder fine-tuned with low rank adaptation (LoRA) (Hu et al. 2021a; Zhang and Liu 2023). It outperformed state-of-the-art DL models on the Synapse dataset, by achieving a DSC of 72.15% for pancreas segmentation vs. 65.67% of a hybrid transformer which is the highest DSC found in the reviewed studies (Zhang and Liu 2023; Huang et al. 2023). MedSAM, a recent fine-tuned version of SAM using bounding box prompts, underwent internal and external validation on 86 and 60 segmentation tasks, respectively (Ma et al. 2024). A dataset of 1.5 million medical images from 10 modalities was curated, including datasets with annotations of the pancreas, namely AbdomenCT-1k, AMOS-CT, and MSD for internal, and WORD for external validation. It outperformed specialist DL models like nnU-Net on pancreas tumor segmentation in terms of DSC (77.80% vs. 75.80%) while achieving a similar score on pancreas parenchyma (85.50% vs. 85.70%) during internal validation. In external validation, MedSAM scored better than nnU-Net on pancreas parenchyma (87.00% vs. 85.00%). Unfortunately there was no external validation on pancreas tumor segmentation (Ma et al. 2024). SAM-2 was recently introduced to segment both images and videos (Ravi et al. 2024). Initial works explored the capability of SAM 2 to segment 3D medical images by transferring its video segmentation capability, thus treating each slice as a frame (Zhang and Shen 2024). Based on SAM 2, Medical SAM 2 was recently proposed. Although tested only on BTCV dataset for pancreas segmentation, it outperformed DL models (e.g., nnU-Net, and transformers), SAM-based ones, and Med-SAM (Zhu et al. 2024). The impressive performances of MedSAM and Medical SAM 2 make it spontaneous to ask the following question: *"Did MedSAM and Medical SAM 2 put an end of the DL models developed in the last 10 years?"*. It seems premature to give an answer now for some reasons. First, although MedSAM and Medical SAM 2 were tested on a mixture of known public datasets, a comparison with the reviewed studies on the pancreas is not possible since the samples used for training and validation may be different. Second, a comparison is currently missing among MedSAM, Medical SAM 2 and hybrid transformers which reported the highest DSC on pancreas cancer segmentation.

Future directions: Foundation models can be combined with other models, thus foreshadowing new AI-based medical applications. For instance, the masks generated by MedSAM and Medical SAM 2 or other methods could be used by a classification model. Or, a detection model could provide the initial bounding boxes of anatomical structures used by MedSAM and Medical SAM 2 for subsequent segmentation (Mazurowski et al. 2023). We are of the view that foundation models can benefit from continuous technical advancements in DL. In this regard, it will be interesting to see in the future foundation models with an image encoder inspired by MedNeXt to replace the ViT used by the SAM methods for medical imaging. If the superior performances of MedNeXt over ViT are confirmed also in models like MedSAM and Medical SAM 2, foundation models may soon reach scores at the level or close to those requested by clinicians.

8.2.4 Datasets perspective

The reviewed DL models demonstrated improvements over the years on DSC, Jaccard, and HD metrics. However, the results may show a limited value that is not the marginal entity, as in most cases, of such improvement but the fact that the models were trained and tested on datasets of small size and mostly from a single center. More specifically, by looking at the

Table 18 Reviewed studies on the segmentation of pancreas inflammation

Author	Application	Dataset size (Name)	Model architecture	Learning strategy	Loss	Training, validation, test	GPU	Training time	Results	Main contributions
Deng et al. (2023)	Acute pancreatitis	89 (Internal)	FCN + Region proposal network (Detection) UNet (Segmentation)	Supervised	Focal loss Cross entropy loss L1 regression loss	4-fold cross-validation	Nvidia Tesla V100	–	66.82% (DSC)	FCN for detection of pancreatitis region. The detected region was cropped and sent to the 2D U-Net for segmentation. First study on segmentation on acute pancreatitis
Guo et al. (2022b)	Chronic inflammation of choledoch	76 (Internal)	UNet++	Supervised	Binary cross entropy loss	5-fold cross-validation	Nvidia GeForce RTX 2080 Ti	–	83.90% (DSC)	UNet++ to segment chronic inflammation of cholecdoch in pediatric patients. Then ResUNet is used to classify the degree of severity of inflammation

tables summarizing the studies one remark that stands out is that in the vast majority of studies, the DL models are trained and tested almost exclusively on publicly available datasets, suggesting that there are difficulties in curating internal datasets. In contrast, some institutions were capable of collecting large datasets, e.g. with 1917 CTs from Mayo Clinic and 1150 CTs from John Hopkins Medical Institution for the segmentation of parenchyma and pancreas ductal adenocarcinoma, respectively (Panda et al. 2021; Wang et al. 2021b). Other datasets, like AbdomenCT-1k, merged different datasets like NIH and MSD, and extended them by labeling other organs (liver, spleen, and kidney) in addition to the pancreas, combining data from different institutions, from multiple vendors and acquired with different stages (arterial and venous) (Ma et al. 2022b). This dataset underwent an extensive assessment, including the definition of a benchmark for supervised, semi-supervised, weakly, and continual learning (Ma et al. 2022b). About two-thirds of the 21 reviewed studies on cancer segmentation used the MSD dataset. Although there is no publicly available knowledge on the exact tumor type (PDA, PNET, or other) in the MSD dataset, a joint research between two prominent clinical institutions, i.e. Mayo Clinic and MD Anderson Cancer Center, revealed that 103 out of the 281 CTs of the training set of MSD had lesions with imaging features of PDAC, 36 of PNET, and 65 of IPMN (Suman et al. 2021). There is a paucity of studies using datasets including different types of tumors: five (Si et al. 2021; Zou et al. 2023), and 11 (Qu et al. 2023). In all cases these datasets were private. The curation of new labeled datasets presents logistical, ethical, and privacy challenges. If the need for more public datasets for pancreas segmentation is invoked by many scientists in the field, this need is more urgent for tumors, given its clinical relevance as above discussed. In order to collect datasets of an adequate size, it would be pivotal to define the cardinality to ensure optimal performances. This is a critical point in supervised learning since manual annotation is labor intensive requiring a dedicated team of clinicians and costly. At present there is no proof on public datasets. There are only two on private datasets (Panda et al. 2021; Cavicchioli et al. 2024).

Semi-supervised and unsupervised learning look promising in exploiting datasets with few labeled data or with only unlabeled data, respectively. This review analyzed eight and four works on semi-supervised and unsupervised learning for parenchyma segmentation, respectively. For tumors only two works used semi-supervised learning. According to the literature, it seems that semi-supervised learning methods performed better than supervised ones (Chen et al. 2022d). However, this is not the case for the pancreas segmentation. For this task, supervised learning still provided the higher scores, as documented by this systematic review. While federated learning seems a promising avenue for training DL models across diverse institutions without compromising data privacy, it is noteworthy that no published studies on this approach were available for pancreas segmentation at the time of our systematic review.

One of the most frequent criticisms of AI models is the lack of generalization to data from different institutions, with different models of CT scanners by different vendors, different imaging protocols, and different patient demographics, all resulting in different statistics distributions. In fact, using the same datasets for development and validation increases the risk of overfitting and a lack of generalizability (Isensee et al. 2024). Hence, further work is required to prove the robustness of DL models to external institutions.

Future directions: Future developments for the curation of an effective benchmark dataset should consider two requirements to measure methodological differences: low standard

deviation of DSC on the same method on 5-fold cross-validation to ensure statistical stability and low noise (intra-method); and a high standard deviation across different methods to ascertain methodological differences (inter-method) (Isensee et al. 2024). Additionally, a pool of experts should establish a robust protocol for annotation, revision, and validation (Ansari et al. 2022). Furthermore, there is a need for new datasets on pancreas tumors and vascular structures, in addition to setting a benchmark for each of them as in the case of AbdomenCT-1k. In this regard the PANORAMA study dataset was recently curated for PDAC (Alves et al. 2022). At the time of this writing it consists of 2738 CTs, of which 400 for validation and 100 for test, from three centers in the Netherlands, one in Norway, and one in Sweden. It also includes 80 healthy cases from the NIH dataset, 194 with PDAC from the MSD, and The Clinical Proteomic Tumor Analysis Consortium Pancreatic Ductal Adenocarcinoma Collection. It was manually annotated by one of two trained investigators supervised by an expert radiologist with over 20 years of pancreatic cancer experience (Alves et al. 2022). Although it will be used for the first challenge on detection of PDAC, thus not segmentation, it contains annotations of the pancreas parenchyma, pancreatic duct, veins, arteries, and common bile duct for the training and test portions. Therefore, in the future it may be used for PDAC segmentation. Lastly, synthetic data generation may help to create realistic, and diverse datasets for pancreas imaging, without compromising privacy or ethical standards. In addition, research on federated learning should be encouraged by considering the benefits in terms of patients privacy.

8.2.5 Clinical translation perspective

Despite the vast number of published studies on DL for pancreas segmentation, there has not been clinical translation due to the small sizes and limited heterogeneity of labeled training datasets that have precluded clinical-grade performance and generalizability of the models (Suman et al. 2021). Overall, by considering the published literature, the following question arises: *"Should resources be directed towards refining existing models or harnessing established networks with extensive, meticulously curated datasets?"* The answer still remains complex. From a clinical standpoint, the tangible benefits for patients might be elusive if the DL models are not rigorously tested in real-world clinical settings. In fact, an increase of DSC or other metrics reflects technical progress, but it does not impact the clinical practice if there is no evidence of superiority during clinical trials (Chen et al. 2022c). Developers of DL models should not forget that improving patients outcomes is the priority of clinicians. According to the most recent guidelines on minimally invasive pancreatic surgery, there exists no corpus of evidence on the impact of AI in laparoscopic or robotic pancreatic surgery. Most of the published studies assessed the technical feasibility of utilizing AI. However, there is no demonstration of clinical implementation and validation at multiple centers (Abu Hilal et al. 2023). Given the complexities of DL models it would be difficult to obtain informed consent from patients since they need to understand how their data will be used, how the DL models work, and the potential risks involved. Similarly, the application of explainable AI to pancreas segmentation remains largely unexplored. Moreover, transitioning from research to market poses formidable challenges, including model generalizability, explainable AI, data privacy safeguards, regulations, and certification.

Future directions: Developers should provide healthcare providers and patients with clear, transparent, and simplified explanations of how DL models work. Patients will also

have the right to know how their data will be used, how their privacy will be protected, and the risks of using their data (e.g., AI model bias). More collaboration between developers of DL models and clinicians is encouraged, i.e., by collaborating within projects or in the writing of papers or by attending conferences of mutual interest (Chen et al. 2022c).

8.3 Limitations

This work has some limitations. First the systematic review is focused on DL approches based on CT imaging modality. As stated in the respective section inside the methods, the articles retrieved by our literature search on MRI and PET did not introduce technical advancements in the DL arena. Overall, the number of excluded studies on MRI were very low in comparison with those on CT, which formed the backbone of the present work. When a substantial number of publications on DL methods from MRI and other modalities will be reached, a comparison with CT could be addressed by a subsequent review. Second, the analyzed studies by the present review are limited to articles published only in peer-reviewed journals, while some researchers prefer other venues to first disseminate their work. For instance, in computer science, and its branches like computer vision, there are scientists who opt for the publication of their research findings as preprints on arXiv to share them with the scientific community as soon as possible. We decided to direct our efforts on works which underwent rigorous peer-reviewed. This is reflected by the high average impact factor of the journals where the reviewed studies were published, i.e., 5.4 according to the 2023 statistics by the Journal Citation Reports ^{TM}. Lastly, the present review did not include the gray literature, which generally publishes the most recent methods.

9 Conclusions

This systematic review of DL applications for segmenting the pancreas and its lesions elucidates significant advancements and identifies important areas of improvement. The review highlights several critical challenges. From a clinical point of view there is an urgent need to improve segmentation of pancreas cancer, given its aggressive nature, and its highly complex surgical treatment. Other clinical applications concern the segmentation of pancreas subregions, especially the head where about two thirds of pancreas cancers occur. From a technical point of view, emphasis should be placed more on method configuration than architectural variations, following the successful paradigm of the nnU-Net architecture. Moreover, the DL models should demonstrate competitive or superior performances over state-of-the-art not just on a single (and small) dataset, but on a wide range of datasets. Model size should also be considered when comparing different approaches to avoid bias in claiming that one outperformed others, which may be smaller. Metrics specific to tiny and tortuous structures like the pancreas should be designed to prevent oversegmentation which, if translated clinically, may lead to inadvertent resection of healthy tissue. There is also the need for new and large datasets, especially with annotations of pancreas tumors. The definition of benchmarks on new datasets is encouraged, following the archetypal instance of the AbdomenCT-1k one. For the clinical translation, patient outcomes are the priority. Therefore, it will be crucial to assess whether or not the improvements of DL architectures

obtained in silico have an impact on real patients. Finally, clinicians may be reluctant to adopt DL-based methods in case transparency and explainability are not ensured.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10462-024-11050-4.

Acknowledgements This study was partially funded by the Future Artificial Intelligence Research (FAIR) project, PNRR-PE - Italian Ministry of University and Research, and PNRR Program n. 352 - Italian Ministry of University and Research.

Author Contributions Andrea Moglia had the idea of the systematic review. Andrea Moglia and Matteo Cavicchioli performed the literature search and data analysis. Andrea Moglia and Matteo Cavicchioli prepared a draft. Luca Mainardi and Pietro Cerveri critically revised the work. Pietro Cerveri acquired the funding. All authors read and approved the final manuscript. Andrea Moglia has full access to all the data in the work and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Declaration

Generative AI and AI-assisted technologies in the writing process During the preparation of this work, the authors did not use any tool or software based on generative AI and AI-assisted technologies.

Conflict of interest This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue. The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

References

Abu Hilal M, van Ramshorst TM, Boggi U et al (2023) The Brescia internationally validated European guidelines on minimally invasive pancreatic surgery (egumips). Ann Surg 279(1):45–57. https://doi.org/10.1097/sla.0000000000006006

Adsay NV (2008) Cystic neoplasia of the pancreas: pathology and biology. J Gastrointest Surg 12(3):401–404

Aljabri M, AlGhamdi M (2022) A review on the use of deep learning for medical images segmentation. Neurocomputing 506:311–335. https://doi.org/10.1016/j.neucom.2022.07.070

Alves N, Schuurmans M, Litjens G et al (2022) Fully automatic deep learning framework for pancreatic ductal adenocarcinoma detection on computed tomography. Cancers 14(2):376. https://doi.org/10.3390/cancers14020376

Ansari MY, Abdalla A, Ansari MY et al (2022) Practical utility of liver segmentation methods in clinical surgeries and interventions. BMC Med Imaging 22(1):97

Azad R, Kazerouni A, Heidari M et al (2024) Advances in medical image analysis with vision transformers: a comprehensive review. Med Image Anal 91:103000. https://doi.org/10.1016/j.media.2023.103000

Bagheri MH, Roth H, Kovacs W et al (2020) Technical and clinical factors affecting success rate of a deep learning method for pancreas segmentation on ct. Acad Radiol 27(5):689–695. https://doi.org/10.1016/j.acra.2019.08.014

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. Preprint at https://doi.org/10.48550/ARXIV.1409.0473

Beger HG, Buchler MW, Kozarek R et al (2009) The pancreas: an integrated textbook of basic science, medicine, and surgery. Wiley, Hoboken

Boers TGW, Hu Y, Gibson E et al (2020) Interactive 3d u-net for the segmentation of the pancreas in computed tomography scans. Phys. Med. Biol. 65(6):065002. https://doi.org/10.1088/1361-6560/ab6f99

Burns WR, Edil BH (2012) Neuroendocrine pancreatic tumors: guidelines for management and update. Curr Treat Options Oncol 13(1):24–34

Cao H, Wang Y, Chen J, et al (2023a) Swin-unet: Unet-like pure transformer for medical image segmentation. In: Computer Vision – ECCV 2022 Workshops. Springer Nature Switzerland, pp 205–218, https://doi.org/10.1007/978-3-031-25066-8_9

Cao L, Li J (2024) Strongly representative semantic-guided segmentation network for pancreatic and pancreatic tumors. Biomed Signal Process Control 87:105562. https://doi.org/10.1016/j.bspc.2023.105562

Cao L, Li J, Chen S (2023) Multi-target segmentation of pancreas and pancreatic tumor based on fusion of attention mechanism. Biomed Signal Process Control 79:104170. https://doi.org/10.1016/j.bspc.2022.104170

Cavicchioli M, Moglia A, Pierelli L et al (2024) Main challenges on the curation of large scale datasets for pancreas segmentation using deep learning in multi-phase ct scans: Focus on cardinality, manual refinement, and annotation quality. Comput Med Imaging Graph 117:102434

Çiçek O, Abdulkadir A, Lienkamp SS, et al (2016) 3d u-net: Learning dense volumetric segmentation from sparse annotation. In: Lecture Notes in Computer Science. Springer International Publishing, pp 424–432, https://doi.org/10.1007/978-3-319-46723-8_49

Chaitanya K, Karani N, Baumgartner CF et al (2021) Semi-supervised task-driven data augmentation for medical image segmentation. Med Image Anal 68:101934. https://doi.org/10.1016/j.media.2020.101934

Chaudhari S, Mithal V, Polatkan G et al (2021) An attentive survey of attention models. ACM Transact. Intell. Syst. Technol. 12(5):1–32. https://doi.org/10.1145/3465055

Chen H, Liu Y, Shi Z (2022) Fpf-net: feature propagation and fusion based on attention mechanism for pancreas segmentation. Multimedia Syst 29(2):525–538. https://doi.org/10.1007/s00530-022-00963-1

Chen H, Liu Y, Shi Z et al (2022) Pancreas segmentation by two-view feature learning and multi-scale supervision. Biomed Signal Process Control 74:103519. https://doi.org/10.1016/j.bspc.2022.103519

Chen L, Wan L (2022) Ctunet: automatic pancreas segmentation using a channel-wise transformer and 3d u-net. Vis Comput 39(11):5229–5243. https://doi.org/10.1007/s00371-022-02656-2

Chen L, Bentley P, Mori K et al (2018) Drinet for medical image segmentation. IEEE Trans Med Imaging 37(11):2453–2462. https://doi.org/10.1109/tmi.2018.2835303

Chen L, Wang W, Jin K et al (2023) Special issue "the advance of solid tumor research in china": Prediction of sunitinib efficacy using computed tomography in patients with pancreatic neuroendocrine tumors. Int J Cancer 152(1):90–99

Chen LC, Papandreou G, Kokkinos I, et al (2016) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. Preprint at https://doi.org/10.48550/ARXIV.1606.00915

Chen X, Wang X, Zhang K et al (2022) Recent advances and clinical applications of deep learning in medical image analysis. Med Image Anal 79:102444. https://doi.org/10.1016/j.media.2022.102444

Chen Y, Ruan D, Xiao J et al (2020) Fully automated multiorgan segmentation in abdominal magnetic resonance imaging with deep neural networks. Med Phys 47(10):4971–4982. https://doi.org/10.1002/mp.14429

Chen Y, Xu C, Ding W et al (2022) Target-aware u-net with fuzzy skip connections for refined pancreas segmentation. Appl Soft Comput 131:109818. https://doi.org/10.1016/j.asoc.2022.109818

Chen Z, Wang X, Yan K et al (2020) Deep multi-scale feature fusion for pancreas segmentation from ct images. Int J Comput Assist Radiol Surg 15(3):415–423. https://doi.org/10.1007/s11548-020-02117-y

Conroy T, Pfeiffer P, Vilgrain V et al (2023) Pancreatic cancer: Esmo clinical practice guideline for diagnosis, treatment and follow-up. Ann Oncol 34(11):987–1002. https://doi.org/10.1016/j.annonc.2023.08.009

Cooke A, Smith D, Booth A (2012) Beyond pico: the spider tool for qualitative evidence synthesis. Qual Health Res 22(10):1435–1443. https://doi.org/10.1177/1049732312452938

Cui H, Pan H, Zhang K (2022) Scu-net++: a nested u-net based on sharpening filter and channel attention mechanism. Wirel Commun Mob Comput 2022:1–8. https://doi.org/10.1155/2022/2848365

Dai S, Zhu Y, Jiang X et al (2023) Td-net: Trans-deformer network for automatic pancreas segmentation. Neurocomputing 517:279–293. https://doi.org/10.1016/j.neucom.2022.10.060

Deng Y, Lan L, You L et al (2023) Automated ct pancreas segmentation for acute pancreatitis patients by combining a novel object detection approach and u-net. Biomed Signal Process Control 81:104430. https://doi.org/10.1016/j.bspc.2022.104430

Ding J, Zhang Y, Amjad A et al (2022) Automatic contour refinement for deep learning auto-segmentation of complex organs in mri-guided adaptive radiation therapy. Adv Radiat Oncol 7(5):100968. https://doi.org/10.1016/j.adro.2022.100968

Dogan RO, Dogan H, Bayrak C et al (2021) A two-phase approach using mask r-cnn and 3d u-net for high-accuracy automatic segmentation of pancreas in ct imaging. Comput Methods Programs Biomed 207:106141. https://doi.org/10.1016/j.cmpb.2021.106141

Dosovitskiy A, Beyer L, Kolesnikov A, et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. Preprint at https://doi.org/10.48550/ARXIV.2010.11929

Du Y, Zuo X, Liu S et al (2023) Segmentation of pancreatic tumors based on multi-scale convolution and channel attention mechanism in the encoder-decoder scheme. Med Phys 50(12):7764–7778. https://doi.org/10.1002/mp.16561

Ducreux M, Cuhna AS, Caramella C et al (2015) Cancer of the pancreas: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. Ann Oncol 26:v56–v68

Duh MM, Torra-Ferrer N, Riera-Marín M et al (2023) Deep learning to detect pancreatic cystic lesions on abdominal computed tomography scans: Development and validation study. JMIR AI 2:e40702. https://doi.org/10.2196/40702

Falconi M, Tamburrino D, Buzzetti E et al (2016) Il carcinoma pancreatico. Recenti Prog Med 107(6):337–340

Fang K, He B, Liu L et al (2023) Umrformer-net: a three-dimensional u-shaped pancreas segmentation method based on a double-layer bridged transformer network. Quant Imaging Med Surg 13(3):1619–1630. https://doi.org/10.21037/qims-22-544

Farag A, Lu L, Turkbey E, et al (2014) A bottom-up approach for automatic pancreas segmentation in abdominal ct scans. In: Abdominal Imaging. Computational and Clinical Applications: 6th International Workshop, ABDI 2014, Held in Conjunction with MICCAI 2014, Cambridge, MA, USA, September 14, 2014. 6, Springer, pp 103–113

Farag A, Lu L, Roth HR et al (2017) A bottom-up approach for pancreas segmentation using cascaded superpixels and (deep) image patch labeling. IEEE Trans Image Process 26(1):386–399. https://doi.org/10.1109/tip.2016.2624198

Fleurentin A, Mazellier JP, Meyer A et al (2023) Automatic pancreas anatomical part detection in endoscopic ultrasound videos. Comput Methods Biomech Biomed Eng Imaging Visual 11(4):1136–1142. https://doi.org/10.1080/21681163.2022.2154274

Francis S, Jayaraj PB, Pournami PN et al (2023) Contourgan: auto-contouring of organs at risk in abdomen computed tomography images using generative adversarial network. Int J Imaging Syst Technol 33(5):1494–1504. https://doi.org/10.1002/ima.22901

Fu Y, Mazur TR, Wu X et al (2018) A novel mri segmentation method using cnn-based correction network for mri-guided adaptive radiotherapy. Med Phys 45(11):5129–5137. https://doi.org/10.1002/mp.13221

Ge R, Shi F, Chen Y et al (2023) Improving anisotropy resolution of computed tomography and annotation using 3d super-resolution network. Biomed Signal Process Control 82:104590. https://doi.org/10.1016/j.bspc.2023.104590

Ghorpade H, Jagtap J, Patil S et al (2023) Automatic segmentation of pancreas and pancreatic tumor: A review of a decade of research. IEEE Access 11:108727–108745. https://doi.org/10.1109/access.2023.3320570

Gibson E, Giganti F, Hu Y et al (2018) Automatic multi-organ segmentation on abdominal ct with dense v-networks. IEEE Trans Med Imaging 37(8):1822–1834. https://doi.org/10.1109/tmi.2018.2806309

Gong Z, Guo W, Zhou W et al (2020) A deep learning based level set model for pancreas segmentation. J Med Imaging Health Inform 10(11):2681–2685. https://doi.org/10.1166/jmihi.2020.3200

Goodfellow IJ, Pouget-Abadie J, Mirza M, et al (2014) Generative adversarial networks. Preprint at https://doi.org/10.48550/ARXIV.1406.2661

Guo MH, Xu TX, Liu JJ et al (2022) Attention mechanisms in computer vision: A survey. Comput Visual Media 8(3):331–368. https://doi.org/10.1007/s41095-022-0271-y

Wl Guo, Ak Geng, Geng C et al (2022) Combination of unet++ and resnest to classify chronic inflammation of the choledochal cystic wall in patients with pancreaticobiliary maljunction. Brit J Radiol. https://doi.org/10.1259/bjr.20201189

He J, Xu C (2023) Hybrid transformer-cnn with boundary-awareness network for 3d medical image segmentation. Appl Intell 53(23):28542–28554. https://doi.org/10.1007/s10489-023-05032-2

Heinrich MP, Blendowski M, Oktay O (2018) Ternarynet: faster deep model inference without gpus for medical 3d segmentation using sparse and binary convolutions. Int J Comput Assist Radiol Surg 13(9):1311–1320. https://doi.org/10.1007/s11548-018-1797-4

Hu EJ, Shen Y, Wallis P, et al (2021a) Lora: Low-rank adaptation of large language models. Preprint at arXiv:2106.09685

Hu J, Shen L, Albanie S, et al (2017) Squeeze-and-excitation networks. Preprint at https://doi.org/10.48550/ARXIV.1709.01507

Hu P, Li X, Tian Y et al (2021) Automatic pancreas segmentation in ct images with distance-based saliency-aware denseaspp network. IEEE J Biomed Health Inform 25(5):1601–1611. https://doi.org/10.1109/jbhi.2020.3023462

Huang B, Lin X, Shen J et al (2021) Accurate and feasible deep learning based semi-automatic segmentation in ct for radiomics analysis in pancreatic neuroendocrine neoplasms. IEEE J Biomed Health Inform 25(9):3498–3506. https://doi.org/10.1109/jbhi.2021.3070708

Huang B, Huang H, Zhang S et al (2022) Artificial intelligence in pancreatic cancer. Theranostics 12(16):6931–6954. https://doi.org/10.7150/thno.77949

Huang M, Huang C, Yuan J et al (2021) A semiautomated deep learning approach for pancreas segmentation. J Healthc Eng 2021:1–10. https://doi.org/10.1155/2021/3284493

Huang ML, Wu YZ (2022) Semantic segmentation of pancreatic medical images by using convolutional neural network. Biomed Signal Process Control 73:103458. https://doi.org/10.1016/j.bspc.2021.103458

Mx Huang, Yj Wang, Cf Huang et al (2022) Learning a discriminative feature attention network for pancreas ct segmentation. Appl Math J Chin Univ 37(1):73–90. https://doi.org/10.1007/s11766-022-4346-4

Huang X, Deng Z, Li D et al (2023) Missformer: an effective transformer for 2d medical image segmentation. IEEE Trans Med Imaging 42(5):1484–1494. https://doi.org/10.1109/tmi.2022.3230943

Huang Y, Wen J, Wang Y et al (2022) Subset selection strategy-based pancreas segmentation in ct. Quant Imaging Med Surg 12(6):3061–3077. https://doi.org/10.21037/qims-21-798

Huttenlocher D, Klanderman G, Rucklidge W (1993) Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Mach Intell 15(9):850–863. https://doi.org/10.1109/34.232073

Irshad S, Gomes DPS, Kim ST (2023) Improved abdominal multi-organ segmentation via 3d boundary-constrained deep neural networks. IEEE Access 11:35097–35110. https://doi.org/10.1109/access.2023.3264582

Isensee F, Jäger PF, Kohl SAA, et al (2019) Automated design of deep learning methods for biomedical image segmentation. Preprint at https://doi.org/10.48550/ARXIV.1904.08128

Isensee F, Jaeger PF, Kohl SAA et al (2020) nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18(2):203–211. https://doi.org/10.1038/s41592-020-01008-z

Isensee F, Wald T, Ulrich C, et al (2024) nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. Preprint at arXiv:2404.09556

Iwasa Y, Iwashita T, Takeuchi Y et al (2021) Automatic segmentation of pancreatic tumors using deep learning on a video image of contrast-enhanced endoscopic ultrasound. J Clin Med 10(16):3589. https://doi.org/10.3390/jcm10163589

Jaccard P (1912) The distribution of the flora in the alpine zone.1. New Phytol 11(2):37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Jaderberg M, Simonyan K, Zisserman A, et al (2015) Spatial transformer networks. Preprint at https://doi.org/10.48550/ARXIV.1506.02025

Jain S, Sikka G, Dhir R (2023) An automatic cascaded approach for pancreas segmentation via an unsupervised localization using 3d ct volumes. Multimedia Syst 29(4):2337–2349. https://doi.org/10.1007/s00530-023-01115-9

Javed S, Qureshi TA, Deng Z et al (2022) Segmentation of pancreatic subregions in computed tomography images. J Imaging 8(7):195. https://doi.org/10.3390/jimaging8070195

Ji Y, Bai H, Yang J, et al (2022) Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Preprint at https://doi.org/10.48550/ARXIV.2206.08023

Jiang J, Hong J, Tringale K et al (2023) Progressively refined deep joint registration segmentation (prorseg) of gastrointestinal organs at risk: Application to mri and cone-beam ct. Med Phys 50(8):4758–4774. https://doi.org/10.1002/mp.16527

Johnston ME, Carter MM, Wilson GC et al (2020) Surgical management of primary pancreatic neuroendocrine tumors. J Gastrointest Oncol 11(3):578

Ju J, Li J, Chang Z et al (2023) Incorporating multi-stage spatial visual cues and active localization offset for pancreas segmentation. Pattern Recogn Lett 170:85–92. https://doi.org/10.1016/j.patrec.2023.05.004

Kamnitsas K, Ledig C, Newcombe VF et al (2017) Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Med Image Anal 36:61–78. https://doi.org/10.1016/j.media.2016.10.004

Karimi D, Salcudean SE (2020) Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks. IEEE Trans Med Imaging 39(2):499–513. https://doi.org/10.1109/tmi.2019.2930068

Kart T, Fischer M, Küstner T et al (2021) Deep learning-based automated abdominal organ segmentation in the UK biobank and German national cohort magnetic resonance imaging studies. Invest Radiol 56(6):401–408. https://doi.org/10.1097/RLI.0000000000000755

Khan A, Khosa F, Eisenberg RL (2011) Cystic lesions of the pancreas. Am J Roentgenol 196(6):W668–W677

Khasawneh H, Patra A, Rajamohan N et al (2022) Volumetric pancreas segmentation on computed tomography: accuracy and efficiency of a convolutional neural network versus manual segmentation in 3d slicer in the context of interreader variability of expert radiologists. J Comput Assist Tomogr 46(6):841–847. https://doi.org/10.1097/rct.0000000000001374
Kirillov A, Mintun E, Ravi N, et al (2023) Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 4015–4026
Klein S, Staring M, Murphy K et al (2010) elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging 29(1):196–205. https://doi.org/10.1109/tmi.2009.2035616
Knolle M, Kaissis G, Jungmann F et al (2021) Efficient, high-performance semantic segmentation using multi-scale feature extraction. PLoS ONE 16(8):e0255397. https://doi.org/10.1371/journal.pone.0255397
Kumar H, DeSouza SV, Petrov MS (2019) Automated pancreas segmentation from computed tomography and magnetic resonance images: a systematic review. Comput Methods Programs Biomed 178:319–328. https://doi.org/10.1016/j.cmpb.2019.07.002
Landman B, Xu Z, Igelsias J, et al (2015) Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge, p 12, https://doi.org/10.7303/syn3193805
Li C, Mao Y, Guo Y et al (2022) Multi-dimensional cascaded net with uncertain probability reduction for abdominal multi-organ segmentation in ct sequences. Comput Methods Programs Biomed 221:106887. https://doi.org/10.1016/j.cmpb.2022.106887
Li F, Li W, Shu Y et al (2020) Multiscale receptive field based on residual network for pancreas segmentation in ct images. Biomed Signal Process Control 57:101828. https://doi.org/10.1016/j.bspc.2019.101828
Li J, Lin X, Che H et al (2021) Pancreas segmentation with probabilistic map guided bi-directional recurrent unet. Phys Med Biol 66(11):115010. https://doi.org/10.1088/1361-6560/abfce3
Li J, Feng C, Shen Q et al (2022) Pancreatic cancer segmentation in unregistered multi-parametric MRI with adversarial learning and multi-scale supervision. Neurocomputing 467:310–322. https://doi.org/10.1016/j.neucom.2021.09.058
Li J, Qi L, Chen Q et al (2022) A dual meta-learning framework based on idle data for enhancing segmentation of pancreatic cancer. Med Image Anal 78:102342. https://doi.org/10.1016/j.media.2021.102342
Li J, Chen T, Qian X (2023) Generalizable pancreas segmentation modeling in ct imaging via meta-learning and latent-space feature flow generation. IEEE J Biomed Health Inform 27(1):374–385. https://doi.org/10.1109/jbhi.2022.3207597
Li J, Yin W, Wang Y (2023) Papnet: Convolutional network for pancreatic cyst segmentation. J Xray Sci Technol 31(3):655–668. https://doi.org/10.3233/xst-230011
Li J, Zhu H, Chen T et al (2023) Generalizable pancreas segmentation via a dual self-supervised learning framework. IEEE J Biomed Health Inform 27(10):4780–4791. https://doi.org/10.1109/jbhi.2023.3294278
Li L, Zhao H, Wang H et al (2023) Automatic abdominal segmentation using novel 3d self-adjustable organ aware deep network in ct images. Biomed Signal Process Control 84:104691. https://doi.org/10.1016/j.bspc.2023.104691
Li M, Lian F, Guo S (2020) Pancreas segmentation based on an adversarial model under two-tier constraints. Phys Med Biol 65(22):225021. https://doi.org/10.1088/1361-6560/abb6bf
Li M, Lian F, Guo S (2021) Automatic pancreas segmentation using double adversarial networks with pyramidal pooling module. IEEE Access 9:140965–140974. https://doi.org/10.1109/access.2021.3118718
Li M, Lian F, Guo S (2021) Multi-scale selection and multi-channel fusion model for pancreas segmentation using adversarial deep convolutional nets. J Digit Imaging 35(1):47–55. https://doi.org/10.1007/s10278-021-00563-x
Li M, Lian F, Wang C et al (2021) Accurate pancreas segmentation using multi-level pyramidal pooling residual u-net with adversarial mechanism. BMC Med Imaging. https://doi.org/10.1186/s12880-021-00694-1
Li M, Lian F, Wang C et al (2021) Dual adversarial convolutional networks with multilevel cues for pancreatic segmentation. Phys Med Biol 66(17):175025. https://doi.org/10.1088/1361-6560/ac155f
Li M, Lian F, Li Y et al (2022) Attention-guided duplex adversarial u-net for pancreatic segmentation from computed tomography images. J Appl Clin Med Phys. https://doi.org/10.1002/acm2.13537
Li Q, Li X, Liu W et al (2023) Non-enhanced magnetic resonance imaging-based radiomics model for the differentiation of pancreatic adenosquamous carcinoma from pancreatic ductal adenocarcinoma. Front Oncol 13:1108545. https://doi.org/10.3389/fonc.2023.1108545
Li Q, Liu X, He Y et al (2023) Temperature guided network for 3d joint segmentation of the pancreas and tumors. Neural Netw 157:387–403. https://doi.org/10.1016/j.neunet.2022.10.026
Li Q, Zhou Z, Chen Y et al (2023) Fully automated magnetic resonance imaging-based radiomics analysis for differentiating pancreatic adenosquamous carcinoma from pancreatic ductal adenocarcinoma. Abdom Radiol 48(6):2074–2084. https://doi.org/10.1007/s00261-023-03801-8
Li W, Qin S, Li F et al (2020) Mad-unet: a deep u-shaped network combined with an attention mechanism for pancreas segmentation in ct images. Med Phys 48(1):329–341. https://doi.org/10.1002/mp.14617

Li X, Sun X, Meng Y, et al (2019) Dice loss for data-imbalanced NLP tasks. Preprint at https://doi.org/10.48550/ARXIV.1911.02855
Liang Y, Schott D, Zhang Y et al (2020) Auto-segmentation of pancreatic tumor in multi-parametric MRI using deep convolutional neural networks. Radiother Oncol 145:193–200. https://doi.org/10.1016/j.radonc.2020.01.021
Lim SH, Kim YJ, Park YH et al (2022) Automated pancreas segmentation and volumetry using deep neural network on computed tomography. Sci Rep. https://doi.org/10.1038/s41598-022-07848-3
Lin TY, Goyal P, Girshick R, et al (2017) Focal loss for dense object detection. Preprint at https://doi.org/10.48550/ARXIV.1708.02002
Liu P, Zheng G (2023) Cvcl: context-aware voxel-wise contrastive learning for label-efficient multi-organ segmentation. Comput Biol Med 160:106995. https://doi.org/10.1016/j.compbiomed.2023.106995
Liu S, Yuan X, Hu R et al (2020) Automatic pancreas segmentation via coarse location and ensemble learning. IEEE Access 8:2906–2914. https://doi.org/10.1109/access.2019.2961125
Liu S, Liang S, Huang X et al (2022) Graph-enhanced u-net for semi-supervised segmentation of pancreas from abdomen ct scan. Phys Med Biol 67(15):155017. https://doi.org/10.1088/1361-6560/ac80e4
Liu Y, Duan Y, Zeng T (2022) Learning multi-level structural information for small organ segmentation. Signal Process 193:108418. https://doi.org/10.1016/j.sigpro.2021.108418
Liu Y, Yang B, Chen X et al (2023) Efficient segmentation using domain adaptation for MRI-guided and CBCT-guided online adaptive radiotherapy. Radiother Oncol 188:109871. https://doi.org/10.1016/j.radonc.2023.109871
Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. Preprint at https://doi.org/10.48550/ARXIV.2103.14030
Liu Z, Mao H, Wu CY, et al (2022c) A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11976–11986
Liu Z, Su J, Wang R et al (2022) Pancreas co-segmentation based on dynamic ROI extraction and VGGU-net. Expert Syst Appl 192:116444. https://doi.org/10.1016/j.eswa.2021.116444
Long J, Song X, An Y et al (2021) Parallel multi-scale network with attention mechanism for pancreas segmentation. IEEJ Trans Electr Electron Eng 17(1):110–119. https://doi.org/10.1002/tee.23493
Lu L, Jian L, Luo J et al (2019) Pancreatic segmentation via ringed residual u-net. IEEE Access 7:172871–172878. https://doi.org/10.1109/access.2019.2956550
Luo X, Liao W, Xiao J et al (2022) Word: a large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. Med Image Anal 82:102642. https://doi.org/10.1016/j.media.2022.102642
Ma H, Zou Y, Liu PX (2021) Mhsu-net: a more versatile neural network for medical image segmentation. Comput Methods Programs Biomed 208:106230. https://doi.org/10.1016/j.cmpb.2021.106230
Ma J, He J, Yang X (2021) Learning geodesic active contours for embedding object global information in segmentation CNNS. IEEE Trans Med Imaging 40(1):93–104. https://doi.org/10.1109/tmi.2020.3022693
Ma J, Zhang Y, Gu S et al (2022) Fast and low-GPU-memory abdomen CT organ segmentation: the flare challenge. Med Image Anal 82:102616
Ma J, Zhang Y, Gu S et al (2022) Abdomenct-1k: is abdominal organ segmentation a solved problem? IEEE Trans Pattern Anal Mach Intell 44(10):6695–6714. https://doi.org/10.1109/tpami.2021.3100536
Ma J, He Y, Li F et al (2024) Segment anything in medical images. Nat Commun 15(1):654
Mahmoudi T, Kouzahkanan ZM, Radmard AR et al (2022) Segmentation of pancreatic ductal adenocarcinoma (PDAC) and surrounding vessels in CT images using deep convolutional neural networks and texture descriptors. Sci Rep. https://doi.org/10.1038/s41598-022-07111-9
Man Y, Huang Y, Feng J et al (2019) Deep Q learning driven CT pancreas segmentation with geometry-aware U-net. IEEE Trans Med Imaging 38(8):1971–1980. https://doi.org/10.1109/tmi.2019.2911588
Mazor N, Dar G, Lederman R et al (2024) Mc3du-net: a multisequence cascaded pipeline for the detection and segmentation of pancreatic cysts in MRI. Int J Comput Assist Radiol Surg 19(3):423–432. https://doi.org/10.1007/s11548-023-03020-y
Mazurowski MA, Dong H, Gu H et al (2023) Segment anything model for medical image analysis: an experimental study. Med Image Anal 89:102918
McGuigan A, Kelly P, Turkington RC et al (2018) Pancreatic cancer: a review of clinical diagnosis, epidemiology, treatment and outcomes. World J Gastroenterol 24(43):4846
Milletari F, Navab N, Ahmadi SA (2016) V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE, pp 565–571, https://doi.org/10.1109/3dv.2016.79
Miyamoto R, Shiihara M, Shimoda M et al (2024) Laparoscopic distal pancreatectomy using three-dimensional computer graphics for surgical navigation with a deep learning algorithm: a case report. Cureus. https://doi.org/10.7759/cureus.55907

Mo J, Zhang L, Wang Y et al (2020) Iterative 3d feature enhancement network for pancreas segmentation from CT images. Neural Comput Appl 32(16):12535–12546. https://doi.org/10.1007/s00521-020-04710-3
Moglia A, Georgiou K, Cerveri P et al (2024) Large language models in healthcare: from a systematic review on medical examinations to a comparative analysis on fundamentals of robotic surgery online test. Artif Intell Rev 57(9):1–54
Mukherjee S, Korfiatis P, Khasawneh H et al (2023) Bounding box-based 3d ai model for user-guided volumetric segmentation of pancreatic ductal adenocarcinoma on standard-of-care cts. Pancreatology 23(5):522–529. https://doi.org/10.1016/j.pan.2023.05.008
Ni H, Zhou G, Chen X et al (2023) Predicting recurrence in pancreatic ductal adenocarcinoma after radical surgery using an AX-Unet pancreas segmentation model and dynamic nomogram. Bioengineering 10(7):828. https://doi.org/10.3390/bioengineering10070828
Nikolov S, Blackwell S, Zverovitch A et al (2021) Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. J Med Internet Res 23(7):e26151. https://doi.org/10.2196/26151
Ning Y, Han Z, Zhong L et al (2020) Dran: deep recurrent adversarial network for automated pancreassegmentation. IET Image Proc 14(6):1091–1100. https://doi.org/10.1049/iet-ipr.2019.0399
Nishio M, Noguchi S, Fujimoto K (2020) Automatic pancreas segmentation using coarse-scaled 2d model of deep learning: usefulness of data augmentation and deep u-net. Appl Sci 10(10):3360. https://doi.org/10.3390/app10103360
Oktay O, Schlemper J, Folgoc LL, et al (2018) Attention U-net: Learning where to look for the pancreas. Preprint at https://doi.org/10.48550/ARXIV.1804.03999
Page MJ, McKenzie JE, Bossuyt PM et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. https://doi.org/10.1136/bmj.n71
Pradip Paithane, Kakarwal S (2023) LMNS-net: lightweight multiscale novel semantic-net deep learning approach used for automatic pancreas image segmentation in ct scan images. Expert Syst Appl 234:121064. https://doi.org/10.1016/j.eswa.2023.121064
Pan S, Chang C, Wang T et al (2022) Abdomen CT multi-organ segmentation using token-based MLP-mixer. Med Phys 50(5):3027–3038. https://doi.org/10.1002/mp.16135
Panda A, Korfiatis P, Suman G et al (2021) 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. Med Phys 48(5):2468–2481. https://doi.org/10.1002/mp.14782
Park S, Chu L, Fishman E et al (2020) Annotated normal CT data of the abdomen for deep learning: challenges and strategies for implementation. Diagn Interv Imaging 101(1):35–44. https://doi.org/10.1016/j.diii.2019.05.008
Petit O, Thome N, Soler L (2021) 3d spatial priors for semi-supervised organ segmentation with deep convolutional neural networks. Int J Comput Assist Radiol Surg 17(1):129–139. https://doi.org/10.1007/s11548-021-02494-y
Qiao Y, van Lew B, Lelieveldt BPF et al (2016) Fast automatic step size estimation for gradient descent optimization of image registration. IEEE Trans Med Imaging 35(2):391–403. https://doi.org/10.1109/tmi.2015.2476354
Qiu C, Song Y, Liu Z et al (2022) CMFCUNet: cascaded multi-scale feature calibration UNet for pancreas segmentation. Multimedia Syst 29(2):871–886. https://doi.org/10.1007/s00530-022-01020-7
Qiu C, Xue J, Liu X et al (2022) Deep dynamic spiking neural p systems with applications in organ segmentation. J Membr Comput 4(4):329–340. https://doi.org/10.1007/s41965-022-00115-4
Qiu C, Liu Z, Song Y et al (2023) RTUnet: Residual transformer UNet specifically for pancreas segmentation. Biomed Signal Process Control 79:104173. https://doi.org/10.1016/j.bspc.2022.104173
Qu T, Wang X, Fang C et al (2022) M3net: a multi-scale multi-view framework for multi-phase pancreas segmentation based on cross-phase non-local attention. Med Image Anal 75:102232. https://doi.org/10.1016/j.media.2021.102232
Qu T, Li X, Wang X et al (2023) Transformer guided progressive fusion network for 3D pancreas and pancreatic mass segmentation. Med Image Anal 86:102801. https://doi.org/10.1016/j.media.2023.102801
Qureshi TA, Lynch C, Azab L et al (2022) Morphology-guided deep learning framework for segmentation of pancreas in computed tomography images. J Med Imaging. https://doi.org/10.1117/1.jmi.9.2.024002
Ravi N, Gabeur V, Hu YT, et al (2024) Sam 2: Segment anything in images and videos. Preprint at arXiv:2408.00714
Rawla P, Sunkara T, Gaduputi V (2019) Epidemiology of pancreatic cancer: global trends, etiology and risk factors. World J Oncol 10(1):10
Rehman A, Khan FG (2020) A deep learning based review on abdominal images. Multimed Tools Appl 80(20):30321–30352. https://doi.org/10.1007/s11042-020-09592-0

Reinke A, Tizabi MD, Sudre CH, et al (2021) Common limitations of image processing metrics: a picture story. Preprint at arXiv:2104.05642

Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. Springer International Publishing, pp 234–241, https://doi.org/10.1007/978-3-319-24574-4_28

Roth HR, Lu L, Farag A, et al (2015) Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, pp 556–564, https://doi.org/10.1007/978-3-319-24553-9_68

Roth HR, Lu L, Lay N et al (2018) Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Med Image Anal 45:94–107. https://doi.org/10.1016/j.media.2018.01.006

Roth HR, Oda H, Zhou X et al (2018) An application of cascaded 3D fully convolutional networks for medical image segmentation. Comput Med Imaging Graph 66:90–99. https://doi.org/10.1016/j.compmedimag.2018.03.001

Roy S, Koehler G, Ulrich C, et al (2023) Mednext: Transformer-driven scaling of convnets for medical image segmentation. Preprint at arXiv:2303.09975

Russell TB, Aroori S (2022) The pancreas from a surgical perspective: an illustrated overview. Art Surg 6

Schlemper J, Oktay O, Schaap M et al (2019) Attention gated networks: learning to leverage salient regions in medical images. Med Image Anal 53:197–207. https://doi.org/10.1016/j.media.2019.01.012

Seidlitz S, Sellner J, Odenthal J et al (2022) Robust deep learning-based semantic organ segmentation in hyperspectral images. Med Image Anal 80:102488. https://doi.org/10.1016/j.media.2022.102488

Senkyire IB, Liu Z (2021) Supervised and semi-supervised methods for abdominal organ segmentation: a review. Int J Autom Comput 18(6):887–914. https://doi.org/10.1007/s11633-021-1313-0

Seo K, Lim JH, Seo J et al (2022) Semantic segmentation of pancreatic cancer in endoscopic ultrasound images using deep learning approach. Cancers 14(20):5111. https://doi.org/10.3390/cancers14205111

Shamshad F, Khan S, Zamir SW et al (2023) Transformers in medical imaging: a survey. Med Image Anal 88:102802. https://doi.org/10.1016/j.media.2023.102802

Shan T, Yan J (2021) SCA-Net: a spatial and channel attention network for medical image segmentation. IEEE Access 9:160926–160937. https://doi.org/10.1109/access.2021.3132293

Shen C, Roth HR, Hayashi Y et al (2022) A cascaded fully convolutional network framework for dilated pancreatic duct segmentation. Int J Comput Assist Radiol Surg 17(2):343–354. https://doi.org/10.1007/s11548-021-02530-x

Shen N, Wang Z, Li J et al (2023) Multi-organ segmentation network for abdominal CT images based on spatial attention and deformable convolution. Expert Syst Appl 211:118625. https://doi.org/10.1016/j.eswa.2022.118625

Shi G, Xiao L, Chen Y et al (2021) Marginal loss and exclusion loss for partially supervised multi-organ segmentation. Med Image Anal 70:101979. https://doi.org/10.1016/j.media.2021.101979

Shi Y, Zhang J, Ling T et al (2022) Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. IEEE Trans Med Imaging 41(3):608–620. https://doi.org/10.1109/tmi.2021.3117888

Shi Y, Wang H, Ji H et al (2023) A deep weakly semi-supervised framework for endoscopic lesion segmentation. Med Image Anal 90:102973

Si K, Xue Y, Yu X et al (2021) Fully end-to-end deep-learning-based diagnosis of pancreatic tumors. Theranostics 11(4):1982–1990. https://doi.org/10.7150/thno.52508

Siegel RL, Giaquinto AN, Jemal A (2024) Cancer statistics. CA: Cancer J Clin 74(1):12–49. https://doi.org/10.3322/caac.21820

Simpson AL, Antonelli M, Bakas S, et al (2019) A large annotated medical image dataset for the development and evaluation of segmentation algorithms. https://doi.org/10.48550/ARXIV.1902.09063

Suman G, Patra A, Korfiatis P et al (2021) Quality gaps in public pancreas imaging datasets: implications & challenges for ai applications. Pancreatology 21(5):1001–1008

Sundar LKS, Yu J, Muzik O et al (2022) Fully automated, semantic segmentation of whole-body18F-FDG PET/CT images based on data-centric artificial intelligence. J Nucl Med 63(12):1941–1948. https://doi.org/10.2967/jnumed.122.264063

Sureka B, Jha S, Yadav A et al (2021) Mdct evaluation of pancreatic contour variations in head, neck, body and tail: surgical and radiological significance. Surg Radiol Anat 43:1405–1412

Tang A, Gong P, Fang N et al (2023) Endoscopic ultrasound diagnosis system based on deep learning in images capture and segmentation training of solid pancreatic masses. Med Phys 50(7):4197–4205. https://doi.org/10.1002/mp.16390

Tang A, Tian L, Gao K et al (2023) Contrast-enhanced harmonic endoscopic ultrasound (CH-EUS) master: a novel deep learning-based system in pancreatic mass diagnosis. Cancer Med 12(7):7962–7973. https://doi.org/10.1002/cam4.5578

Tian L, Zou L, Yang X (2023) A two-stage data-model driven pancreas segmentation strategy embedding directional information of the boundary intensity gradient and deep adaptive pointwise parameters. Phys Med Biol 68(14):145005. https://doi.org/10.1088/1361-6560/ace099
Tian M, He J, Yu X et al (2021) MCMC guided CNN training and segmentation for pancreas extraction. IEEE Access 9:90539–90554. https://doi.org/10.1109/access.2021.3070391
Tong N, Gou S, Niu T et al (2020) Self-paced densenet with boundary constraint for automated multi-organ segmentation on abdominal CT images. Phys Med Biol 65(13):135011. https://doi.org/10.1088/1361-6560/ab9b57
Tong N, Xu Y, Zhang J et al (2023) Robust and efficient abdominal CT segmentation using shape constrained multi-scale attention network. Physica Med 110:102595. https://doi.org/10.1016/j.ejmp.2023.102595
Touvron H, Cord M, Douze M, et al (2020) Training data-efficient image transformers & distillation through attention. Preprint at https://doi.org/10.48550/ARXIV.2012.12877
Turečková A, Tureček T, Komínková Oplatková Z et al (2020) Improving CT image tumor segmentation through deep supervision and attentional gates. Front Robot AI. https://doi.org/10.3389/frobt.2020.00106
Vareedayah AA, Alkaade S, Taylor JR (2018) Pancreatic adenocarcinoma. Mo Med 115(3):230
Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Preprint at https://doi.org/10.48550/ARXIV.1706.03762
Viriyasaranon T, Woo SM, Choi JH (2023) Unsupervised visual representation learning based on segmentation of geometric pseudo-shapes for transformer-based medical tasks. IEEE J Biomed Health Inform 27(4):2003–2014. https://doi.org/10.1109/jbhi.2023.3237596
Wang F, Jiang M, Qian C, et al (2017) Residual attention network for image classification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 3156–3164, https://doi.org/10.1109/cvpr.2017.683
Wang F, Cheng C, Cao W et al (2023) MFCNet: a multi-modal fusion and calibration networks for 3D pancreas tumor segmentation on PET-CT images. Comput Biol Med 155:106657. https://doi.org/10.1016/j.compbiomed.2023.106657
Wang L, Chen R, Wang S et al (2019) Nested dilation network (NDN) for multi-task medical image segmentation. IEEE Access 7:44676–44685. https://doi.org/10.1109/access.2019.2908386
Wang Y, Zhou Y, Shen W et al (2019) Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. Med Image Anal 55:88–102. https://doi.org/10.1016/j.media.2019.04.005
Wang Y, Gong G, Kong D et al (2021) Pancreas segmentation using a dual-input V-mesh network. Med Image Anal 69:101958. https://doi.org/10.1016/j.media.2021.101958
Wang Y, Tang P, Zhou Y et al (2021) Learning inductive attention guidance for partially supervised pancreatic ductal adenocarcinoma prediction. IEEE Trans Med Imaging 40(10):2723–2735. https://doi.org/10.1109/tmi.2021.3060066
Wang Y, Zhang J, Cui H et al (2021) View adaptive learning for pancreas segmentation. Biomed Signal Process Control 66:102347. https://doi.org/10.1016/j.bspc.2020.102347
Wei K, Hackert T (2021) Surgical treatment of pancreatic ductal adenocarcinoma. Cancers 13(8):1971
Woo S, Park J, Lee JY, et al. (2018) Convolutional block attention module. Preprint at https://doi.org/10.48550/ARXIV.1807.06521
Wu H, Li X, Lin Y et al (2023) Compete to win: enhancing pseudo labels for barely-supervised medical image segmentation. IEEE Trans Med Imaging 42:3244–3255. https://doi.org/10.1109/TMI.2023.3279110
Xia E, He J, Liao Z (2023) MFA-ICPS: semi-supervised medical image segmentation with improved cross pseudo supervision and multi-dimensional feature attention. Med Phys 51(3):1918–1930. https://doi.org/10.1002/mp.16740
Xia Y, Yang D, Yu Z et al (2020) Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Med Image Anal 65:101766. https://doi.org/10.1016/j.media.2020.101766
Xie L, Yu Q, Zhou Y et al (2020) Recurrent saliency transformation network for tiny target segmentation in abdominal CT scans. IEEE Trans Med Imaging 39(2):514–525. https://doi.org/10.1109/tmi.2019.2930679
Xue J, He K, Nie D et al (2021) Cascaded multitask 3-D fully convolutional networks for pancreas segmentation. IEEE Transactions on Cybernetics 51(4):2153–2165. https://doi.org/10.1109/tcyb.2019.2955178
Yan Y, Zhang D (2021) Multi-scale U-like network with attention mechanism for automatic pancreas segmentation. PLoS ONE 16(5):e0252287. https://doi.org/10.1371/journal.pone.0252287
Yang JZ, Zhao J, Nemati R et al (2022) An adapted deep convolutional neural network for automatic measurement of pancreatic fat and pancreatic volume in clinical multi-protocol magnetic resonance images: a retrospective study with multi-ethnic external validation. Biomedicines 10(11):2991. https://doi.org/10.3390/biomedicines10112991

Yang M, Zhang Y, Chen H et al (2022) AX-UNet: a deep learning framework for image segmentation to assist pancreatic tumor diagnosis. Front Oncol. https://doi.org/10.3389/fonc.2022.894970
Yao L, Zhang J, Liu J et al (2021) A deep learning-based system for bile duct annotation and station recognition in linear endoscopic ultrasound. EBioMedicine. https://doi.org/10.1016/j.ebiom.2021.103238
Yao X, Song Y, Liu Z (2019) Advances on pancreas segmentation: a review. Multimed Tools Appl 79(9–10):6799–6821. https://doi.org/10.1007/s11042-019-08320-7
Yeghiazaryan V, Voiculescu I (2018) Family of boundary overlap metrics for the evaluation of medical image segmentation. J Med Imaging 5(01):1. https://doi.org/10.1117/1.jmi.5.1.015006
You C, Zhou Y, Zhao R et al (2022) SimCVD: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. IEEE Trans Med Imaging 41(9):2228–2237. https://doi.org/10.1109/tmi.2022.3161829
Yuan F, Tang Z, Wang C et al (2023) A multiple gated boosting network for multi-organ medical image segmentation. IET Image Proc 17(10):3028–3039. https://doi.org/10.1049/ipr2.12852
Zeng G, Zheng G (2019) Holistic decomposition convolution for effective semantic segmentation of medical volume images. Med Image Anal 57:149–164. https://doi.org/10.1016/j.media.2019.07.003
Zeng X, Huang R, Zhong Y et al (2022) A reciprocal learning strategy for semisupervised medical image segmentation. Med Phys 50(1):163–177. https://doi.org/10.1002/mp.15923
Zhang D, Zhang J, Zhang Q et al (2021) Automatic pancreas segmentation based on lightweight DCNN modules and spatial prior propagation. Pattern Recogn 114:107762. https://doi.org/10.1016/j.patcog.2020.107762
Zhang G, Yang Z, Huo B et al (2021) Multiorgan segmentation from partially labeled datasets with conditional NNU-Net. Comput Biol Med 136:104658. https://doi.org/10.1016/j.compbiomed.2021.104658
Zhang G, Bao C, Liu Y et al (2023) 18F-FDG-PET/CT-based deep learning model for fully automated prediction of pathological grading for pancreatic ductal adenocarcinoma before surgery. EJNMMI Res 13(1):49. https://doi.org/10.1186/s13550-023-00985-4
Zhang J, Xie Y, Wang Y et al (2021) Inter-slice context residual learning for 3D medical image segmentation. IEEE Trans Med Imaging 40(2):661–672. https://doi.org/10.1109/tmi.2020.3034995
Zhang K, Liu D (2023) Customized segment anything model for medical image segmentation. Preprint at arXiv:2304.13785
Zhang Y, Shen Z (2024) Unleashing the potential of sam2 for biomedical images and videos: A survey. Preprint at arXiv:2408.12889
Zhang Y, Wu J, Liu Y et al (2021) A deep learning framework for pancreas segmentation with multi-atlas registration and 3D level-set. Med Image Anal 68:101884. https://doi.org/10.1016/j.media.2020.101884
Zhang Y, Liang Y, Ding J et al (2022) A prior knowledge-guided, deep learning-based semiautomatic segmentation for complex anatomy on magnetic resonance imaging. Int J Radiat Oncol Biol Phys 114(2):349–359. https://doi.org/10.1016/j.ijrobp.2022.05.039
Zhang Y, Yang Y, Chen S et al (2023) Clinical application of 3D reconstruction in pancreatic surgery: a narrative review. J Pancreatol 6(01):18–22
Zhao C, Duan Y, Yang D (2022) Contour interpolation by deep learning approach. J Med Imaging. https://doi.org/10.1117/1.jmi.9.6.064003
Zhao Y, Li J, Hua Z (2022) MPSHT: multiple progressive sampling hybrid model multi-organ segmentation. IEEE J Translat Eng Health Med 10:1–9. https://doi.org/10.1109/jtehm.2022.3210047
Zheng H, Qian L, Qin Y et al (2020) Improving the slice interaction of 2.5d CNN for automatic pancreas segmentation. Med Phys 47(11):5543–5554. https://doi.org/10.1002/mp.14303
Zheng Y, Luo J (2023) Extension-contraction transformation network for pancreas segmentation in abdominal CT scans. Comput Biol Med 152:106410. https://doi.org/10.1016/j.compbiomed.2022.106410
Zhou Z, Siddiquee MMR, Tajbakhsh N et al (2020) UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging 39(6):1856–1867. https://doi.org/10.1109/tmi.2019.2959609
Zhou Z, Bian Y, Pan S et al (2023) A dual branch and fine-grained enhancement network for pancreatic tumor segmentation in contrast enhanced CT images. Biomed Signal Process Control 82:104516. https://doi.org/10.1016/j.bspc.2022.104516
Zhu J, Qi Y, Wu J (2024) Medical sam 2: Segment medical images as video via segment anything model 2. Preprint at arXiv:2408.00874
Zhu JY, Park T, Isola P, et al (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, pp 2223–2232, https://doi.org/10.1109/iccv.2017.244
Zhu Q, Li L, Hao J et al (2020) Selective information passing for MR/CT image segmentation. Neural Comput Appl 35(18):13007–13020. https://doi.org/10.1007/s00521-020-05407-3

Zhu Y, Hu P, Li X et al (2022) Multiscale unsupervised domain adaptation for automatic pancreas segmentation in CT volumes using adversarial learning. Med Phys 49(9):5799–5818. https://doi.org/10.1002/mp.15827

Zhu Y, Hu P, Li X et al (2023) An end-to-end data-adaptive pancreas segmentation system with an image quality control toolbox. J Healthc Eng 2023:1–12. https://doi.org/10.1155/2023/3617318

Zou L, Cai Z, Qiu Y et al (2023) CTG-Net: an efficient cascaded framework driven by terminal guidance mechanism for dilated pancreatic duct segmentation. Phys Med Biol 68(21):215006. https://doi.org/10.1088/1361-6560/acf110

Şolea SF, Brisc MC, Orăşeanu A et al (2024) Revolutionizing the pancreatic tumor diagnosis: emerging trends in imaging technologies: a systematic review. Medicina 60(5):695. https://doi.org/10.3390/medicina60050695

Authors and Affiliations

Andrea Moglia[1] · Matteo Cavicchioli[1,2] · Luca Mainardi[1] · Pietro Cerveri[1,3]

✉ Andrea Moglia
andrea.moglia@polimi.it

Matteo Cavicchioli
matteo.cavicchioli@polimi.it

Luca Mainardi
luca.mainardi@polimi.it

Pietro Cerveri
pietro.cerveri@unipv.it; pietro.cerveri@polimi.it

[1] Department of Electronics, Information, and Bioengineering, Politecnico di Milano, 34 Giuseppe Ponzio St., 20133 Milan, Lombardy, Italy

[2] Fondazione AIMS Academy, Ospedale Niguarda, Piazza dell'Ospedale Maggiore 3, 20162 Milan, Lombardy, Italy

[3] Department of Industrial and Information Engineering, University of Pavia, Via Adolfo Ferrata 5, 27100 Pavia, Lombardy, Italy