PanSAM: Zero-Shot, Prompt-Free Pancreas Segmentation in CT Imaging

Abolfazl Malekahmadi [1] **MohammadTaha TeimuriJervakani** [1] **Armin Behnamnia** [1] **Zahra Dehghanian** [1]
Amir Shamloo [1] **Hamid R. Rabiee** [1]

Abstract

Segmentation of the pancreas in CT images is crucial in multiple pancreatic diagnostic tasks, such as the detection, classification, and prognosis of pancreatic cancer. We present a segmentation model to find pancreatic tissue accurately in abdominal CT images. We utilize the Segment-Anything Model (SAM), a prompt-based 2D segmentation transformer model, and adapt it to 3D CT images to build a model that can segment the pancreas automatically without any prompts. To our knowledge, this is the first prompt-free work to segment the pancreas on a CT image based on the generalizable SAM model. We achieve a DICE score of 87.01% and a Jaccard score of 81.42% on the NIH dataset. We also performed zero-shot segmentation on the Abdominal-1K dataset. We achieved a DICE score of 83.20%, which shows the generalizability and applicability of our method to new unseen samples. Our study put together the zero-shot performance of SAM and the 3D nature of CT images to provide an automatic, real-time model that provides consistent segmentation throughout CT slices without the need for expert intervention. Our code is available at:
https://github.com/teimuri/PSDDSAM

1. Introduction

The pancreas is a crucial organ in the human body, and its segmentation represents one of the most challenging tasks in medical image segmentation (Roth et al., 2015). In addition, the soft tissue low contrast problem in CT scans amplifies the challenges with the pancreas (Zheng & Luo, 2023). Our work introduces PanSAM, a novel prompt-free model based on SAM for pancreas segmentation in CT images that demonstrates superior performance even on out-of-domain data samples (Kirillov et al., 2023). We design a novel point prompt generator that extracts prompts from the output of an arbitrary segmentation model. This is the first study to evaluate zero-shot learning capabilities in pancreatic segmentation from out-of-domain CT images. The high zero-shot performance of our model, which is discussed and experimentally shown in B.4, indicates the generalizability of our method. This is because of the choice of SAM as our backbone, which shows exceptional zero-shot performance. Keeping the encoder part of SAM unchanged prevents overfitting to noise in the dataset and allows rich pre-trained features to participate in solving the segmentation task.

2. Related Work

Significant progress has been made recently in developing deep learning methods to enhance pancreatic segmentation in medical imaging. These advances are vital to improving the diagnostic accuracy in this field. Pancreas segmentation techniques are broadly categorized into two approaches: single-stage methods and coarse-to-fine methods.

Single-stage: These methods directly utilize full-resolution images to define the pancreas mask as the output. Fu et al. (2018) developed a model that leverages Richer Convolutional Features (RCF) to enhance the utilization of textures of various sizes, significantly improving segmentation accuracy with a DICE score of 76.36%. Nishio et al. (2020) employed the widely recognized U-net architecture, optimized with data augmentation strategies to handle segmentation directly, bypassing the need for image cropping. Also, Fang et al. (2019) presents a 3D method for pancreas segmentation, utilizing a Progressive Fusion Network that merges local 3D contextual details with a Global Guidance Branch. This approach achieves an 85.5% DICE score. However, a significant challenge persists across these methodologies: the imbalance learning problem. Given that the pancreas typically accounts for less than 5% of the image slices and occupies only a tiny portion of each slice, models frequently face challenges with accuracy and consistency (Deng et al., 2023). This significant data imbalance poses a critical challenge for single-stage segmentation techniques, complicating the

*Equal contribution [1]Sharif University of Technology. Correspondence to: Hamid R. Rabiee <rabiee@sharif.edu>.

Published at ICML 2024 Workshop on Foundation Models in the Wild.

achievement of reliable segmentation results.

Coarse-to-Fine: Conversely, the coarse-to-fine methods employ a two-stage approach, initially identifying a broader region likely containing the pancreas and refining this detection to delineate a detailed mask. (Man et al., 2019) combine reinforcement learning with a deformable U-Net for precise localization and segmentation, achieving a DICE score of 86.9%. (Zhao et al., 2019) use a U-Net model for initial coarse segmentation of down-sampled volumes, followed by a refined analysis of localized regions. (Zhang et al., 2021) start with a multi-atlas 3D diffeomorphic registration for coarse segmentation, then apply 2D and 3D CNNs for fine details, and a 3D level-set method for final refinement, achieving an 84.61% DICE score . (Qiu et al., 2023b) develop CMFCUNet, which identifies the pancreas area and then applies detailed segmentation enhanced by a Dual Enhancement Module for effective multi-scale feature calibration. This hierarchical strategy can be implemented with varying degrees of human involvement, from manual interventions to fully automated processes. It requires substantial effort and expertise, making it less scalable. Automatic preliminary stages like (Zhou et al., 2016), although less labor intensive, introduce the risk of cascade errors, where inaccuracies in the first stage spread and adversely affect the precision of detailed segmentation in the subsequent stage.

3. Method

One of the primary challenges many models encounter is generalizability. Previous models were trained and evaluated on a single dataset, often failing to achieve satisfactory results on out-of-domain data, which is a critical issue for real-world applications. To address this problem, we use the Segment-Anything Model (SAM) (Kirillov et al., 2023) as our base architecture, a vision transformer that can provide a wide range of image segmentation based on given prompts in the form of a bounding box, point, and mask. SAM comprises an image encoder, a prompt encoder, and a mask decoder.

Although the SAM model offers generalizability and robust results, it faces significant limitations for our task. SAM is designed for 2D images, whereas each CT image consists of multiple slices that form a 3D structure. This mismatch in dimensionality poses a significant challenge. SAM is also a prompt-based model optimized for RGB images, which excels with appropriate prompts. However, its performance significantly diminishes in a prompt-free setup, and locating the pancreas in non-contrast CT images as a prompt is time-consuming. These challenges underscore the need for a new, more tailored solution for pancreas segmentation.

To address these challenges, we present the PanSAM model, a novel approach to pancreas segmentation. PanSAM is a coarse-to-fine model that comprises three key components: a prompt generator, a prompt-based segmentor, and a 3D aggregator. The process begins with the prompt generator, which identifies the pancreatic area within the image and then generates point prompts from this initial estimate. The prompt-based model then uses these prompts to conduct detailed per-slice pancreas segmentation, offering a unique and innovative solution to the pancreas segmentation problem.

In the final stage, the segmentations of each slice are processed by the 3D aggregator. This component enhances the model's accuracy by applying corrections across slices, mainly focusing on slices where the pancreas is difficult to detect individually but can be inferred from neighboring slices. The architecture of our model is illustrated in Figure 1.

The prompt generator and the prompt-based segmentor use a shared encoder in our model. Since most parameters are in the SAM encoder, using a single encoder significantly reduces the computational overhead compared to two separate SAM models. The image is encoded once using the shared SAM encoder, and encoded features are used in both the prompt generator and the prompt-based model.

The training procedure is as follows. First, the prompt generator is trained. Then, the prompt-based model is trained with the prompt generator's parameters frozen. In the last step of the training, the 3D aggregator is trained while the other two models' parameters are frozen.

These parameters freeze during training, making our model more generalizable. The model is less biased toward the noise in the training dataset, and high-level visual features learned from the large-scale dataset on which SAM is trained are kept. In the rest of this section, we explain each of these modules in more detail.

We train the PanSAM model using a combination of focal loss (Lin et al., 2018) and DICE loss (Sudre et al., 2017). The final loss function is as follows.

$$L = \sum_{i=1}^{n} L_f(f(x_i), y_i) + \beta L_D(f(x_i), y_i) \qquad (1)$$

where L_f is the focal loss between the model's output and ground truth segmentation mask and L_D is the DICE loss.

3.1. Mask Prompt Generation

To create point prompts, the input image and the predicted mask from the trained prompt-free model are fed into the prompt generator model, which samples points based on the output probabilities generated by its operation. The prompt generator is designed to produce two prompts: foreground and background.

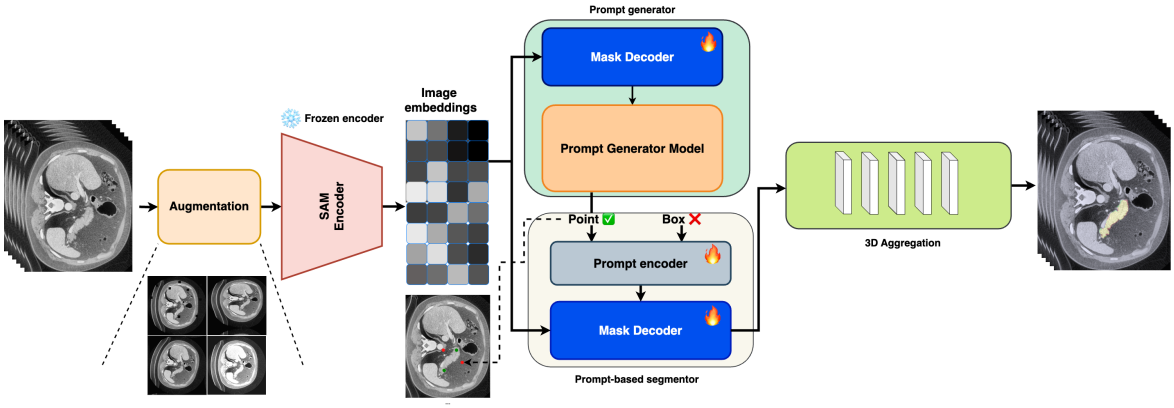

Figure 1. The architecture of PanSAM, which consists of a prompt-generator model, a prompt-based per-slice segmentor, and a 3D convolutional aggregator.

For foreground prompts, the process begins by selecting the regions in which the previous model was more certain. Then, to enhance the reliability of these prompts, a series of morphological operations is applied to the binary mask. This step helps reduce noise and eliminate the boundaries of the segmented regions, thereby ensuring that the prompts are extracted from the core of the estimated pancreatic area with greater confidence.

For background prompts, our goal is to find difficult but confident points in the image. As there are many candidate points for the background, and most of them are trivial, an algorithm is needed to find more informative background prompts. To achieve this, we create two different masks and sample background prompts from their intersection. The first mask is designed to closely approximate the borders of the pancreas, capturing maximal information from these critical transitional areas. The second mask selectively includes points where the model demonstrates high confidence in its predictions, ensuring the relevance and reliability of the prompts. Section 3.4 provides more details on these operations and the methodologies employed.

3.2. Prompt-based Segmentation

To create pancreas segmentation on CT images, we rely on the prompts generated in the prompt-creation phase and input them into the prompt-based segmentation model. This model efficiently estimates the pancreas segmentation by utilizing the pre-computed image embedding as encoded features and the generated prompts. By the end of this stage, the model can estimate the pancreas region by processing each slice of the CT image separately without the need for any manual prompt. Figure 2 clearly illustrates the efficiency of the prompt-based model compared to the SAM model, which operates without any estimated prompt. More comparisons can be found in Figure 5 in the appendix.

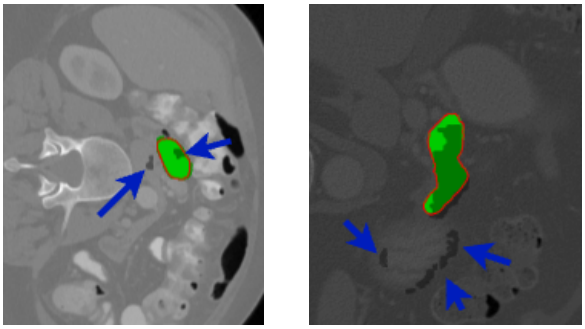

Figure 2. The image illustrates the label delineated by a red line, while the output from the prompt-based module is highlighted in glowing green, and the darker parts indicate the results from the prompt-generator module. Blue arrows demonstrate the impact of the prompt-based segmentor module, which significantly enhances the accuracy of corrections and segmentations, as evident in the image.

3.3. 3D Aggregation

After creating per-slice segmentations, we give the pixel-wise scores of the prompt-based model as input to the 3D aggregator model. The 3D aggregator aims to merge the per-slice predictions and improve the final result. It consists of several 3D convolution layers that use nearby slice information to predict a specific slice. The effect of 3D aggregation is illustrated in Figure 3. Our results demonstrate that 3D aggregation has several benefits. One is smooth, continuous predictions as we move through the slices. Another benefit is the correction in the starting and ending slices where the pancreas begins to appear and disappears in the per-slice view. The output of the 3D aggregator will be the final prediction of our model. More results show these effects can be found in Figure 3. In the appendix, more pairs can be found in Figure 8.

3.4. Details of Point Prompt Generation

For foreground prompts, after thresholding with 0.5 (which is used to create the model output's binary mask), in order to create more confident prompts and remove noisy segmentation and the border of the segmented region, we apply these operations on the output mask of a without prompt trained SAM. To extract prompts from inside the estimated pancreas area, we apply this operation:

$$M_f = dilate(erode(th(O, 0.5), (10, 10)), (5, 5)),$$

where O is the output probabilities of the prompt generator model, th is the thresholding function, and erode and dilate apply the operations given an input image and a tuple of size 2 as kernel size. We uniformly sampled from this new mask to create foreground prompts. We also choose two points from the border of the thresholded mask as additional foreground prompts.

For background prompts, our goal is to find difficult but confident points in the image. As there are many candidate points for the background, and most of them are trivial, an algorithm is needed to find more informative background prompts. To achieve this, we create two different masks and sample background prompts from their intersection. The first mask is the complement of a dilated version of the model's prediction of the pancreas region:

$$M_b^1 = 1 - dilate(th(O, 0.5), (5, 5)).$$

This mask indicates pixels that are not in the region of the detected pancreas. The second mask is created by thresholding the model's prediction by a lower value:

$$M_b^2 = 1 - th(O, 0.4),$$

which indicates the areas where the model is fairly confident that they do not contain the pancreas tissue. The final background mask is formed by the intersection of the above-mentioned masks. We sample from M_b with probabilities proportional to the model's output probabilities:

$$P(x, y) = \frac{(M_b \times O)_{xy}}{\sum_{x',y'} (M_b \times O)_{x'y'}} \tag{2}$$

Our prompt sampler tries to avoid finding incorrect background or foreground points, caused by the prompt-generator's incorrect detections while trying to sample from the points that are hard to detect. We also add two random points from the background of the 0.5-thresholded mask as trivial background prompts.

4. Experiments

In our research, we utilized the ViT-B version of the Segment-Anything Model (SAM) for efficient pancreas segmentation in CT images, applying two innovative methods

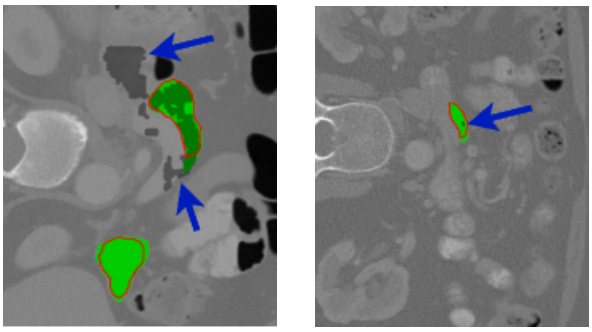

Figure 3. In the image, the label is outlined with a red line, while the output from the 3D aggregator module is displayed in glowing green. The results of the segmentator model appear in the darker areas, highlighted with blue arrows for clearer visibility. The effectiveness of the 3D aggregator module is particularly notable in regions where the pancreas is bifurcated or at the initial and terminal segments, areas that are typically challenging to identify.

to prove its clinical utility. Initially, we fine-tuned SAM's mask decoder without prompts, reaching a score of 83.75% for the DICE value. Then, use our full approach to refine segmentation, achieving an improved DICE coefficient of 87.01% on NIH Pancreas-CT dataset (Roth et al., 2016) seen in table 1. This result shows an increase in the accuracy and stability of our model. Furthermore, to improve the robustness of our model in diverse conditions, we implemented various augmentations; These augmentations bolstered our model's strength and resilience, ensuring its efficacy across a spectrum of scenarios. Additionally, the incorporation of these enhancements further underscores the versatility and reliability of our segmentation approach. To find further experiments in different setups, ablation study of our approach, zero-shot benchmarking on AbdomenCT-1K dataset and computational cost, check the appendix B. Our segmentation model demonstrated superior performance, particularly in challenging areas such as where the pancreas is divided into distinct regions, at the organ's extremities, or at the initial or terminal slices of the pancreas, as illustrated in Figure 8 in Appendix A.

4.1. Experiment Details

Dataset We analyzed the pancreas in abdominal CT scans using two datasets: NIH Pancreas-CT (Roth et al., 2016) and AbdomenCT-1K (Ma et al., 2021). The NIH Pancreas-CT dataset contains 82 contrast-enhanced abdominal CT volumes, each presenting a variety of longitudinal slices. This dataset is particularly designed for in-depth pancreatic studies. In contrast, the AbdomenCT-1K dataset broadens our scope to multi-organ segmentation, containing over 1000 CT scans sourced from 12 medical centers. These scans cover a variety of phases, equipment vendors and diseases, with annotations for the liver, kidney, spleen, and pancreas

Table 1. Comparison of PanSAM with Single-Stage and Coarse-to-Fine Models

	Dimension	Name	DICE	Recall	Jaccard	Precision	F1-Score
Single-Stage	2D	Fu et al (Fu et al., 2018)	76.36±14.34	79.12±16.27	63.72±17.05	77.36±17.96	78.23
	2D	Nishio (Nishio et al., 2020)	78.90±8.60	76.00±12.00	65.00±10.00	**100**	86.36
	2D	Jinzheng Cai (Cai et al., 2017)	82.40±6.70	–	70.60±9.00	–	–
	2D	Huang et al(Huang & Wu, 2022)	82.87±1.00	77.37±1.41	70.97 ± 1.39	89.29 ± 0.98	82.90
	2D	Oktay et al (Oktay et al., 2018)	82.25±4.33	82.81±6.42	–	82.41±7.01	82.60
	3D	Zhang et al (Zhang & Bagci, 2022)	85.50±3.70	88.20±4.00	–	84.00±8.30	86.05
	3D	Fang et al. (Fang et al., 2019)	85.50±4.80	–	–	–	–
Coarse-to-Fine	2D	BRIEFnet (Heinrich & Oktay, 2017)	64.00±4.00	–	–	–	–
	2D	Deng et al (Deng et al., 2023)	81.70±9.41	87.10±8.47	–	75.52±6.45	80.90
	2D	Li et al. (Li et al., 2022)	82.80±6.30	–	–	82.51±1.03	–
	2D	Zhou et al(Zhou et al., 2017)	83.20±4.80	–	–	–	–
	2D	RTUNet (Qiu et al., 2023a)	86.25±4.52	–	–	–	–
	2D	CKS (Tang et al., 2023)	85.42±4.39	–	–	–	–
	2D	CMFCUNet (Qiu et al., 2023b)	86.30±4.03	86.85±5.23	76.26±5.01	85.91±3.97	86.38
	2D	Man et al (Man et al., 2019)	86.93±4.92	86.91±4.85	–	–	–
	2.5D	Yan et al (Yan & Zhang, 2021)	86.61±3.47	–	–	–	–
	2.5D	PBR-UNet (Li et al., 2021)	85.35±4.13	–	–	–	–
	3D	Zhao et al (Zhao et al., 2019)	85.90±4.51	–	–	–	–
	3D	Zhang et al (Zhang et al., 2021)	84.61±5.21	–	–	–	–
	3D	ECTN (Zheng & Luo, 2023)	85.58±3.98	85.11±5.96	74.99±5.86	86.59±6.14	85.84
	3D	**PanSAM**	**87.01±1.93**	**89.38±0.73**	**81.42±2.29**	89.86±2.54	**89.45**

(Ma et al., 2021). We used the NIH dataset both to train our model and to compare our work with other related studies. In contrast, the AbdomenCT-1K dataset was used for external validation.

Preprocessing

The preprocessing of abdominal CT images plays a pivotal role in enhancing model performance. To achieve optimal results, we employ Contrast-limited Adaptive Histogram Equalization (CLAHE) (Mishra, 2021). CLAHE works by normalizing the image pixel values based on their local histogram, effectively improving the contrast of the images.

Our preprocessing pipeline involves several key steps. First, each slice of the CT image is normalized to a range between [0, 1]. Next, we apply CLAHE to adjust the contrast. Finally, we normalize each slice once again to ensure that the pixel values remain within the range of [0, 1]. This dual normalization process, as shown in Figure 4, significantly enhances the clarity and quality of the CT images, thus improving the overall performance of the model.

We also improved the robustness of PanSAM by using the Albummentations library for image enhancement, which simulates a variety of medical imaging conditions (Buslaev et al., 2020). In this experiment, we introduced rotational variation with a rotation of 30 degrees and a 50% chance of applying it. The brightness and contrast of each image were perturbed by up to 30%. By randomly scaling the image size to 90-110% of the original images and then resizing them to 1024x1024 pixels, we maintain the required size while also introducing further enhancement. We also use CLAHE as an augmentation with a clipping threshold of 2 and a tile grid size of 8×8. CLAHE was applied with a probability of 0.5 to each image during training. With a 50% chance, up to 8 random square mask regions with height and width in the range of $[8, 16]$ pixels were applied to simulate artifacts. Also, with a probability of 0.5, the images were randomly scaled with a limit of 10%. The diversity of augmentations exposed the model to various clinically relevant image changes.

5. Conclusion

In this study, we proposed PanSAM, a prompt-free model that utilizes the strong Segment-Anything transformer model to extract the pancreas from 3D CT images. During training, we applied extensive data augmentation to ensure our model's robustness across different imaging sources and conditions. This strategy significantly enhances the model's ability to generalize and perform well on various datasets. To evaluate its generalizability, we tested the trained model on a different dataset without any additional training. The zero-shot performance of the model in segmenting pancreatic tissue was impressive, demonstrating its applicability in situations where the model must be used in data from other sites or hospitals. The success of our model in external validation (Abdominal-1K dataset) further underscores its potential as a versatile and effective tool for clinical application in various imaging conditions. This provides confidence in its deployment in real-world diagnostic settings, where the model must perform accurately on data it has not been explicitly trained on.

References

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A. Albumentations: fast and flexible image augmentations. *Information*, 11(2): 125, 2020.

Cai, J., Lu, L., Xie, Y., Xing, F., and Yang, L. Improving deep pancreas segmentation in ct and mri images via recurrent neural contextual learning and direct loss function. *arXiv preprint arXiv:1707.04912*, 2017.

Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

Deng, Y., Lan, L., You, L., Chen, K., Peng, L., Zhao, W., Song, B., Wang, Y., Ji, Z., and Zhou, X. Automated ct pancreas segmentation for acute pancreatitis patients by combining a novel object detection approach and u-net. *Biomedical Signal Processing and Control*, 81:104430, 2023.

Fang, C., Li, G., Pan, C., Li, Y., and Yu, Y. Globally guided progressive fusion network for 3d pancreas segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pp. 210–218. Springer, 2019.

Fu, M., Wu, W., Hong, X., Liu, Q., Jiang, J., Ou, Y., Zhao, Y., and Gong, X. Hierarchical combinatorial deep learning architecture for pancreas segmentation of medical computed tomography cancer images. *BMC systems biology*, 12:119–127, 2018.

Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.-A., and Dou, Q. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. *arXiv preprint arXiv:2306.13465*, 2023.

Heinrich, M. P. and Oktay, O. Briefnet: Deep pancreas segmentation using binary sparse convolutions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 329–337. Springer, 2017.

Huang, M.-L. and Wu, Y.-Z. Semantic segmentation of pancreatic medical images by using convolutional neural network. *Biomedical Signal Processing and Control*, 73: 103458, 2022.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything, 2023.

Li, J., Lin, X., Che, H., Li, H., and Qian, X. Pancreas segmentation with probabilistic map guided bi-directional recurrent unet. *Physics in Medicine & Biology*, 66(11): 115010, 2021.

Li, M., Lian, F., and Guo, S. Multi-scale selection and multi-channel fusion model for pancreas segmentation using adversarial deep convolutional nets. *Journal of Digital Imaging*, 35(1):47–55, 2022.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. 2018.

Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2021.

Man, Y., Huang, Y., Feng, J., Li, X., and Wu, F. Deep q learning driven ct pancreas segmentation with geometry-aware u-net. *IEEE transactions on medical imaging*, 38 (8):1971–1980, 2019.

Mishra, A. Contrast limited adaptive histogram equalization (clahe) approach for enhancement of the microstructures of friction stir welded joints. *arXiv preprint arXiv:2109.00886*, 2021.

Nishio, M., Noguchi, S., and Fujimoto, K. Automatic pancreas segmentation using coarse-scaled 2d model of deep learning: usefulness of data augmentation and deep u-net. *Applied Sciences*, 10(10):3360, 2020.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Qiu, C., Liu, Z., Song, Y., Yin, J., Han, K., Zhu, Y., Liu, Y., and Sheng, V. S. Rtunet: Residual transformer unet specifically for pancreas segmentation. *Biomedical Signal Processing and Control*, 79:104173, 2023a.

Qiu, C., Song, Y., Liu, Z., Yin, J., Han, K., and Liu, Y. Cmfcunet: cascaded multi-scale feature calibration unet for pancreas segmentation. *Multimedia Systems*, 29(2): 871–886, 2023b.

Roth, H., Farag, A., Turkbey, E. B., Lu, L., Liu, J., and Summers, R. M. Data From Pancreas-CT. 2016. doi: 10.7937/K9/TCIA.2016.TNB1KQBU.

Roth, H. R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E. B., and Summers, R. M. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International*

Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18, pp. 556–564. Springer, 2015.

Smith, L. N. and Topin, N. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pp. 369–386. SPIE, 2019.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. pp. 240–248, 2017. ISSN 1611-3349. doi: 10.1007/978-3-319-67558-9_28.

Tang, Y., Zhan, K., Tian, Z., Zhang, M., Wang, S., and Wen, X. Curriculum knowledge switching for pancreas segmentation. *arXiv preprint arXiv:2306.12651*, 2023.

Yan, Y. and Zhang, D. Multi-scale u-like network with attention mechanism for automatic pancreas segmentation. *Plos one*, 16(5):e0252287, 2021.

Zhang, Y., Wu, J., Liu, Y., Chen, Y., Chen, W., Wu, E. X., Li, C., and Tang, X. A deep learning framework for pancreas segmentation with multi-atlas registration and 3d level-set. *Medical Image Analysis*, 68:101884, 2021.

Zhang, Z. and Bagci, U. Dynamic linear transformer for 3d biomedical image segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pp. 171–180. Springer, 2022.

Zhao, N., Tong, N., Ruan, D., and Sheng, K. Fully automated pancreas segmentation with two-stage 3d convolutional neural networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pp. 201–209. Springer, 2019.

Zheng, Y. and Luo, J. Extension–contraction transformation network for pancreas segmentation in abdominal ct scans. *Computers in Biology and Medicine*, 152:106410, 2023.

Zhou, Y., Xie, L., Shen, W., Fishman, E., and Yuille, A. Pancreas segmentation in abdominal ct scan: a coarse-to-fine approach. *arXiv preprint arXiv:1612.08230*, 10, 2016.

Zhou, Y., Xie, L., Shen, W., Wang, Y., Fishman, E. K., and Yuille, A. L. A fixed-point model for pancreas segmentation in abdominal ct scans. In *International conference on medical image computing and computer-assisted intervention*, pp. 693–701. Springer, 2017.

A. Image Results

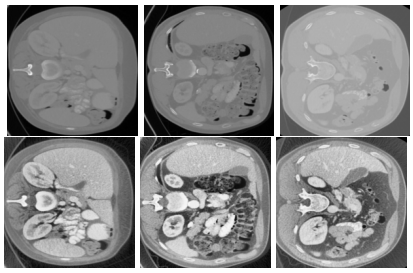

Figure 4. The first row displays the original abdominal CT images prior to applying the preprocessing pipeline, showcasing lower contrast and clarity. The second row demonstrates the same CT images after preprocessing, where Contrast-limited Adaptive Histogram Equalization (CLAHE) has been applied. This preprocessing significantly enhances image contrast and detail, making anatomical structures more discernible

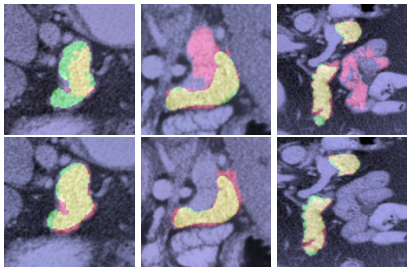

Figure 5. Comparison of the fine-tuned SAM without prompts and the PanSAM model for estimating the pancreatic region. Each image pair displays the output of the fine-tuned SAM at the top and the PanSAM model's output below. Although the PanSAM model leverages outputs from the fine-tuned SAM to generate its prompts, it effectively amends errors from the initial SAM model, resulting in more precise pancreas segmentation. In the visualizations, red signifies model predictions, green represents the ground truth labels, and yellow illustrates the overlap between the predictions and the ground truth.

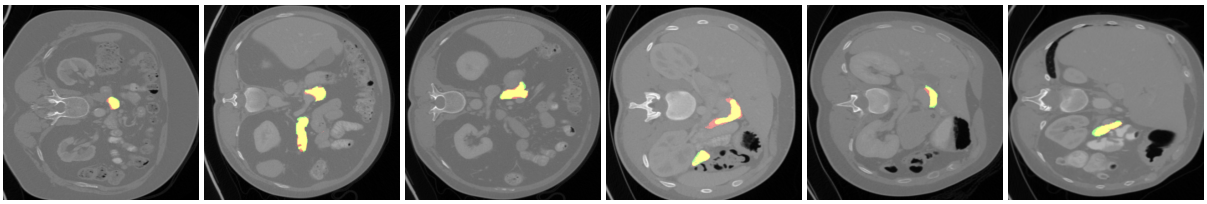

Figure 6. Model efficacy with challenging and complex samples. In the images, red signifies model predictions, green represents ground truth labels, and yellow marks their intersection.

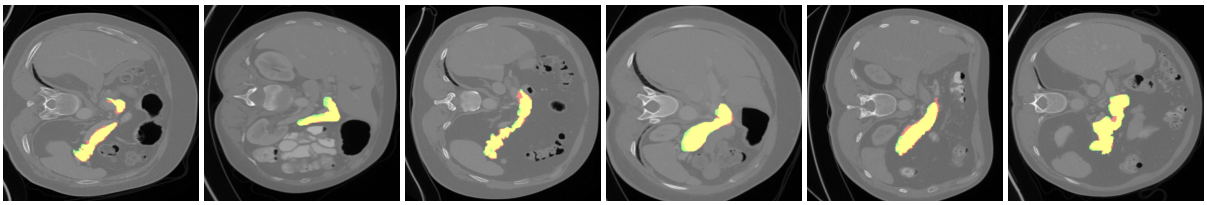

Figure 7. Performance of the model across various samples. In the images, red signifies model predictions, green represents ground truth labels, and yellow marks their intersection.

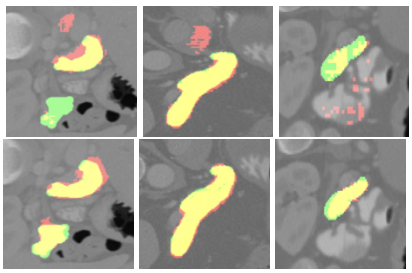

Figure 8. Comparison between the PanSAM(2D) and PanSAM(3D) models, demonstrating the impact of the 3D aggregator in enhancing pancreatic segmentation. The PanSAM(3D) model uniquely incorporates a 3D aggregator that utilizes information from neighboring slices to improve accuracy. This addition allows the PanSAM(3D) to achieve more precise segmentation, particularly in challenging areas such as where the pancreas is divided into distinct regions, at the organ's extremities, or at the initial parts of the organ. In contrast, the PanSAM(2D) model does not include this feature. In each pair of images, the top one shows the output from the PanSAM(2D) model, while the bottom one illustrates the output from the PanSAM(3D) model, highlighting the enhanced segmentation capabilities provided by the 3D aggregator. Each color signifies the same meaning mentioned in Figure 5.

B. Further Experiments

B.1. Experiments Setup

In our experiments, we used an Adam optimizer with a learning rate of 5×10^{-4} and weight decay of 5×10^{-5}. We also set $\beta = 0.05$, the weight of the DICE loss. We used a batch size of 8 for training. Additionally, a OneCycleLR scheduler was implemented for fast convergence and to avoid overfitting, ensuring accuracy and efficiency(Smith & Topin, 2019).

B.2. Ablation Study

In this section, we present an ablation study to assess the contribution of individual modules within our architecture. Table 2 delineates the impact of each module, illustrating how they influence the overall performance of the system. The results clearly demonstrate that each component plays a significant role in improving the efficacy of the final model.

Table 2. Ablation Study on PanSAM Model

Approaches	DICE	Recall	Jaccard	Precision	F1-Score
SAM	17.57±7.31	68.43±6.19	11.00±6.72	55.36±8.25	60.88
fine-tuned SAM	83.31±1.27	88.32±1.66	78.51±1.59	85.84±1.8	87.11
fine-tuned SAM + 3D aggregator	83.75±1.12	88.74±1.46	79.41±1.22	86.58±1.53	87.66
fine-tuned SAM + prompt generator	84.62±1.18	88.50±1.45	80.25±1.61	88.42±1.47	88.51
PanSAM	**87.01±1.93**	**89.38±0.73**	**81.42±2.29**	**89.86±2.54**	**89.45**

B.3. Discussion

In this section, we discuss the performance of the PanSAM model in a comprehensive way. As detailed in the ablation study section B.2, each component of PanSAM plays a crucial role in enhancing the model's overall effectiveness, with the final PanSAM demonstrating superior results compared to its individual modules. Here, we specifically focus on evaluating the performance of our prompt generator.

Ideally, the most accurate method to generate prompts would involve an expert manually selecting both interior and exterior points with a high degree of confidence. However, this approach is not feasible due to the high costs associated with expert time, the time-consuming nature of the task, and the necessity for ongoing expert involvement with each model run. Consequently, we have opted to exclude human experts from the loop, aiming instead for full automation of the prompt generation process.

To assess the quality of the automatically generated prompts, we compare them against a baseline where points are directly derived from the ground truth. This comparison is quantitatively presented in Table 3, where we measure the impact of using ground truth points versus points from our automated prompt generator module. This evaluation helps us understand the efficacy of the prompt generator and its contribution to the overall performance of the PanSAM model.

Table 3. Comparison of DICE Scores between Ground Truth and Automatically Generated Prompts in PanSAM model. The GroundTruth extensions choose point prompts from real mask instead of getting it from the prompt generator.

Dimension	Name	Prompt	DICE
2D	SAM (Kirillov et al., 2023)	5 points	17.57±7.31
2D	SAM MED2D (Cheng et al., 2023)	5 points	60.43±5.42
3D	3DSAM-adapter (Gong et al., 2023)	10 points	53.12±5.11
2D	**PanSAM-GroundTruth**	**4 points**	**87.24±1.93**
3D	**PanSAM-GroundTruth**	**4 points**	**89.15±1.07**
3D	**PanSAM**	**4 points**	**87.01±1.93**

Table 3 illustrates that while our prompt generator does not achieve the same level of accuracy as expert-provided ground truth points, the difference in performance is relatively minimal, resulting in only a 2-point decrease in the DICE score. Despite this, the PanSAM model still outperforms all other state-of-the-art (SOTA) models currently available. This outcome underscores the effectiveness of our model, although it also highlights potential areas for improvement.

B.4. Zero-shot Benchmark

The robustness of PanSAM was further evaluated through a zero-shot generalization on the Abdomen1K dataset, which consists of multiorgan CT scans. For external validation purposes, we exclusively analyzed the model's ability to segment the pancreas. Interestingly, PanSAM achieved a DICE score of 83.20%, which is indicative of its powerful generalization capabilities when applied to unseen conditions.

This high level of zero-shot performance is a testament to two pivotal strategies implemented during model training. First, we used diverse data augmentation techniques to foster model adaptability to new datasets. Second, by freezing the encoder, we preserved the learned features from the initial training phase, enabling the model to maintain its learned generalization without additional training or fine-tuning.

The success of PanSAM in external validation, particularly in zero-shot scenarios, demonstrates its potential as a versatile and effective tool for clinical application across various imaging conditions, providing confidence in its deployment in real-world diagnostic settings where the model must perform accurately on data it has not explicitly been trained on.

B.5. Computational Cost

As our proposed model is multistage, the computation time of the complete pipeline can become undesirably high. In this section, we compare the size and inference time of our model with other SAM integrations. The image encoder of SAM ViT-H, which is used in our experiments, has 637M parameters, and the mask decoder has 4M parameters. In the first stage, a single pass from the image encoder and the first mask decoder takes place, with a total of 641M parameters. In the second stage, the already calculated features of the image encoder are passed to the second mask decoders, with 4M parameters. In the end, the outputs are fed into the 3D aggregator which has 14K parameters. The total number of parameters of our model is less than 646M. Using the SAM model in its original form, with manually-fed prompts, the number of parameters would be 641M. Hence, with an only additional 0.78% number of parameters compared to SAM, we omitted the requirement for manual prompts. The runtime of our model is 0.145s per slice on an RTX 3090 GPU, estimated over 10 runtimes of the model.