How Well Do Supervised 3D Models Transfer to Medical Imaging Tasks?

Wenxuan Li **Alan Yuille** **Zongwei Zhou**[*]
Johns Hopkins University
https://github.com/MrGiovanni/SuPreM

Abstract

The pre-training and fine-tuning paradigm has become prominent in transfer learning. For example, if the model is pre-trained on ImageNet and then fine-tuned to PASCAL, it can significantly outperform that trained on PASCAL from scratch. While ImageNet pre-training has shown enormous success, it is formed in 2D and the learned features are for classification tasks; when transferring to more diverse tasks, like 3D image segmentation, its performance is inevitably compromised due to the deviation from the original ImageNet context. A significant challenge lies in the lack of large, annotated 3D datasets rivaling the scale of ImageNet for model pre-training. To overcome this challenge, we make two contributions. Firstly, we construct AbdomenAtlas 1.1 that comprises 9,262 three-dimensional computed tomography (CT) volumes with high-quality, per-voxel annotations of 25 anatomical structures and pseudo annotations of seven tumor types. Secondly, we develop a suite of models that are pre-trained on our AbdomenAtlas 1.1 for transfer learning. Our preliminary analyses indicate that the model trained only with 21 CT volumes, 672 masks, and 40 GPU hours has a transfer learning ability similar to the model trained with 5,050 (unlabeled) CT volumes and 1,152 GPU hours. More importantly, the transfer learning ability of supervised models can further scale up with larger annotated datasets, achieving significantly better performance than preexisting pre-trained models, irrespective of their pre-training methodologies or data sources. We hope this study can facilitate collective efforts in constructing larger 3D vision datasets and more releases of supervised pre-trained models.

1 Introduction

Pre-training and fine-tuning is a widely adopted transfer learning paradigm (Zoph et al., 2020). Given the relationship across different vision tasks, a model pre-trained on one dataset is expected to benefit another. Over the past few decades, pre-training has been important in AI development (Radford et al., 2021; Kumar, 2017). For 2D vision tasks, there are two available options: (*i*) supervised pre-training and (*ii*) self-supervised pre-training, but for 3D vision tasks, option (*i*) is often not available simply due to the lack of large, annotated 3D volumetric datasets (Wang et al., 2022).

Supervised pre-training can learn image features that are transferable to many target tasks. It has been common practice to pre-train models using ImageNet and then fine-tune the model on target tasks that often have less training data, e.g., PASCAL. However, two challenges arise in ImageNet pre-training. Firstly, ImageNet predominantly comprises 2D images, leaving a palpable void in large-scale 3D datasets and investigation in 3D transfer learning (Huang et al., 2023). Secondly, ImageNet is intended for image classification, so the benefit for segmentation (and other vision tasks) can be somewhat compromised (He et al., 2019). If such an ImageNet-like dataset exists—formed in 3D and annotated per voxel—supervised pre-trained models are expected to transfer better to 3D image segmentation than self-supervised ones for two reasons.

1. **Supervised pre-training is more efficient in data and computation because of its explicit learning objective.** While self-supervised pre-training can learn features without manual annotation, it often requires a large corpus of datasets (Xiao et al., 2022). Extracting meaningful features

[*] Correspondence to Zongwei Zhou (zzhou82@jh.edu).

directly from raw, unlabeled data is inherently challenging. Unlabeled data have a high degree of redundancy (Haghighi et al., 2020; 2021) and noise (Mahajan et al., 2018), which can complicate the learning process. Therefore, self-supervised pre-training often calls for greater computational resources and time to match the outcomes achieved by supervised pre-training (Chen et al., 2020a; Tang et al., 2022). We have quantified the improved data and computational efficiency from perspectives of both pre-training (Figure 2a; 99.6% fewer data) and fine-tuning (Figure 2b; 66% less computation). Specifically, the model trained with 21 CT volumes, 672 masks, and 40 GPU hours shows transfer learning ability similar to that trained with 5,050 CT volumes and 1,152 GPU hours, highlighting the remarkable efficiency of supervised pre-training.

2. **Supervised pre-training enables the model to learn image features that are relevant to image segmentation.** Self-supervised pre-training must extract images features from raw, unlabeled data using pretext tasks such as mask image modeling (Zhou et al., 2021b; He et al., 2022; Tao et al., 2020; Chen et al., 2019a), instance discrimination (Xie et al., 2020; Shekoofeh et al., 2021; Chaitanya et al., 2020), etc. Despite their efficacy in pre-training, these pretext tasks share no obvious relation to the target image segmentation. In contrast, supervised pre-training uses semantically meaningful annotations (e.g., organ/tumor segmentation) as supervision, with which the model can mimic the behavior of medical professionals—identifying the edge and boundary of specific anatomical structures. As a result, the pre-training is interpretable, and the learned features are expected to be relevant to image segmentation tasks (Zamir et al., 2018; Ilharco et al., 2022; You et al., 2022). We have demonstrated that the learned features can be *direct inference* for organ segmentation on CT volumes collected from hospitals worldwide (Table 3; evaluated on three novel hospitals). The features learned by supervision can also be *fine-tuned* to perform novel class segmentation (unseen in the pre-training) with higher accuracy and less annotated data than the features learned by self-supervision (Table 4; evaluated on 63 novel classes).

This paper seeks to answer the question *how well the model transfers to 3D medical imaging tasks* IF it is pre-trained on large, annotated 3D datasets. Naturally, we start with creating an *IF* dataset at a massive scale. **Firstly**, we construct a dataset (termed AbdomenAtlas 1.1[1]) of 9,262 CT volumes with per-voxel annotations of 25 anatomical structures and pseudo annotations of seven types of tumors. This large-scale, fully-annotated dataset enables us to train models in a fully supervised manner using multi-organ segmentation as the pretext task. As reviewed in Table 1, this dataset is much more extensive (considering both the number of CT volumes and annotated classes) than public datasets (Wasserthal et al., 2022; Ma et al., 2022; Qu et al., 2023). Scaling experiments in §3.1 suggested that pre-training models on more annotated datasets can further improve the transfer learning ability. **Secondly**, we develop a suite of **Su**pervised **Pre**-trained **M**odels, termed SuPreM, that combined the good of large-scale datasets and per-voxel annotations, demonstrating the efficacy across a range of target segmentation tasks. As reported in §3.2, some of the dominant segmentation backbones have been pre-trained and will be available to the public. Current pre-trained backbones are U-Net (CNN-type) (Ronneberger et al., 2015), SegResNet (CNN-type) (Chen et al., 2016), and Swin UNETR (Transformer-type) (Tang et al., 2022), and more backbones will be added along time.

In prospective endeavors, we anticipate that the expansion of datasets and annotations will not only enhance feature learning, as demonstrated in this study, but also promote the development of advanced AI algorithms and benchmark the state of the art in terms of segmentation performance, inference efficiency, and domain generalizability.

2 Brief History: Supervised Pre-Training

In a major initiative aimed at developing widely transferable AI models—known as Foundation Models in the medical domain (Moor et al., 2023; Butoi et al., 2023; Ma & Wang, 2023a)—one faces a critical decision: *should the focus of pre-training be supervised or self-supervised?* While human annotations undeniably improve task-specific performance, such as semantic segmentation, the best strategy for learning generic image features that can be transferable across a spectrum of tasks has yet to be determined. For 2D vision tasks, the advent of ImageNet (Deng et al., 2009) makes it possible to debate the merits and limitations of supervised pre-trained models for transfer learning compared

[1] Segmentation is fundamental in the medical domain (Ma & Wang, 2023b). It can be viewed as a per-voxel classification task. Therefore, the per-voxel supervision used in our pre-training (**272.7B** annotated voxels) is much stronger than the per-image supervision used in ImageNet pre-training (**14M** images).

with self-supervised ones. We refer the readers to Yang et al. (2020) and Tendle & Hasan (2021) for a plethora of viewpoints from either side. In essence, the debates are about clarifying the learning objective (loss function) of emulating human vision (Zhou, 2021).

The learning objective of supervised pre-training is to minimize the discrepancy between AI predictions and semantic labels annotated by humans. Over the years, supervised pre-training on ImageNet has shown marked success in transfer learning (Yosinski et al., 2014). Moreover, the transfer learning ability can be further enhanced when models are trained on increasingly expansive datasets, such as ImageNet-21K (Kolesnikov et al., 2020), Instagram (Mahajan et al., 2018), JFT-300M (Sun et al., 2017), and JFT-3B (Zhai et al., 2022). In general, supervised pre-training exhibits clear advantages over self-supervised pre-training when sizable annotated datasets are available (Steiner et al., 2021; Ridnik et al., 2021). However, acquiring millions of manual annotations is labor-intensive, time-consuming, and challenging to scale—but certainly not impossible—evidenced by several recent influential endeavors (Kuznetsova et al., 2020; Mei et al., 2022; Kirillov et al., 2023; Bai et al., 2023).

On the other hand, self-supervised pre-training offers an alternative by enabling AI models to learn from raw, unlabeled data (Jing & Tian, 2020; Zoph et al., 2020; Ren et al., 2022; 2023), thus reducing the need for manual annotation. Self-supervised pre-training has historically lagged behind the state-of-the-art supervised pre-training in ImageNet benchmarks (Pathak et al., 2016; Noroozi & Favaro, 2016). The recent pace of progress in self-supervised pre-training has yielded models whose performance not only matches but, at times, surpasses those achieved by supervised pre-training (Chen et al., 2020a; Grill et al., 2020; Chen et al., 2020b; Zhou et al., 2021a; Wei et al., 2022). This has raised hopes that self-supervised pre-training could indeed replace the ubiquitous supervised pre-training in advanced computer vision going forward. The caveat, however, is the significant demand for both data and computational power, often exceeding the resources available in academic settings. For example, He et al. (2020) have demonstrated that self-supervised features trained on 1B images (a factor of 714× larger) can transfer comparably or better than ImageNet features.

Supervised pre-training on ImageNet has demonstrated benefit for 2D medical image tasks after transfer learning (Tajbakhsh et al., 2016; Shin et al., 2016; Zhou et al., 2017). Unfortunately, it has been constrained for 3D medical imaging tasks due to the lack of a 3D counterpart to ImageNet. Although there are a great number of raw, unlabeled medical images available (Team, 2011; Baxter et al., 2023; Zhao et al., 2023; Saenz et al., 2024), annotating these images is a labor-intensive undertaking for professionals. Our contribution to a large, annotated 3D dataset could spark the debate of whether self-supervised or supervised pre-training leads to better performance and data/computational efficiency, which would not be possible without the invention of a dataset of such scale.

3 Material & Method

We constructed an AbdomenAtlas 1.1 dataset comprising 9,262 three-dimensional CT volumes and over 300K masks spanning 25 anatomical structures and seven types of tumors. In addition, we released a suite of supervised pre-trained models (SuPreM) to benefit 3D medical imaging tasks.

3.1 Extensive Dataset: AbdomenAtlas 1.1

Interactive segmentation, an integration of AI algorithms and human expertise, was used to create AbdomenAtlas 1.1 in a semi-automatic procedure. We recruited a team of ten radiologists to perform manual annotations to ensure the annotation quality[2]. Given the complexity of 3D data, rather than annotating the entire dataset voxel by voxel, we asked the radiologists to focus on the most important CT volumes and regions therein. In doing so, an importance score for each volume was computed, derived from the uncertainty, consistency, and overlap (Qu et al., 2023). Six junior radiologists revised the annotations predicted by AI under the supervision of four senior radiologists, and in turn, AI improved its predictions by learning from these revised annotations. This interactive procedure continued to enhance the quality of annotations until no major revision was required from the radiologists. Subsequently, four senior radiologists went through the final visualizations for all the annotations, detecting and revising major errors as needed before the dataset was released. Annotation tools employed included a licensed version from Pair and an open-source MONAI Label.

[2] Ensuring high-quality annotations is costly and time-consuming, yet it is critical for transfer learning, as quantified in Appendix B.4, and for reducing ambiguity when training AI models for image segmentation.

Table 1: **Contribution #1: An extensive dataset of 9,262 CT volumes with per-voxel annotations of 25 anatomical structures.** This dataset is unprecedented in terms of data and annotation scales, providing over 300K organ/tumor masks and 3.7M annotated images that are taken from 88 hospitals worldwide. In 2009, before the advent of ImageNet (Deng et al., 2009), it was challenging to empower an AI model with generalized image representation using a small or even medium size of labeled data, the same situation, we believe, that presents in 3D medical image analysis today. As seen in the table, the annotations of public datasets are limited, partial, and incomplete, and the CT volumes in these datasets are often biased toward specific populations, medical centers, and countries. Our constructed dataset mitigates these gaps, representing a significant leap forward in the field. The CT volumes in datasets 1–17 are used to construct AbdomenAtlas 1.1. Detailed information can be found in Appendix B.1, and the domain gap across these datasets is illustrated in Appendix B.2.

dataset (year) [source]	# of class	# of† volume	# of center	dataset (year) [source]	# of class	# of† volume	# of center
1. Pancreas-CT (2015) [link]	1	42	1	2. BTCV (2015) [link]	12	47	1
3. AbdomenCT-1K (2021) [link]	4	1,050	12	4. CHAOS (2018) [link]	4	20	1
5. Trauma Detect. (2023) [link]	0	4,714	23	6. LiTS (2019) [link]	1	131	7
7-12. MSD CT Tasks (2021) [link]	9	945	1	13. KiTS (2020) [link]	1	489	1
14. AMOS22 (2022) [link]	15	200	2	15. WORD (2021) [link]	16	120	1
16. CT-ORG (2020) [link]	5	140	8	17. FLARE'23 (2022) [link]	13	4,100	30
18. AbdomenAtlas (2023) [link]	8	5,195	26	19. AbdomenAtlas 1.1	25	9,262‡	88

†Our reported number of CT volumes may differ from original publications, as some CT volumes are reserved for validation purposes.

‡The number of CT volumes in AbdomenAtlas 1.1 is lower than the sum of datasets 1–17 due to overlaps within these public datasets.

AbdomenAtlas 1.1 is a composite dataset that unifies CT volumes from public datasets 1–17 as summarized in Table 1 and Appendix B.1. We provide per-voxel annotations for 25 anatomical structures, including 16 abdominal organs, two thorax organs, five vascular structures, and two skeletal structures. We also provide pseudo annotations for seven types of tumors, namely liver, kidneys, pancreatic, hepatic vessel, lung, colon tumors, and kidney cysts. In total, more than 272.7B voxels are annotated in AbdomenAtlas 1.1, marking a significant leap compared with the 4.3B voxels annotated in the existing public datasets, amplifying the annotations by a factor of 63.4× (illustrated in Appendix Figure 5). AbdomenAtlas 1.1 presents a level of diversity because the CT volumes are sourced from 88 hospitals worldwide. The gap between these CT volumes includes changes in image quality due to different acquisition parameters, reconstruction kernels, and contrast enhancement, exampled in Appendix B.2. The CT volumes in AbdomenAtlas 1.1 include pre, portal, arterial, and delayed phases. *We commit to releasing the entire AbdomenAtlas 1.1 to the public.* This dataset, the largest public per-voxel annotated CT collection by far, accounts for only 0.01% of the CT volumes annually acquired in the United States (Papanicolas et al., 2018). Therefore, cross-institutional collaboration is crucial for accelerating data sharing, annotation, and AI development.

3.2 A Suite of Pre-trained Models: SuPreM

The magnitude of our AbdomenAtlas 1.1 is unprecedented in terms of data and annotations. One of the advantages is that it enables us to train AI models in both a supervised and self-supervised manner. At the time this paper is written, neither supervised nor self-supervised pre-training has been performed on this scale of dataset (9,262 volumetric data)[3]. We have developed models (termed SuPreM) pre-trained on data and annotations in AbdomenAtlas 1.1, which leverage established CNN backbones, such as U-Net and SegResNet, as well as Transformer backbones, such as Swin UNETR. With the growing trend of using pre-trained models, we have maintained a standardized, accessible project page to sharing public model weights as well as a suite of supervised pre-trained models (SuPreM) released by us. Releasing pre-trained models should be considered a marked contribution as they offer an alternative way of knowledge sharing while protecting patient privacy (Zhang & Metaxas, 2023; Sellergren et al., 2022; Ma et al., 2023a). In this study, all of the models in SuPreM follow pre-training and fine-tuning configurations as below.

[3] For supervised pre-training, the largest study to date was by Liu et al. (2023b), which was developed on 3,410 (2,100 for training and 1,310 for validation) annotated CT volumes. For self-supervised pre-training, the largest one was by Tang et al. (2022), which was trained on 5,050 unannotated CT volumes. Concurrently, Valanarasu et al. (2023) pre-trained a model on 50K volumes of CT and MRI using self-supervised learning.

Table 2: **Contribution #2: A suite of pre-trained models (termed SuPreM) comprising several widely recognized AI models.** We provide pre-trained AI models based on CNN, Transformer, and their hybrid versions, and more AI models will be added. Each model was supervised pre-trained on large datasets and per-voxel annotations from AbdomenAtlas 1.1. Compared with learning from scratch and publicly available models, fine-tuning the models in SuPreM consistently achieve the state-of-the-art organ and tumor segmentation performance on two datasets. All of the result, including the mean and standard deviation (mean±s.d.) across ten trials. In addition, we have further performed an independent two-sample t-test between learning from scratch and fine-tuning models in our SuPreM. The performance gain is statistically significant at the $P = 0.05$ level, with highlighting in a light red box. Detailed per-class performance can be found in Appendix §C.1.

			TotalSegmentator			proprietary dataset	
model (# of param)	pre-training	organ	muscle	cardiac	organ	gastro	cardiac
U-Net (2015) family (19.08M)	scratch	88.9±0.6	92.9±0.4	88.8±0.7	85.6±0.5	69.8±1.2	38.1±1.1
	Zhou et al. (2019)	85.9	90.1	86.3	80.1	65.5	36.9
	Chen et al. (2019b)	86.9	91.4	87.4	79.0	66.2	36.7
	Xie et al. (2022)	88.5	92.9	89.0	-	-	-
	Zhang et al. (2021)	89.3	93.8	89.1	85.7	72.7	38.3
	SuPreM	92.1±0.3	95.4±0.1	92.2±0.3	90.8±0.2	76.2±0.8	70.5±0.5
Swin UNETR (2021) (62.19M)	scratch	86.4±0.5	88.8±0.5	84.5±0.6	77.3±0.9	65.9± 1.7	35.5±1.4
	Tang et al. (2022)	89.3	93.8	88.3	87.9	72.5	38.9
	Liu et al. (2023b)	89.7	94.1	89.4	89.1	74.6	67.6
	SuPreM	91.3±0.3	94.6±0.2	90.3±0.3	90.4±0.7	75.9±1.2	69.8±0.9
SegResNet (2016) (470.13M)	scratch	88.6±0.5	91.3±0.4	89.8±0.4	80.6±0.8	67.0±1.4	36.0±1.3
	SuPreM	91.3±0.5	94.0±0.1	91.3±0.5	86.6±0.3	73.7±1.0	67.9±0.8

To perform a fair and rigorous comparison, we benchmarked with public pre-training methods by pre-training SuPreM using 2,100 CT volumes (same as Liu et al. (2023b) and fewer than Tang et al. (2022)) in Tables 2, 4 and Figures 1, 2b, 3. Then, we scaled up the number of CT volumes for pre-training to 9,262 CT volumes to perform direct inference in Table 3. Lastly, we scaled down the number of CT volumes to 21 to explore the edge of our SuPreM in Figure 2a. All these pre-trained models and configurations have been summarized in Appendix Table 8. The best-performing model was selected based on the highest average DSC score over 32 classes on a validation set of 1,310 CT volumes. Implementation details of both pre-training and fine-tuning can be found in Appendix C.2.

The transfer learning ability is assessed by segmentation performance on two datasets, i.e., TotalSegmentator and a proprietary dataset. Benchmarking results in Table 2 indicate that, in comparison with learning from scratch and with existing public models, those fine-tuned from our SuPreM consistently attain superior organ, muscle, cardiac, and gastro segmentation performance on both datasets. U-Net, as a simple and lightweight segmentation backbone, still performs competitively compared with alternative choices like Swin UNETR and SegResNet. This observation is aligned with the majority of the medical imaging community (Isensee et al., 2021; Eisenmann et al., 2023), suggesting that more exploration is needed for advancing segmentation backbones. Moreover, in the scenarios of either small data regimes shown in Figure 1 or large data regimes shown in Appendix Figure 8a–d, supervised models transfer better than their self-supervised counterparts. In summary, our SuPreM surpasses all existing 3D pre-trained models by a large margin in transfer learning performance, irrespective of their pre-training methodologies or data sources.

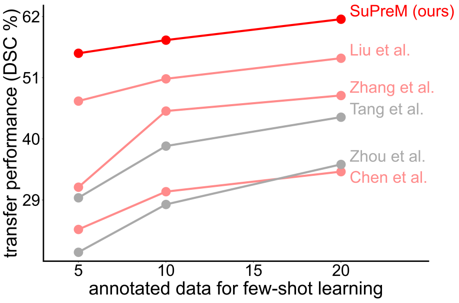
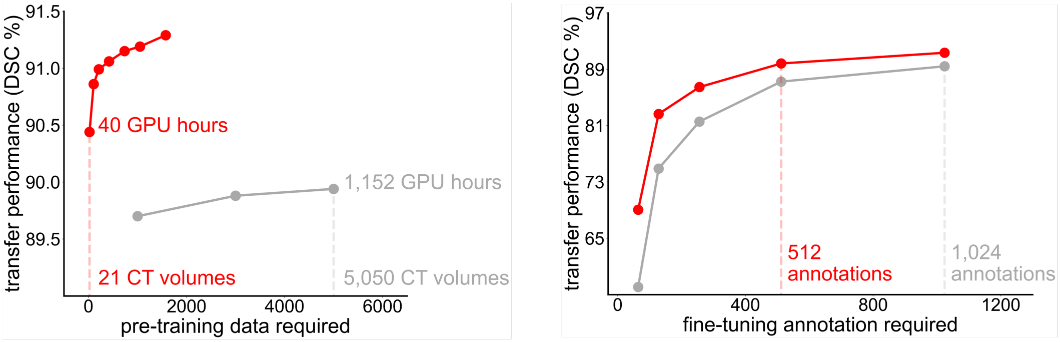


Figure 1: We present the transfer performance on a proprietary dataset with few-shot examples ($N = 5, 10, 20$). The transfer performance (Y-axis) stands for the average DSC score over 20-class organ segmentation and 3-class tumor segmentation. Overall, in few-shot learning, supervised pre-trained models (in red) transfer better than self-supervised pre-trained models (in gray). Notably, our SuPreM achieves the best transfer performance over other well-known publicly available models.



(a) data & computational efficiency in *pre-training*

(b) annotation & learning efficiency in *fine-tuning*

Figure 2: **Analysis of pre-training and fine-tuning efficiency.** For a fair comparison, both supervised (in red) and self-supervised (in gray) models use Swin UNETR as the backbone, and the compared self-supervised pre-training is the current state of the art (Tang et al., 2022). The target task was on the TotalSegmentator dataset. **(a)** scales the model transfer learning ability when pre-trained on varying numbers of images. The results indicate a consistent improvement in transfer learning ability when pre-training on more images. The model trained with 21 CT volumes, 672 masks, and 40 GPU hours shows a transfer learning ability similar to that trained with 5,050 CT volumes and 1,152 GPU hours. Specifically, supervised pre-training is more efficient, requiring 99.6% fewer data and 96.5% less computation. **(b)** assesses the annotation & learning efficiency by fine-tuning models on different number of annotated CT volumes from TotalSegmentator. Specifically, SuPreM, fine-tuned on 512 per-voxel annotated CT volumes, can achieve a segmentation performance on par with self-supervised models fine-tuned on 1,024 volumes, reducing 50% manual annotation cost for target tasks.

4 Experiment & Analysis

4.1 Data, Annotation, and Computational Efficiency

Summary. We demonstrate the remarkable efficiency: (1) SuPreM trained with 21 CT volumes, 672 masks, and 40 GPU hours shows transfer learning ability similar to that trained with 5,050 CT volumes and 1,152 GPU hours. (2) SuPreM requires 50% fewer manual annotations for organ/tumor segmentation than self-supervised pre-training.

Data efficiency for pre-training. As shown in Figure 2a, supervised pre-training requires less data (21 vs. 5,050 CT volumes) for the pretext task than self-supervised pre-training. This discrepancy arises from the inherent differences in their learning learning objectives and the information they leverage. Supervised pre-training benefits from explicit annotations, which provide direct guidance for the task, i.e., segmentation in this study. The model learns features from both data and annotations, which offer strong and precise supervision. On the other hand, self-supervised learning relies on pretext tasks derived from the raw data, which may offer a more ambiguous learning signal, therefore requiring more examples to capture meaningful features. Importantly, our finding suggests that supervised pre-training is more scalable with increased data. When data are increased from 21 to 1,575 volumes, the transfer learning performance on TotalSegmentator improves from 90.4% to 91.3%. In comparison, for self-supervised pre-training, an increase in data from 1,000 to 5,050 volumes only marginally improves performance from 89.7% to 89.9%. Therefore, supervised pre-training requires significantly less data than self-supervised and is more scalable and effective with increased data.

Annotation efficiency for fine-tuning. We have assessed the annotation efficiency by fine-tuning SuPreM and self-supervised models (Tang et al., 2022) on the TotalSegmentator dataset. Figure 2(b) suggests that fine-tuning SuPreM can reduce annotation costs for the segmentation task by 50%, averaged over the classes that were not used for pre-training (per-class performance can be found in Appendix Figure 9a–d). Specifically, SuPreM fine-tuned on 512 per-voxel annotated CT volumes can achieve segmentation performance similar to Tang et al. (2022) fine-tuned on 1,024 annotated CT volumes. The fine-tuning performance improvement gets bigger when the number of annotated

Table 3: **Direct inference on three external datasets.** We conduct external validation across numerous hospitals worldwide. Specifically, our SuPreM—trained on 9,262 CT volumes—is directly inferred (inf.) on three external datasets, i.e., TotalSegmentator (representing the Central European population from Switzerland), FLARE'23 (the East Asian population from China), and the proprietary dataset (the North American population from the United States) measured by DSC scores. For every dataset, we compare the *out-of-distribution* (OOD) performance obtained by SuPreM with *independently and identically distributed* (IID) performance obtained by AI models trained on the data and annotations of that specific dataset, which are considered as upper bound performance. We find that SuPreM can be generalized well across external datasets without additional fine-tuning, yielding comparable or even superior performance to the IID counterparts, evidenced by the t-test results. Appendix B.3 and Appendix E.1 visualizes examples of anatomical structures rendering and anatomical structures segmentation, respectively.

class	TotalSegmentator		FLARE'23		our proprietary dataset	
	inf. (SuPreM)	Wasserthal et al.	inf. (SuPreM)	Liu et al.	inf. (SuPreM)	Wang et al.
spleen	95.2±0.0	93.2	96.5±0.0	96.6	95.0±0.0	89.6
kidney right	92.5±0.2	91.2	93.7±0.1	90.7	92.2±0.0	88.0
kidney left	89.0±0.3	89.4	93.0±0.0	91.0	91.6±0.1	83.9
gall bladder	82.8±0.2	82.2	84.8±0.2	82.1	83.6±0.2	85.4
liver	94.7±0.2	94.0	96.8±0.1	97.8	95.0±0.3	91.4
stomach	85.2±0.3	82.8	90.7±0.6	92.9	92.2±0.1	90.1
aorta	75.6±0.2	72.1	87.0±0.7	84.5	73.9±0.3	87.0
IVC	74.2±0.2	73.7	85.5±0.4	87.8	77.7±0.4	80.8
pancreas	83.5±0.2	80.6	85.4±0.2	81.7	79.0±0.3	79.3
average	85.9±0.2	84.4	90.4±0.3	89.5	86.7±0.2	86.1

CT volumes is limited in the target task (e.g., 64, 128, 256). In addition, similar levels of annotation efficiency (reduced 50% cost) are observed when fine-tuning SuPreM on the three-class tumor segmentation task using the proprietary dataset, as presented in Appendix Figure 9e–g.

Computational efficiency for both pre-training and fine-tuning. This efficiency stems, in part, from the reduced data requirements inherent to supervised pre-training, as discussed above. As shown in Figure 2(a), supervised pre-training only needs 40 GPU hours to achieve a transfer learning performance comparable to that of self-supervised pre-training, which requires 1,152 GPU hours—a factor increase of 28.8×. When fine-tuning on target tasks, such as on a 10% subset of TotalSegmentator in Figure 10, the supervised pre-trained model converges much faster than the self-supervised one, reducing the GPU hours needed from 60 to 20. This implies that image features learned by supervised pre-training are intrinsically more expressive, enabling the model to seamlessly adapt across a myriad of 3D image segmentation tasks with minimal annotated data for fine-tuning. This computational efficiency makes supervised pre-training a compelling choice for 3D image segmentation without compromising model performance, especially when the large, annotated dataset is available.

4.2 Enhanced Features for Novel Datasets, Classes, and Tasks

Summary. The learned features manifest considerable generalizability and adaptability. The features can *direct inference* for organ segmentation on external datasets of CT volumes taken from different hospitals. The features can also be *fine-tuned* to segment novel organ/tumor classes and classify tumor sub-types with higher accuracy and less annotated data than those learned by self-supervision.

Direct inference on external datasets. AI models trained on a specific dataset often encounter challenges in generalizing to novel datasets when a marked difference—referred to as a *domain gap*—exists between them (Zhang & Metaxas, 2023). While domain adaptation and generalization are prevalent research strategies to mitigate this challenge (Guan & Liu, 2021; Zhou et al., 2022a), we choose to address this issue by training a model on an expansive and diverse dataset (elaborated in Appendix B.2). We assume the domain gap between CT volumes from different hospitals is not as pronounced as those in computer vision. This is because of the relatively standardized nature of computer tomography as an imaging modality, where pixel intensity conveys consistent anatomical significance (Zhou et al., 2022b). AbdomenAtlas 1.1 presents impressive diversity, covering CT volumes with variations in contrast enhancement, reconstruction kernels, CT scanner types, and acquisition parameters. This breadth and diversity are imperative for developing an AI model with the robustness required to accommodate the variations present in novel datasets. We conduct external

Table 4: **Fine-tuning SuPreM on 66 novel classes.** Following the standard transfer learning paradigm, we fine-tune our SuPreM on the segmentation task of novel classes. These tasks include segmenting 19 muscles, 15 cardiac structures, 5 organs, and 24 vertebrae from TotalSegmentator, as well as three fine-grained pancreatic tumor types from the proprietary dataset. It is important to note that these classes were not part of the pre-training of SuPreM. We observe that SuPreM, supervised pre-trained on only a few classes, can transfer better than those self-supervised pre-trained on raw, unlabeled data measured by DSC scores (per-class results in Appendix E.3). In other words, it is the task of segmentation itself that can enhance the model's capability of segmenting novel-class objects. This benefit is much more straightforward and understandable than such self-supervised tasks as contextual prediction, mask image modeling, and instance discrimination in the context of transfer learning. We hypothesize that it is because the model learns to understand the concept of *objectness* in a broader sense through full supervision, as suggested by Kirillov et al. (2023), but this certainly deserves further exploration. In addition, an independent two-sample t-test was performed between the self-supervised pre-trained model and the supervised pre-trained model. The performance gain (Δ) is statistically significant at the $P = 0.05$ level, with highlighting in a light red box.

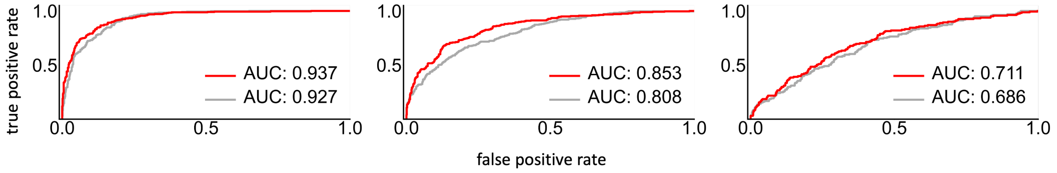
novel class	self-super.	super.	Δ	novel class	self-super.	super.	Δ
humerus left	92.6±0.3	93.1±0.2	0.5	vertebrae L5	89.6±0.7	89.0±0.7	-0.6
humerus right	87.3±1.0	94.9±0.1	7.6	vertebrae L4	90.4±0.7	93.0±0.2	2.5
··· (15 more classes)				··· (20 more classes)			
iliopsoas left	84.5±0.4	85.9±0.4	1.5	vertebrae C2	86.3±0.5	86.5±1.9	0.2
iliopsoas right	87.6±0.4	88.8±0.2	1.1	vertebrae C1	79.5±2.3	78.9±1.1	-0.6
average (muscle)	93.9±0.3	94.3±0.1	0.4	**average (vertebrae)**	84.3±1.3	85.4±0.9	1.1
trachea	93.4±0.1	93.3±0.1	-0.1				
heart myocardium	88.9±0.2	89.7±0.2	0.8				
··· (11 more classes)				PDAC	53.4±0.3	53.6±0.4	0.2
urinary bladder	90.1±0.9	91.2±0.5	1.1	Cyst	41.6±0.4	49.2±0.5	7.6
face	75.3±0.8	85.0±0.4	9.7	PanNet	35.4±0.8	45.7±0.8	10.2
average (cardiac)	88.9±0.5	90.7±0.3	1.8	**average (tumor)**	48.9±0.4	53.1±0.4	4.2

validation on several novel datasets sourced from Switzerland and East Asia to challenge the AI model on the data distribution that it has not encountered during the training. This result is referred to as *out-of-distribution* (OOD) performance. For comparison, we also collect the result achieved by dataset-specific AI models—those individually trained on the specific datasets—referred to as *independently and identically distributed* (IID) performance. As shown in Table 3, our SuPreM can be generalized well to novel data distribution without the need for further fine-tuning or adaptation, consistently offering OOD performance that matches or even exceeds that of its IID counterparts.

Fine-tuning on novel classes. The value of transfer learning lies in fine-tuning the pre-trained models on novel scenarios (Zhou et al., 2021b), such as novel classes, image modalities, and vision tasks that are completely unseen during the pre-training. In this study, we evaluate the proficiency of SuPreM when transferred to a wide variety of novel classes for 3D image segmentation tasks[4]. These novel classes include 19 muscles, 15 cardiac structures, 5 organs, and 24 vertebrae from the TotalSegmentator dataset, as well as three fine-grained pancreatic tumor types from the proprietary dataset. As shown in Table 4, our SuPreM, supervised pre-trained on 25 classes, can transfer better to novel classes than those self-supervised models pre-trained on raw, unlabeled data. We find that the pretext task of segmentation itself can enhance the model capability of segmenting novel classes. Correlation analysis in Appendix Figure 14 reveals a strong positive correlation (r = 0.81; p = 0.0031) in segmentation performance between the pretext and target tasks. The benefit of same-task transfer learning, i.e., segmentation as pretext and target tasks, is much more straightforward and understandable than other pretext tasks such as contextual prediction, mask image modeling, and instance discrimination. Through full supervision in segmentation tasks, the model learns to understand the concept of *objectness*[5], wherein the model gains a more profound understanding of what characterizes an object. The model does not just recognize predefined objects but begins to understand the foundational factors of objects in general. Such factors include texture, boundary, shape, size, and other low-level visual cues that are often deemed essential for image segmentation.

[4] The fine-tuning performance of 17 seen classes, detailed in Appendix E.2, is promising, but this is expected given that the model is exposed to more examples of these classes in both pre-training and fine-tuning phases.

[5] Objectness refers to the inherent attributes that distinguish something as an object within an image, differentiating it from the background or other entities.



(a) PDAC

(b) Cyst

(c) PanNET

Figure 3: **Fine-tuning SuPreM on fine-grained tumor classification.** We plot receiver operating characteristic (ROC) curves to evaluate the transfer learning performance of tumor classification. Detecting Cysts and PanNETs raises additional challenges for AI because these lesions exhibit a greater variety of texture patterns than PDACs. This diversity in texture patterns is reflected in the values of the Area Under the Curve (AUC) that we obtained. For all three sub-types of pancreatic tumors, SuPreM (in red) demonstrates superior performance over the self-supervised model (Tang et al., 2022) (in gray), showcasing its effectiveness in fine-grained tumor classification.

This resonates with our assertion in the introduction: just as classification-based features from ImageNet transfer optimally for classification tasks (Huh et al., 2016; He et al., 2019; Zoph et al., 2020; Ridnik et al., 2021), segmentation-based features are optimal for segmentation tasks.

Fine-tuning on novel tasks. We have investigated the cross-task transfer learning ability of SuPreM between organ segmentation and fine-grained tumor classification. The distance between the two tasks is much larger than transferring among segmentation tasks. It is challenging to benchmark fine-grained tumor classification, particularly due to the scarcity of annotations in public datasets (often limited to hundreds of tumors). To overcome this limitation, we employed our proprietary dataset (Xia et al., 2022), which comprises 3,577 annotated pancreatic tumors, including detailed sub-types: 1,704 PDACs, 945 Cysts, and 928 PanNets. This extensive dataset enabled us to thoroughly assess the transfer learning ability of SuPreM in tumor-related tasks. Figure 3 shows that supervised models (SuPreM) transfer better to target classification tasks than self-supervised models (Tang et al., 2022), leading to improved Area Under the Curve (AUC) for identifying each tumor type. Notably, the transfer learning results detailed in Appendix E.4 reveal a sensitivity of 86.1% and specificity of 95.4% for PDAC detection. This performance surpasses the average radiologist's performance in PDAC identification by 27.6% in sensitivity and 4.4% in specificity, as reported in Cao et al. (2023). Moreover, Appendix Figure 9 shows that SuPreM requires 50% fewer manual annotations for fine-grained tumor classification than self-supervised pre-training. This is particularly critical for tumor imaging tasks because annotating tumors requires much more effort and often relies on the availability of pathology reports.

5 Conclusion and Discussion

This study examines the transfer learning ability of supervised models that are pre-trained on 3D annotated datasets and fine-tuned on 3D image segmentation tasks. We start by constructing AbdomenAtlas 1.1, an extensive collection of 9,262 three-dimensional CT volumes with high-quality, per-voxel annotations. The magnitude of this dataset is unprecedented regarding data volume (**3.8M images**), granularity of annotations (**300K masks**), and inclusive diversity (**88 hospitals**). This dataset facilitates the development of a suite of pre-trained models, termed SuPreM, that can be effectively transferred to a broad spectrum of 3D image segmentation tasks. Notably, SuPreM transfers better than all existing 3D models by a large margin, irrespective of their pre-training methodologies or data sources; the benefit is more pronounced if the model is transferred to datasets that have limited annotations. The model trained with 21 CT volumes, 672 masks, and 40 GPU hours shows a transfer learning ability similar to that trained with 5,050 CT volumes and 1,152 GPU hours, highlighting the remarkable efficiency of supervised pre-training. We also demonstrate that the learned features can *direct inference* effectively on external datasets and *fine-tune* to segment novel classes and classify multiple types of tumors with higher accuracy and less annotated data than those learned by self-supervision. As open science, we will release both the annotated dataset (AbdomenAtlas 1.1) and pre-trained models (SuPreM) to the public.

ACKNOWLEDGMENTS

This work was supported by the Patrick J. McGovern Foundation Award. This work has utilized the GPUs provided partially by ASU Research Computing and NVIDIA. We appreciate the effort of the MONAI Team to provide open-source code for the community. We thank Chongyu Qu, Yixiong Chen, Junfei Xiao, Jie Liu, Yucheng Tang, Tiezheng Zhang, Yaoyao Liu, Chen Wei, Fengrui Tian, Yu-Cheng Chou, Angtian Wang, and Dora Zhiyu Yang for their constructive suggestions at several stages of the project.

REFERENCES

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021.

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785*, 2023.

Rob Baxter, Thomas Nind, James Sutherland, Gordon McAllister, Douglas Hardy, Ally Hume, Ruairidh MacLeod, Jacqueline Caldwell, Susan Krueger, Leandro Tramma, et al. The scottish medical imaging archive: 57.3 million radiology studies linked to their medical records. *Radiology: Artificial Intelligence*, pp. e220266, 2023.

Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019.

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. *arXiv preprint arXiv:2304.06131*, 2023.

Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature Medicine*, pp. 1–11, 2023.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*, 2020.

Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019a.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019b.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Xinze Chen, Guangliang Cheng, Yinghao Cai, Dayong Wen, and Heping Li. Semantic segmentation with modified deep residual networks. In *Pattern Recognition: 7th Chinese Conference, CCPR 2016, Chengdu, China, November 5-7, 2016, Proceedings, Part II 7*, pp. 42–54. Springer, 2016.

Errol Colak, Hui-Ming Lin, Robyn Ball, Melissa Davis, Adam Flanders, Sabeena Jalal, Kirti Magudia, Brett Marinelli, Savvas Nicolaou, Luciano Prevedello, Jeff Rudie, George Shih, Maryam Vazirabad, and John Mongan. Rsna 2023 abdominal trauma detection, 2023. URL `https://kaggle.com/competitions/rsna-2023-abdominal-trauma-detection`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Yang Deng, Ce Wang, Yuan Hui, Qian Li, Jun Li, Shiwei Luo, Mengke Sun, Quan Quan, Shuxin Yang, You Hao, et al. Ctspine1k: A large-scale dataset for spinal vertebrae segmentation in computed tomography. *arXiv preprint arXiv:2105.14711*, 2021.

Matthias Eisenmann, Annika Reinke, Vivienn Weru, Minu D Tizabi, Fabian Isensee, Tim J Adler, Sharib Ali, Vincent Andrearczyk, Marc Aubreville, Ujjwal Baid, et al. Why is the winner the best? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19955–19966, 2023.

Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601, 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 137–147. Springer, 2020.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Transactions on Medical Imaging*, 2021.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pp. 272–284. Springer, 2021.

Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918–4927, 2019.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Yuting He, Guanyu Yang, Jian Yang, Rongjun Ge, Youyong Kong, Xiaomei Zhu, Shaobo Zhang, Pengfei Shao, Huazhong Shu, Jean-Louis Dillenseger, et al. Meta grayscale adaptive network for 3d integrated renal structures segmentation. *Medical image analysis*, 71:102055, 2021.

Nicholas Heller, Sean McSweeney, Matthew Thomas Peterson, Sarah Peterson, Jack Rickman, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Joel Rosenberg, et al. An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging., 2020.

Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, et al. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv preprint arXiv:2304.06716*, 2023.

Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022.

Yankai Jiang, Mingze Sun, Heng Guo, Xiaoyu Bai, Ke Yan, Le Lu, and Minfeng Xu. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15859–15869, 2023.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 491–507. Springer, 2020.

Siddharth Krishna Kumar. On weight initialization in deep neural networks. *arXiv preprint arXiv:1704.08863*, 2017.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, pp. 12, 2015.

Jianning Li, Antonio Pepe, Christina Gsaxner, Gijs Luijten, Yuan Jin, Narmada Ambigapathy, Enrico Nasca, Naida Solak, Gian Marco Melito, Afaque R Memon, et al. Medshapenet–a large-scale dataset of 3d medical shapes for computer vision. *arXiv preprint arXiv:2308.16139*, 2023.

Jie Liu, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for partially labeled organ and pan-cancer segmentation. In *MICCAI 2023 FLARE Challenge*, 2023a.

Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21152–21164, 2023b.

Xiangde Luo, Wenjun Liao, Jianghong Xiao, Tao Song, Xiaofan Zhang, Kang Li, Guotai Wang, and Shaoting Zhang. Word: Revisiting organs segmentation in the whole abdominal region. *arXiv preprint arXiv:2111.02403*, 2021.

DongAo Ma, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Foundation ark: Accruing and reusing knowledge for superior and robust performance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 651–662. Springer, 2023a.

Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023a.

Jun Ma and Bo Wang. Towards foundation models of biological image segmentation. *Nature Methods*, 20(7):953–955, 2023b.

Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, et al. Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. *Medical Image Analysis*, 82:102616, 2022.

Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023b.

Zhiyu Ma, Chen Li, Tianming Du, Le Zhang, Dechao Tang, Deguo Ma, Shanchuan Huang, Yan Liu, Yihao Sun, Zhihao Chen, et al. Aatct-ids: A benchmark abdominal adipose tissue ct image dataset for image denoising, semantic segmentation, and radiomics evaluation. *arXiv preprint arXiv:2308.08172*, 2023c.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 181–196, 2018.

Mojtaba Masoudi, Hamid-Reza Pourreza, Mahdi Saadatmand-Tarzjan, Noushin Eftekhari, Fateme Shafiee Zargar, and Masoud Pezeshki Rad. A new dataset of computed-tomography angiography images for computer-aided detection of pulmonary embolism. *Scientific data*, 5(1): 1–9, 2018.

Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5): e210315, 2022.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, pp. 7294–7305. PMLR, 2020.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pp. 69–84. Springer, 2016.

Michal Pándy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. Transferability estimation using bhattacharyya class separability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9172–9182, 2022.

Irene Papanicolas, Liana R Woskie, and Ashish K Jha. Health care spending in the united states and other high-income countries. *Jama*, 319(10):1024–1039, 2018.

S Park, LC Chu, EK Fishman, AL Yuille, B Vogelstein, KW Kinzler, KM Horton, RH Hruban, ES Zinreich, D Fadaei Fouladi, et al. Annotated normal ct data of the abdomen for deep learning: Challenges and strategies for implementation. *Diagnostic and interventional imaging*, 101(1): 35–44, 2020.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan Yuille, and Zongwei Zhou. Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks. *Conference on Neural Information Processing Systems*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Sucheng Ren, Huiyu Wang, Zhengqi Gao, Shengfeng He, Alan Yuille, Yuyin Zhou, and Cihang Xie. A simple data mixing prior for improving self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14595–14604, 2022.

Sucheng Ren, Fangyun Wei, Zheng Zhang, and Han Hu. Tinymim: An empirical study of distilling mim pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3687–3697, 2023.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):1–9, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 556–564. Springer, 2015.

Agustina Saenz, Emma Chen, Henrik Marklund, and Pranav Rajpurkar. The maida initiative: establishing a framework for global medical-imaging data sharing. *The Lancet Digital Health*, 6 (1):e6–e8, 2024.

Andrew B Sellergren, Christina Chen, Zaid Nabulsi, Yuanzhen Li, Aaron Maschinot, Aaron Sarna, Jenny Huang, Charles Lau, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, et al. Simplified transfer learning for chest radiography models using less data. *Radiology*, 305(2):454–465, 2022.

Azizi Shekoofeh, Mustafa Basil, Ryan Fiona, Beaver Zachary, Freyberg Jan, Deaton Jonathan, Loh Aaron, Karthikesalingam Alan, Kornblith Simon, Chen Ting, et al. Big self-supervised models advance medical image classification. *arXiv preprint arXiv:2101.05224*, 2021.

Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

Nahian Siddique, Sidike Paheding, Colin P Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *Ieee Access*, 9:82031–82057, 2021.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

Yang Tan, Yang Li, and Shao-Lun Huang. Otce: A transferability metric for cross-domain cross-task representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15779–15788, 2021.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20730–20740, 2022.

Xing Tao, Yuexiang Li, Wenhui Zhou, Kai Ma, and Yefeng Zheng. Revisiting rubik's cube: Self-supervised learning with volume-wise transformation for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 238–248. Springer, 2020.

National Lung Screening Trial Research Team. The national lung screening trial: overview and study design. *Radiology*, 258(1):243–253, 2011.

Atharva Tendle and Mohammad Rashedul Hasan. A study of the generalizability of self-supervised representations. *Machine Learning with Applications*, 6:100124, 2021.

Jeya Maria Jose Valanarasu, Yucheng Tang, Dong Yang, Ziyue Xu, Can Zhao, Wenqi Li, Vishal M Patel, Bennett Landman, Daguang Xu, Yufan He, et al. Disruptive autoencoders: Leveraging low-level features for 3d medical image pre-training. *arXiv preprint arXiv:2307.16896*, 2023.

Vanya V Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 547–556. IEEE, 2018.

Yan Wang, Yuyin Zhou, Wei Shen, Seyoun Park, Elliot K Fishman, and Alan L Yuille. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Medical image analysis*, 55:88–102, 2019.

Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. P2p: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. *Advances in neural information processing systems*, 35:14388–14402, 2022.

Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, Joshy Cyriac, Shan Yang, and Martin Segeroth. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. *arXiv preprint arXiv:2208.05868*, 2022.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14668–14678, 2022.

Yingda Xia, Qihang Yu, Linda Chu, Satomi Kawamoto, Seyoun Park, Fengze Liu, Jieneng Chen, Zhuotun Zhu, Bowen Li, Zongwei Zhou, et al. The felix project: Deep networks to detect pancreatic neoplasms. *medRxiv*, 2022.

Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. *IEEE Winter Conference on Applications of Computer Vision*, 2022.

Yutong Xie, Jianpeng Zhang, Zehui Liao, Yong Xia, and Chunhua Shen. Pgl: Prior-guided local self-supervised learning for 3d medical image segmentation. *arXiv preprint arXiv:2011.12640*, 2020.

Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *European Conference on Computer Vision*, pp. 558–575. Springer, 2022.

Xingyi Yang, Xuehai He, Yuxiao Liang, Yue Yang, Shanghang Zhang, and Pengtao Xie. Transfer learning or self-supervised learning? a tale of two pretraining paradigms. *arXiv preprint arXiv:2007.04234*, 2020.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pp. 3320–3328, 2014.

Chenyu You, Ruihan Zhao, Fenglin Liu, Siyuan Dong, Sandeep Chinchali, Ufuk Topcu, Lawrence Staib, and James Duncan. Class-aware adversarial transformers for medical image segmentation. *Advances in Neural Information Processing Systems*, 35:29582–29596, 2022.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, pp. 12133–12143. PMLR, 2021.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1195–1204, 2021.

Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *arXiv preprint arXiv:2306.05705*, 2023.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329, 2022.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. One model to rule them all: Towards universal segmentation for medical images with text prompts. *arXiv preprint arXiv:2312.17183*, 2023.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021a.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022a.

Zongwei Zhou. *Towards Annotation-Efficient Deep Learning for Computer-Aided Diagnosis*. PhD thesis, Arizona State University, 2021.

Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7340–7351, 2017.

Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International conference on medical image computing and computer-assisted intervention*, pp. 384–393. Springer, 2019.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021b.

Zongwei Zhou, Michael B Gotway, and Jianming Liang. Interpreting medical images. In *Intelligent Systems in Medicine and Health*, pp. 343–371. Springer, 2022b.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33: 3833–3845, 2020.

Appendix

Table of Contents

A Overview **18**

B Extensive Dataset: AbdomenAtlas 1.1 **19**

B.1 Dataset Sources 19
B.2 Domain Transfer Across Datasets 21
B.3 Uniform Annotation Standards 22
B.4 On the Importance of Annotation Quality 23

C A Suite of Pre-trained Models: SuPreM **24**

C.1 Different Backbones in the Suite of Models 24
C.2 Supervised and Self-supervised Benchmarking 25

D Data, Annotation, and Computational Efficiency **29**

D.1 Annotation Efficiency in Fine-tuning 29
D.2 Convergence and Learning Efficiency in Fine-tuning 30

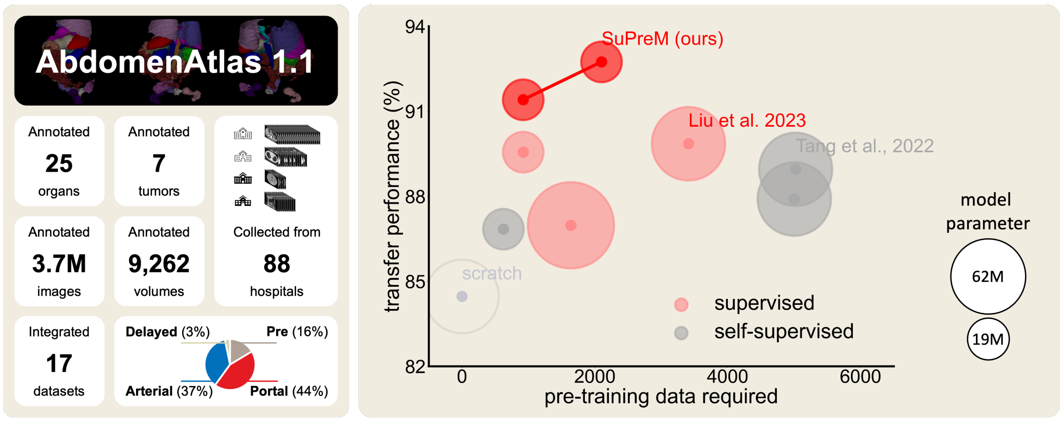
E Enhanced Features for Novel Datasets, Classes, and Tasks **31**

E.1 Direct Inference on Three External Datasets 31
E.2 Fine-tuning SuPreM on 17 Seen Classes 33
E.3 Fine-tuning SuPreM on 63 Novel Classes 34
E.4 Fine-tuning SuPreM on Fine-grained Tumor Classification 35

F More Discussion **36**

F.1 How Transferable Are Features in SuPreM? 36
F.2 Transfer Learning Ability Estimation 37
F.3 Supervised pre-training: image-language or image-mask pairs as supervision? . . 38
F.4 Broader Impact to 3D Vision Tasks? 39

A Overview



(a) an extensive dataset

(b) a suite of pre-trained models

Figure 4: **Our main contributions are as follows: (a)** An extensive dataset of 9,262 CT volumes with per-voxel annotations of 25 anatomical structures, enabling us to perform supervised pre-training of AI models. **(b)** A suite of pre-trained models comprising several widely recognized AI models, each pre-trained on large-scale datasets and per-voxel annotations. In summary, supervised pre-training (in red) strikes as a preferred choice in terms of performance and efficiency compared with self-supervised pre-training (in gray). We anticipate that the release of our annotated dataset (AbdomenAtlas 1.1) and the suite of pre-trained models (SuPreM) will bolster collaborative endeavors in establishing *Foundation Datasets* and *Foundation Models* for the broader applications of 3D volumetric medical image analysis.

We assess the transfer learning ability of our SuPreM, supervised pre-trained by organ segmentation, under five distinct settings.

1. Transfer to external datasets (domains) to segment the same anatomical structures—classes that were used for pre-training.
2. Transfer to segmentation tasks of organs, muscles, vertebrae, and cardiac structures—classes that were not used for pre-training.
3. Transfer to segmentation tasks of pancreatic tumor segmentation—more challenging classes that were not used for pre-training.
4. Transfer to few-shot segmentation tasks using only a limited number of annotated CT volumes—classes that were not used for pre-training.
5. Transfer to classification tasks that identify fine-grained tumors, including PDAC, Cyst, and PanNet in the proprietary dataset.

This evaluation protocol has been widely adopted to assess transfer learning ability in medical imaging (Zhou et al., 2021b; Tang et al., 2022; Jiang et al., 2023) and computer vision (He et al., 2022; Zhai et al., 2022). In the future, we plan to assess the transfer learning ability across imaging modalities and broader 3D vision tasks.

B Extensive Dataset: AbdomenAtlas 1.1

B.1 Dataset Sources

Table 5: **An extensive dataset of 9,262 CT volumes with per-voxel annotations of 25 anatomical structures.** This dataset is unprecedented in terms of data and annotation scales, providing 300K organ/tumor masks and 3.7M annotated images that are taken from 88 hospitals worldwide. In 2009, before the advent of ImageNet (Deng et al., 2009), it was challenging to empower an AI model with generalized image representation using a small or even medium size of labeled data, the same situation, we believe, that presents in 3D medical image analysis today. As seen in the table, the annotations of public datasets are limited, partial, and incomplete, and the CT volumes in these datasets are often biased toward specific populations, medical centers, and countries. Our constructed dataset mitigates these gaps, representing a significant leap forward in the field.

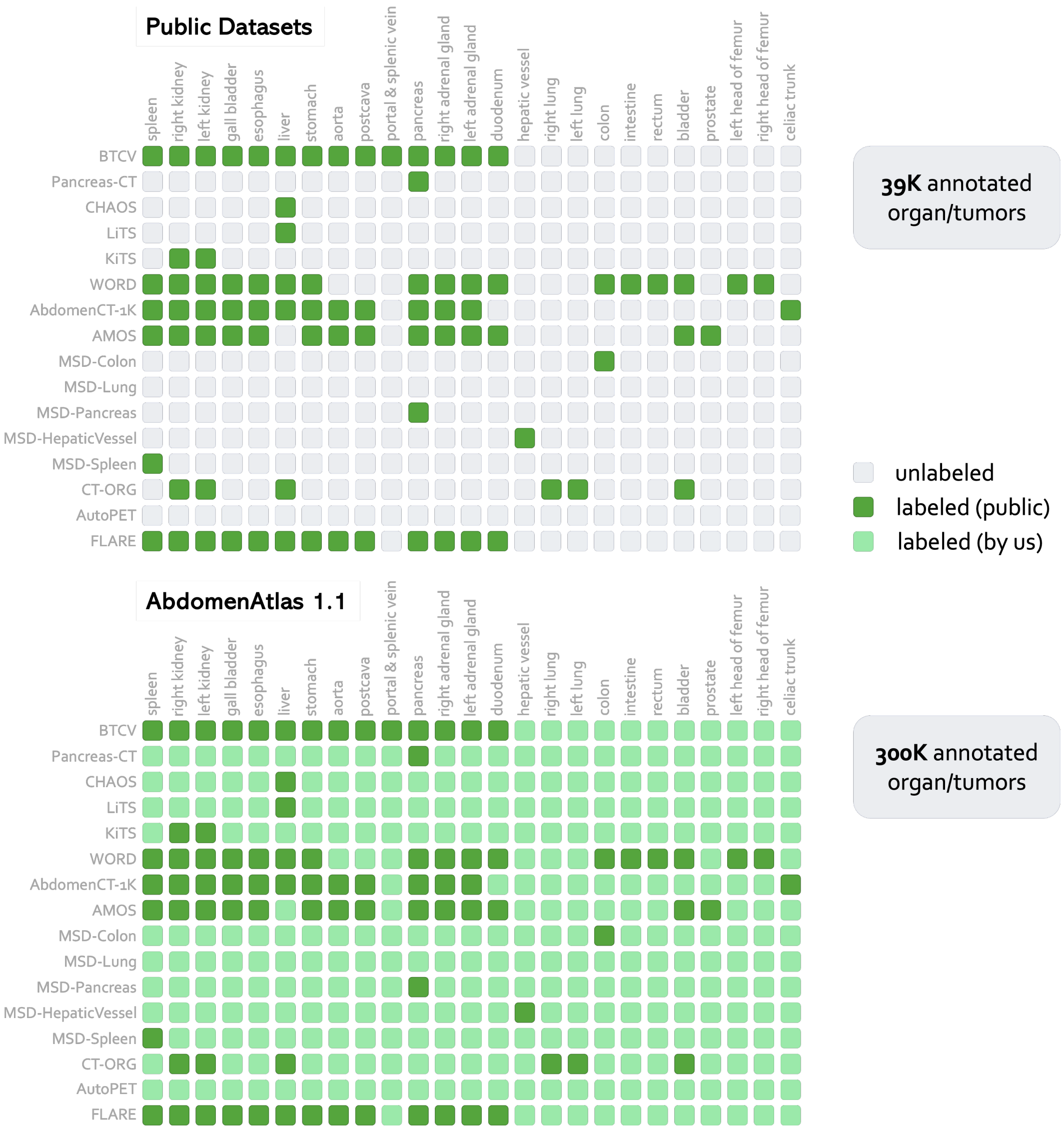
dataset (year) [source]	# of class	# of volume	# of center	source countries	license
1. Pancreas-CT (2015) [link]	1	82	1	US	CC BY 3.0
2. LiTS (2019) [link]	2	201	7	DE, NL, CA, FR, IL	CC BY-SA 4.0
3. AbdomenCT-1K (2021) [link]	4	1,000	12	DE, NL, CA, FR, IL, US, CN	CC BY-NC-SA
4. KiTS (2020) [link]	3	300	1	US	CC BY-NC-SA 4.0
5. AutoPET (2022) [link]	1	1,014	2	DE	TCIA Restricted
6. CHAOS (2018) [link]	4	40	1	TR	CC BY-SA 4.0
7-11. MSD CT Tasks (2021) [link]	9	947	1	US	CC BY-SA 4.0
12. BTCV (2015) [link]	12	50	1	US	CC BY 4.0
13. AMOS22 (2022) [link]	15	500	2	CN	CC BY-NC-SA
14. WORD (2021) [link]	16	150	1	CN	GNU GPL 3.0
15. CT-ORG (2020) [link]	6	140	8	DE, NL, CA, FR, IL, US	CC BY 3.0
16. FLARE'23 (2022) [link]	13	4,000	30	-	CC BY-NC-ND 4.0
17. AATTCT-IDS (2023c) [link]	-	300	1	CN	-
18. KiPA22 (2021) [link]	4	100	1	CN	CC BY-NC-ND 3.0
19. Abdominal Trauma Det. (2023) [link]	-	3,147	23	-	-
20. FUMPE (2018) [link]	1	35	1	IR	CC BY 4.0
21. TotalSegmentator (2022) [link]	104	1,204	1	CH	CC BY 4.0
22. CTSpine1K (2021) [link]	26	1,005	-	-	CC BY 4.0
23. AbdomenAtlas 1.1	25	9,262	88	US, DE, NL, FR, IL, CN, CA, TR, CH	pending

US: United States DE: Germany NL: Netherlands CA: Canada FR: France IL: Israel IR: Iran
CN: China TR: Turkey CH: Switzerland

Our objective in developing AbdomenAtlas 1.1 is to drive algorithmic advancements and set new benchmarks in the field of 3D medical imaging. In many ways, our dataset echoes the early days of ImageNet (Deng et al., 2009), as both datasets emerged at times when large-scale data, diverse classes, and detailed labels were sparse in their respective fields. The limitations of publicly available datasets have been summarized with statistics in Appendix Table 5 and Appendix Figure 5.

Segmentation is often conceptualized as per-voxel classification. In the medical domain, segmentation holds the same fundamental importance as classification does in general computer vision (Ma & Wang, 2023b). We bet that ImageNet-like datasets in the medical domain should be formed as per-voxel segmentation labels. Our dataset aligns with this vision by providing per-voxel labels, offering a level of detail far surpassing ImageNet's per-image labels. Concretely, the per-voxel labels in our dataset (272.7B annotated voxels) are much more extensive than the per-image labels in ImageNet (14M annotated images).

Appendix Table 5 has detailed the source and permissions for data release. Our approach involves disseminating only the annotations of the CT volumes, which users can combine with the original CT volumes obtained from their original sources. All data created and licensed out by us will be in separate files, ensuring no modifications to the original CT volumes. Legal consultations confirm our permission to distribute these annotations under the licenses of each dataset. We will also release the entire AbdomenAtlas 1.1 dataset to the public, providing 300K organ/tumor masks and 3.7M annotated images that are taken from 88 hospitals worldwide. This dataset will continue to expand with the collective effort from the community.



Public Datasets

	spleen	right kidney	left kidney	gall bladder	esophagus	liver	stomach	aorta	postcava	portal & splenic vein	pancreas	right adrenal gland	left adrenal gland	duodenum	hepatic vessel	right lung	left lung	colon	intestine	rectum	bladder	prostate	left head of femur	right head of femur	celiac trunk
BTCV	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
Pancreas-CT	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
CHAOS	⬜	⬜	⬜	⬜	⬜	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
LiTS	⬜	⬜	⬜	⬜	⬜	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
KiTS	⬜	🟩	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
WORD	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬜	⬜	⬜	🟩	🟩	🟩	🟩	⬜	⬜	⬜	🟩	🟩	🟩	🟩	⬜	🟩	🟩	⬜
AbdomenCT-1K	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬜	🟩	🟩	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩
AMOS	🟩	🟩	🟩	🟩	🟩	⬜	🟩	🟩	🟩	⬜	🟩	🟩	🟩	🟩	⬜	⬜	⬜	⬜	⬜	⬜	🟩	🟩	⬜	⬜	⬜
MSD-Colon	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜
MSD-Lung	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
MSD-Pancreas	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
MSD-HepaticVessel	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
MSD-Spleen	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
CT-ORG	⬜	🟩	🟩	⬜	⬜	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	🟩	⬜	⬜	⬜	🟩	⬜	⬜	⬜	⬜
AutoPET	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
FLARE	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬜	🟩	🟩	🟩	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜

39K annotated organ/tumors

⬜ unlabeled
🟩 labeled (public)
▩ labeled (by us)

AbdomenAtlas 1.1

	spleen	right kidney	left kidney	gall bladder	esophagus	liver	stomach	aorta	postcava	portal & splenic vein	pancreas	right adrenal gland	left adrenal gland	duodenum	hepatic vessel	right lung	left lung	colon	intestine	rectum	bladder	prostate	left head of femur	right head of femur	celiac trunk
BTCV	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩
Pancreas-CT	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	🟩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩
CHAOS	▩	▩	▩	▩	▩	🟩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩
LiTS	▩	▩	▩	▩	▩	🟩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩
KiTS	▩	🟩	🟩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩
WORD	🟩	🟩	🟩	🟩	🟩	🟩	🟩	▩	▩	▩	🟩	🟩	🟩	🟩	▩	▩	▩	🟩	🟩	🟩	🟩	▩	🟩	🟩	▩
AbdomenCT-1K	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	▩	🟩	🟩	🟩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	🟩
AMOS	🟩	🟩	🟩	🟩	🟩	▩	🟩	🟩	🟩	▩	🟩	🟩	🟩	🟩	▩	▩	▩	▩	▩	▩	🟩	🟩	▩	▩	▩
MSD-Colon	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	🟩	▩	▩	▩	▩	▩	▩	▩
MSD-Lung	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩
MSD-Pancreas	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	🟩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩
MSD-HepaticVessel	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	🟩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩
MSD-Spleen	🟩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩
CT-ORG	▩	🟩	🟩	▩	▩	🟩	▩	▩	▩	▩	▩	▩	▩	▩	▩	🟩	🟩	▩	▩	▩	🟩	▩	▩	▩	▩
AutoPET	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩
FLARE	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	▩	🟩	🟩	🟩	🟩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩

300K annotated organ/tumors

Figure 5: **Evolution: from a combination of public data to AbdomenAtlas 1.1.** AbdomenAtlas 1.1 is NOT a simple combination of existing datasets. The 9K CT volumes in the combination of public datasets only contain a total of **39K** annotated organ masks, while our AbdomenAtlas 1.1 provides over **300K** annotated organ/tumor masks for these CT volumes, substantially increasing the number of masks by **7.6** times. Creating 300K high-quality organ/tumor masks for 9K CT volumes requires extensive medical knowledge and annotation cost (much more difficult than annotating natural images). Based on our experience and those reported in Park et al. (2020), trained radiologists annotate abdominal organs at a rate of 30–60 minutes per organ per three-dimensional CT volume. This translates to **247K** human hours for completing AbdomenAtlas 1.1. We employed a highly efficient annotation method, combining AI with the expertise of three radiologists using active learning (details in Appendix B.3), to overcome this challenge and produce the largest annotated dataset to date.

B.2 Domain Transfer Across Datasets

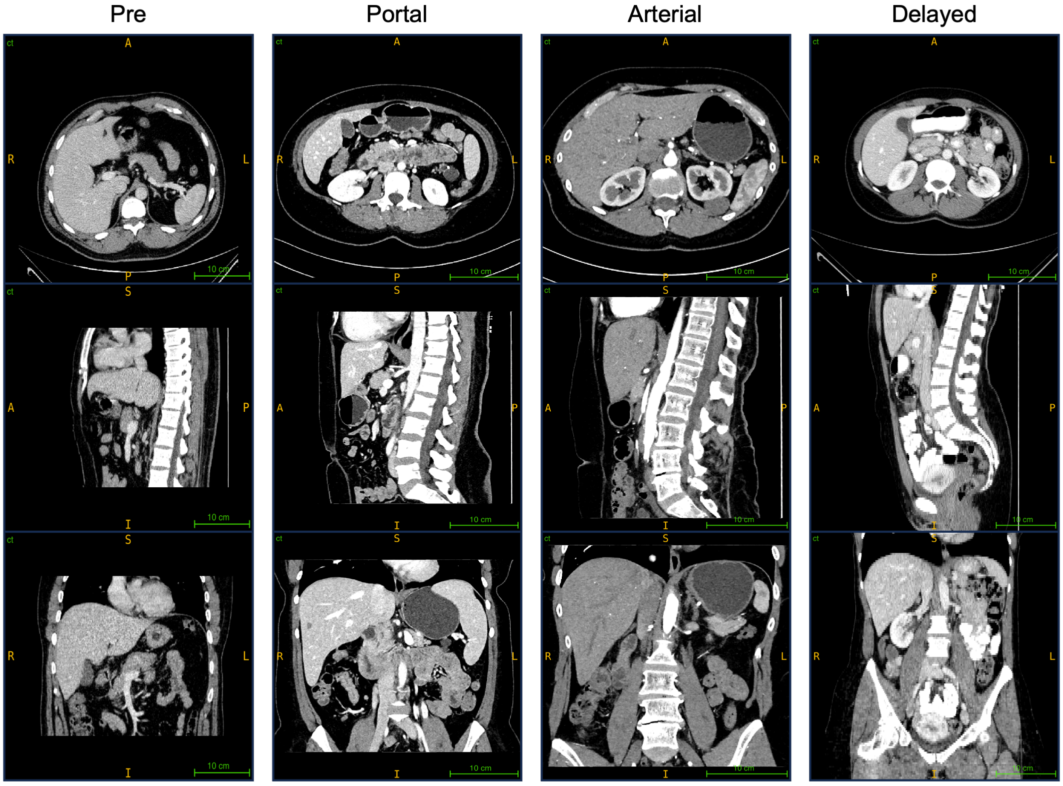


Figure 6: **Domain gaps.** Examples of CT volumes from different domains (e.g., hospitals and countries) illustrate the variability in images. AbdomenAtlas 1.1 are created by a large variety of CT scanners, imaging protocols, and acquired from numerous hospitals worldwide (Table 1). We note that substantial differences in CT volumes occur in image quality and technical display, originating from different acquisition parameters, reconstruction kernels, and contrast enhancements.

Table 3 shows that SuPreM is pretty robust because our AbdomenAtlas 1.1 covers a variety of domains (i.e., 88 hospitals with different scanners and protocols), as shown in Appendix Figure 6; models pre-trained on this dataset are expected to be generalizable for novel domains. Therefore, domain transfer becomes less important if the model is pre-trained on large and diverse datasets, elaborating on the two points below.

1. The domain transfer problem could be solved by methodology innovation, and also by training AI models on enormous datasets. This point has been more clear recently demonstrated by large language models (GPT) and vision foundation models (SAM), which show incredible performance in "novel domain". However, this achievement may not be directly attributed to method-driven solutions for domain transfer, but simply because the AI might have been trained on similar sentences or images. This was also pointed out by Yann Lecun—*beware of testing on the training set*—in response to the incredible results achieved by GPT.
2. In some sense, our paper explores dataset-driven solutions for domain transfer. The robust performance of our models when direct inference on multiple domains could also be attributed to our large-scale, fully-annotated medical dataset—as one of our major contributions. The release of AbdomenAtlas 1.1 can foster AI models that are more robust than the majority of existing models that are only trained on a few hundred CT volumes from limited domains. In addition, existing domain transfer methods could also be supplemented with direct inference and fine-tuning to further improve AI performance.

B.3 Uniform Annotation Standards

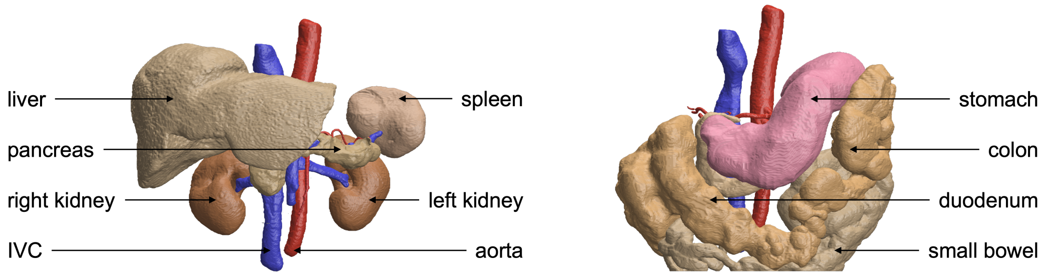


Figure 7: ***Automated organ annotations.*** Our annotation pipeline involved an interactive segmentation approach, an integration of AI algorithms and human expertise, which premises to improve the efficiency while upholding high-quality annotations. *One senior radiologist* revised the annotations predicted by our AI models, and in turn, the AI models improved their predictions by learning from these revised annotations. This interactive procedure continued to enhance the quality of annotations until no major revision is needed. Subsequently, *five junior radiologists* examine the final visualizations for accuracy (examples of the rendered images are shown above). The junior radiologists were responsible for reviewing the correctness of the annotations and marking the patient ID for any major discrepancies. Such cases are then reviewed by the senior radiologist. Our uniform annotation standards, largely overlapping with those in Ma et al. (2023b), require trained radiologists to spend approximately 30–60 minutes annotating each organ in a three-dimensional CT volume.

Automated (pseudo) tumor annotations. We have established uniform annotation standards for tumors, with both senior and junior radiologists actively refining and adhering to these guidelines.

- Liver tumors: Liver tumors include primary tumor lesions and metastases in the liver. Annotations should encompass the entire tumor, including any invasive parts, necrosis, hemorrhage, fibrous scars, and calcifications. Healthy areas or unrelated lesions are not included.
- Kidney tumors: Kidney tumors include both benign and malignant tumor lesions growing in the kidneys. The entire tumor and its invasive parts to surrounding areas, plus internal changes like necrosis and calcification, should be annotated. Exclude healthy structures.
- Pancreatic tumors: Pancreatic tumors include all benign and malignant tumor lesions growing in the pancreas. Annotations cover the whole tumor and its invasive growth into adjacent areas, including changes like cysts, necrosis, and calcification. Exclude healthy structures.
- Colon tumors: Colon tumors include all benign and malignant tumor lesions developing from the colon wall. The entire tumor and its invasion into nearby structures, along with internal changes like necrosis, should be annotated, excluding healthy areas.
- Hepatic vessel tumors: Hepatic vessel tumors include all primary tumor lesions developing from the intrahepatic vessel wall and tumor thrombus in intrahepatic vessels. Annotations should include the tumor within the vessels, excluding external parts and unrelated lesions.

Overall, AbdomenAtlas 1.1 offers 51.8K pseudo tumor masks visually inspected by radiologists, though without biopsy confirmation. While these masks lack pathological validation, we anticipate they will serve as a valuable foundation for expanding precise tumor annotations in future research.

B.4 On the Importance of Annotation Quality

Table 6: **On the importance of annotation quality.** We intentionally introduced deformities to the annotations within our dataset, termed AbdomenAtlas 1.1 (noisy), and trained a noisy AI model using this dataset. For comparison, a clean AI model is trained using the original AbdomenAtlas 1.1 (clean). We then evaluate the noisy model and the clean model using the pretext test set of AbdomenAtlas 1.1 and the target test set of TotalSegmentator. Results show that the suboptimal AI model shows diminished transfer learning ability on the target dataset, leading to negative transfer (Zhang et al., 2022). In summary, annotation quality profoundly impacts AI performance and transfer learning ability. These observations are corroborated by Figure 14, where superior performance in the pretext task indicates enhanced transfer learning ability, and vice versa.

pretext dataset (AbdomenAtlas 1.1)	AbdomenAtlas 1.1 (noisy)	AbdomenAtlas 1.1 (clean)
spleen	92.9	93.2
right kidney	88.9	89.2
left kidney	88.4	87.6
gall bladder	71.0	71.4
esophagus	72.1	71.3
liver	94.5	94.7
stomach	84.8	85.0
aorta	89.1	88.8
postcava	79.5	79.3
portal vein and splenic vein	67.3	61.1
pancreas	76.5	75.6
right adrenal gland	13.3	44.2
left adrenal gland	56.8	58.9
duodenum	59.5	60.8
hepatic vessel	29.2	29.4
right lung	85.8	85.9
left lung	65.0	72.2
colon	70.4	70.7
intestine	81.2	81.0
rectum	64.1	64.8
bladder	67.0	67.6
prostate	71.0	70.0
left head of femur	31.8	64.1
right head of femur	88.9	88.9
celiac truck	77.0	72.9
kidney tumor	20.3	25.6
liver tumor	35.2	45.0
pancreas tumor	15.0	17.2
hepatic vessel tumor	30.6	39.6
lung tumor	0.0	28.8
colon tumor	7.0	8.5
kidney cyst	9.0	9.4
average	58.8	62.6
target dataset (TotalSegmentator)		
spleen	72.2	96.3
kidney right	51.8	93.5
kidney left	47.6	94.9
gallbladder	71.3	84.5
liver	78.1	97.7
stomach	52.6	92.3
aorta	78.9	94.3
inferior vena cava	35.1	90.5
portal vein and splenic vein	67.2	81.6
pancreas	78.5	87.7
adrenal gland right	-	85.5
adrenal gland left	-	79.7
lung upper lobe left	81.2	96.0
lung lower lobe left	75.1	93.4
lung upper lobe right	73.4	89.7
lung middle lobe right	69.1	91.9
lung lower lobe right	73.8	95.2
average (organ)	67.1	90.9

C A Suite of Pre-trained Models: SuPreM

C.1 Different Backbones in the Suite of Models

Table 7: **Detailed results of different backbones on TotalSegmentator.** The performance is measured by the Dice Similarity Coefficient (DSC) and normalized surface dice (NSD) with a tolerance of 1mm. Current backbones included U-Net (CNN-type) (Ronneberger et al., 2015), SegResNet (CNN-type) (Chen et al., 2016), and Swin UNETR (Transformer-type) (Tang et al., 2022), and more backbones will be added along time.

TotalSegmentator	scratch		U-Net		SegResNet		Swin UNETR	
(Wasserthal et al., 2022)	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)
spleen	96.2±0.1	73.8±0.6	95.5±0.3	73.3±0.7	96.5±0.1	73.9±0.6	96.5±0.1	74.5±0.4
kidney right	95.0±0.2	76.3±0.6	95.3±0.4	70.9±0.5	94.8±0.1	75.4±0.8	94.1±0.6	71.6±0.8
kidney left	89.4±1.9	73.8±1.3	95.0±0.3	68.6±1.0	94.4±0.6	71.9±0.8	95.3±0.3	72.8±0.7
gallbladder	78.6±1.3	47.8±1.4	86.6±0.4	49.5±0.9	82.9±1.2	51.7±1.4	85.7±0.6	43.3±1.1
liver	97.6±0.0	72.3±0.3	97.7±0.1	73.0±0.4	97.8±0.0	72.9±0.4	97.7±0.2	72.0±0.4
stomach	88.4±0.2	59.5±0.8	93.3±0.5	60.7±0.7	92.7±0.5	60.3±1.1	93.1±0.5	59.0±0.4
aorta	92.7±0.7	73.1±0.4	95.3±0.2	73.4±0.3	95.2±0.4	72.6±0.3	96.1±0.1	71.9±0.4
inferior vena cava	89.3±0.3	62.8±0.6	89.5±0.6	62.9±0.3	90.9±0.4	61.6±0.6	89.4±0.2	60.5±0.7
portal & splenic vein	75.4±1.3	58.4±1.4	83.4±0.6	58.6±0.8	82.2±0.9	56.6±0.8	81.5±0.5	54.7±0.7
pancreas	86.2±0.5	58.3±1.0	88.5±0.6	59.5±0.8	88.1±0.7	59.9±0.8	88.9±0.4	56.7±0.3
adrenal gland right	82.1±0.5	69.4±0.6	87.8±0.3	69.6±0.8	85.2±0.3	66.4±0.9	87.1±0.1	64.9±0.3
adrenal gland left	75.1±1.2	59.3±1.4	83.1±0.8	60.4±0.9	80.7±0.7	57.1±1.3	79.1±0.5	55.1±0.5
lung upper lobe left	96.6±0.2	74.5±0.4	96.2±0.1	73.4±0.5	96.3±0.2	72.8±0.3	96.3±0.0	72.8±0.2
lung lower lobe left	92.3±0.7	71.5±0.2	95.3±0.3	69.9±0.7	93.6±0.4	69.9±0.7	93.5±0.5	69.3±0.4
lung upper lobe right	91.0±0.5	68.9±0.4	96.1±0.1	64.4±1.0	93.0±0.7	65.4±0.7	91.5±0.5	66.5±0.5
lung middle lobe right	90.2±0.3	60.0±0.4	92.0±0.0	60.0±0.4	92.7±0.3	59.7±0.6	91.6±0.2	56.8±0.4
lung lower lobe right	94.8±0.1	72.7±0.5	95.0±0.0	72.5±0.5	95.3±0.5	72.0±0.3	94.7±0.1	71.4±0.4
average	88.9±0.6	66.6±0.7	92.1±0.3	65.9±0.7	91.3±0.5	65.9±0.7	91.3±0.3	64.3±0.5
humerus left	88.6±1.5	69.1±1.0	92.3±0.2	71.4±0.7	90.0±0.3	48.3±1.2	93.3±0.4	72.8±1.0
humerus right	86.1±1.4	64.0±1.3	96.4±0.2	68.8±0.8	88.7±0.1	32.5±0.7	94.9±0.1	68.6±2.1
scapula left	92.4±0.2	79.4±0.9	95.2±0.2	79.1±1.0	93.4±0.1	73.2±0.7	92.7±0.5	77.1±0.5
scapula right	95.4±0.2	81.2±0.3	95.5±0.2	81.0±0.2	93.5±0.1	74.1±0.2	94.6±0.2	79.3±0.2
clavicula left	95.9±0.0	81.3±0.2	96.4±0.1	83.3±0.3	93.8±0.1	71.2±0.3	95.9±0.0	81.7±0.2
clavicula right	95.1±0.0	77.0±0.2	95.9±0.1	80.0±0.3	93.4±0.1	67.3±0.4	94.8±0.3	76.8±0.2
femur left	85.5±0.5	64.3±1.1	95.0±0.2	66.2±1.2	94.4±0.2	60.6±1.5	94.5±0.2	65.8±2.1
femur right	97.8±0.0	77.8±0.7	97.9±0.1	76.8±0.4	97.3±0.0	72.3±0.3	98.2±0.0	82.5±0.3
hip left	97.5±0.0	83.4±0.4	98.0±0.1	84.3±0.2	95.9±0.6	77.0±0.3	97.8±0.1	82.9±0.4
hip right	98.2±0.0	84.9±0.4	98.4±0.0	85.8±0.1	97.3±0.1	78.2±0.2	98.2±0.0	84.5±0.3
sacrum	96.6±0.2	80.1±0.8	97.0±0.1	80.9±0.2	95.6±0.1	72.0±0.5	96.6±0.0	78.5±0.3
gluteus maximus left	96.8±0.1	63.8±0.6	97.0±0.0	66.1±0.3	96.6±0.0	62.3±0.6	96.3±0.1	59.9±0.3
gluteus maximus right	96.9±0.1	66.2±0.6	97.1±0.1	66.2±0.2	96.9±0.0	64.9±0.4	97.0±0.0	64.8±0.2
gluteus medius left	95.6±0.1	59.2±0.4	95.9±0.1	60.7±0.2	95.2±0.1	54.8±0.5	94.9±0.1	52.8±0.3
gluteus medius right	90.8±0.3	58.8±1.4	96.3±0.1	59.8±1.1	95.6±0.1	53.8±0.5	95.4±0.1	53.9±1.0
gluteus minimus left	93.6±0.1	62.2±0.5	94.1±0.0	63.7±0.3	92.0±0.1	54.4±0.4	92.4±0.2	55.9±0.4
gluteus minimus right	88.8±1.6	64.6±0.7	94.8±0.1	66.9±0.3	93.2±0.1	58.1±0.3	93.2±0.1	59.8±0.7
autochthon left	96.3±0.1	69.5±0.2	96.8±0.0	71.1±0.2	96.4±0.1	67.4±0.2	95.8±0.0	63.3±0.2
autochthon right	96.6±0.0	69.7±0.2	96.8±0.0	70.9±0.3	96.2±0.0	65.9±0.2	95.9±0.0	63.2±0.3
iliopsoas left	79.4±1.2	59.0±0.6	86.4±0.6	61.4±0.5	89.6±0.4	55.5±0.3	85.6±0.6	54.5±0.7
iliopsoas right	87.6±0.5	64.3±0.6	89.6±0.3	61.4±0.7	88.9±0.4	57.9±1.3	89.1±0.6	60.5±0.4
average	92.9±0.4	70.5±0.6	95.4±0.1	71.7±0.5	94.0±0.1	62.9±0.5	94.6±0.2	68.5±0.6
esophagus	93.4±0.1	73.9±0.3	93.8±0.1	75.8±0.3	92.4±0.1	69.4±0.2	90.0±0.2	63.3±0.4
trachea	91.2±1.4	82.2±0.2	95.5±0.4	82.9±0.2	94.6±0.4	78.7±0.3	93.5±0.4	76.3±0.4
heart myocardium	89.7±0.2	56.7±0.5	92.8±0.1	58.5±0.5	91.9±0.1	56.1±0.6	89.9±0.2	48.9±0.3
heart atrium left	93.6±0.2	66.1±0.5	95.6±0.4	65.7±0.5	95.0±0.1	61.8±0.5	94.6±0.0	55.7±0.4
heart ventricle left	94.9±0.1	55.7±0.6	95.6±0.1	57.2±0.6	94.5±0.1	53.9±0.7	93.5±0.4	48.0±0.3
heart atrium right	90.2±0.9	55.8±0.7	94.9±0.1	56.3±0.5	94.7±0.1	54.4±0.5	92.8±0.2	46.6±0.2
heart ventricle right	87.7±0.1	54.3±0.6	95.4±0.2	56.3±0.4	94.8±0.0	54.1±0.4	93.9±0.1	47.3±0.7
pulmonary artery	93.1±0.2	62.4±0.6	93.0±0.3	62.8±0.2	91.8±0.2	59.3±0.5	92.0±0.2	52.8±0.2
brain	87.5±2.8	53.4±0.7	95.0±0.6	55.0±1.5	94.8±0.5	54.0±1.2	95.6±0.6	54.7±1.8
iliac artery left	93.1±0.1	78.9±0.4	93.3±0.1	80.2±0.6	91.2±0.1	73.6±0.6	87.7±0.4	67.7±1.0
iliac artery right	88.4±1.5	78.0±0.5	93.0±0.2	80.1±0.6	90.7±0.2	72.4±0.6	87.7±0.6	67.3±1.1
iliac vena left	94.2±0.1	75.4±0.5	92.8±0.2	74.6±0.7	91.8±0.4	68.7±0.6	90.1±0.4	63.4±0.7
iliac vena right	84.8±1.4	74.4±0.5	92.7±0.3	75.2±0.6	91.3±0.3	68.7±0.7	89.9±0.3	63.3±0.9
small bowel	81.7±1.0	53.8±1.0	85.8±0.5	56.1±0.7	85.4±0.8	54.0±0.8	86.8±0.2	50.6±0.7
duodenum	78.0±1.0	47.0±0.6	84.9±0.3	48.2±0.4	82.5±0.2	45.3±0.7	81.9±0.3	40.7±0.7
colon	91.1±0.3	55.5±0.4	90.9±0.2	57.1±0.5	91.6±0.3	55.5±0.4	88.8±0.3	48.1±0.4
urinary bladder	88.3±0.8	51.0±0.8	93.3±0.9	51.5±0.5	92.6±0.3	47.6±1.1	91.2±0.4	43.5±0.6
face	77.9±0.9	50.7±0.9	80.7±0.9	49.6±0.7	80.5±1.7	45.0±0.6	85.0±0.9	38.0±1.8
average	88.8±0.7	62.5±0.6	92.2±0.3	63.5±0.6	91.2±0.3	59.6±0.6	90.3±0.3	54.2±0.7

C.2 Supervised and Self-supervised Benchmarking

C.2.1 Background

The goal of Table 2 and Appendix Table 8 is to provide a practical benchmark for the transfer learning ability of readily available pre-trained models. Our intent is not to compare the specific pre-training methodologies of each model for two primary reasons.

1. The majority of researchers tend to fine-tune pre-existing models rather than retrain them from scratch due to convenience and accessibility.
2. Reproducing these models would require specialized hyper-parameter tuning and varied computational resources. For example, models like Swin UNETR (Tang et al., 2022) were pre-trained using large-scale GPU clusters at NVIDIA, making them challenging for us to faithfully retrain.

Considering both practical user scenarios and computational constraints, we decided to directly use their released models and fine-tune them with consistent settings on the same datasets.

However, using existing pre-trained models can inevitably lead to certain problems. For example, the U-Net family has seen numerous variations over the years (Siddique et al., 2021). Pre-trained models released before 2021 typically employed a basic version of U-Net (Zhou et al., 2019; Chen et al., 2019b). On the other hand, our U-Net benefits from a more advanced code base, thanks to the MONAI platform at NVIDIA, which includes enhanced architectures and advanced training optimization strategies. Consequently, our U-Net, even trained from scratch, is capable of surpassing the performance of these older baseline models.

Table 8: **Benchmarking all the self-supervised and supervised models.**

	name	backbone	params	pre-trained data	performance†
self-supervised	Models Genesis (Zhou et al., 2019)	U-Net	19.08M	623 CT volumes	90.1
	UniMiSS (Xie et al., 2022)	nnU-Net	61.79M	5,022 CT&MRI volumes	92.9
	NV*	Swin UNETR	62.19M	1,000 CT volumes	93.2
	NV*	Swin UNETR	62.19M	3,000 CT volumes	93.4
	NV (Tang et al., 2022)	Swin UNETR	62.19M	5,050 CT volumes	93.8
	NV*	Swin UNETR	62.19M	5,050 CT volumes	94.2
	NV*	Swin UNETR	62.19M	9,262 CT volumes	94.3
supervised	Med3D (Chen et al., 2019b)	Residual U-Net	85.75M	1,638 CT volumes	91.4
	DoDNet (Zhang et al., 2021)	U-Net	17.29M	920 CT volumes	93.8
	DoDNet*	U-Net	17.29M	920 CT volumes	94.4
	Universal Model (Liu et al., 2023b)	U-Net	19.08M	2,100 CT volumes	-
	Universal Model (Liu et al., 2023b)	Swin UNETR	62.19M	2,100 CT volumes	94.1
	SuPreM*	U-Net	19.08M	2,100 CT volumes	**95.4**
	SuPreM*	Swin UNETR	62.19M	2,100 CT volumes	94.6
	SuPreM*	SegResNet	470.13M	2,100 CT volumes	94.0

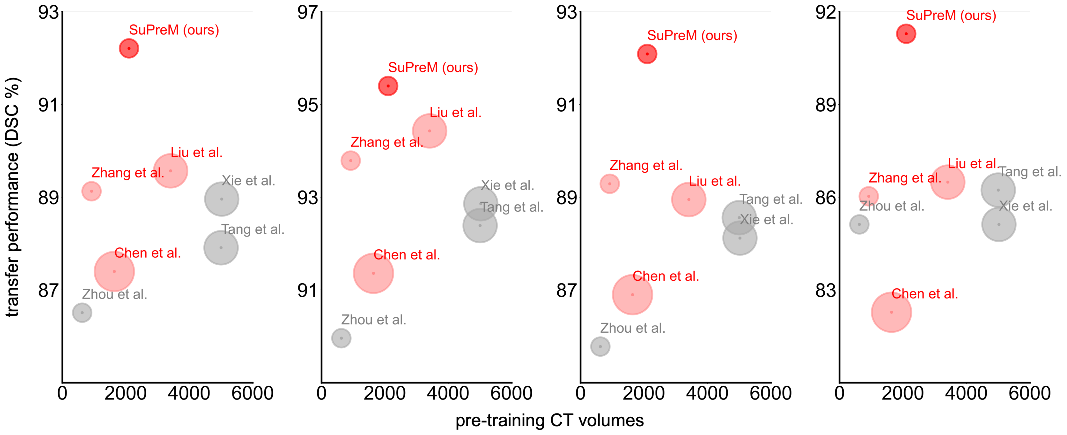
†We report the transfer learning performance of muscle segmentation on TotalSegmentator.
*The name with a star (*) denotes it is implemented by us and pre-trained using our AbdomenAtlas 1.1.

C.2.2 Implementation details of pre-training

For benchmark purposes (Tables 2, 4 and Figures 1, 2b, 3), we pre-trained U-Net, Swin UNETR, and SegResNet on 2,100 fully annotated CT volumes with 25 anatomical structures and pseudo annotations of seven tumors. The best model was selected based on the largest average DSC over the 32 classes on 310 CT volumes as the validation set. We randomly crop sub-volumes, sized 96×96×96 voxels, from the original CT volumes. Our SuPreM is pre-trained with AdamW using $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with a batch size of 2 per GPU and a cosine learning rate schedule with a warm-up for the first 100 epochs. We start with an initial learning rate of $1e^{-4}$ and a decay of $1e^{-5}$. The pre-training has been conducted on four NVIDIA A100 using multi-GPU (4) with distributed data parallel (DDP), implemented in MONAI 0.9.0., with a maximum of 800 epochs. We use the binary cross-entropy and Dice Similarity Coefficient (DSC) losses as the objective function for pre-training.

C.2.3 IMPLEMENTATION DETAILS OF FINE-TUNING

We fine-tune the pre-trained models using TotalSegmentator and the proprietary dataset datasets. During fine-tuning, configurations from pre-training persist, but we adjust the warm-up scheduler to 20 epochs, set a maximum of 200 epochs, and use a single GPU.



(a) cardiac structure (b) muscle (c) organ (d) vertebrae

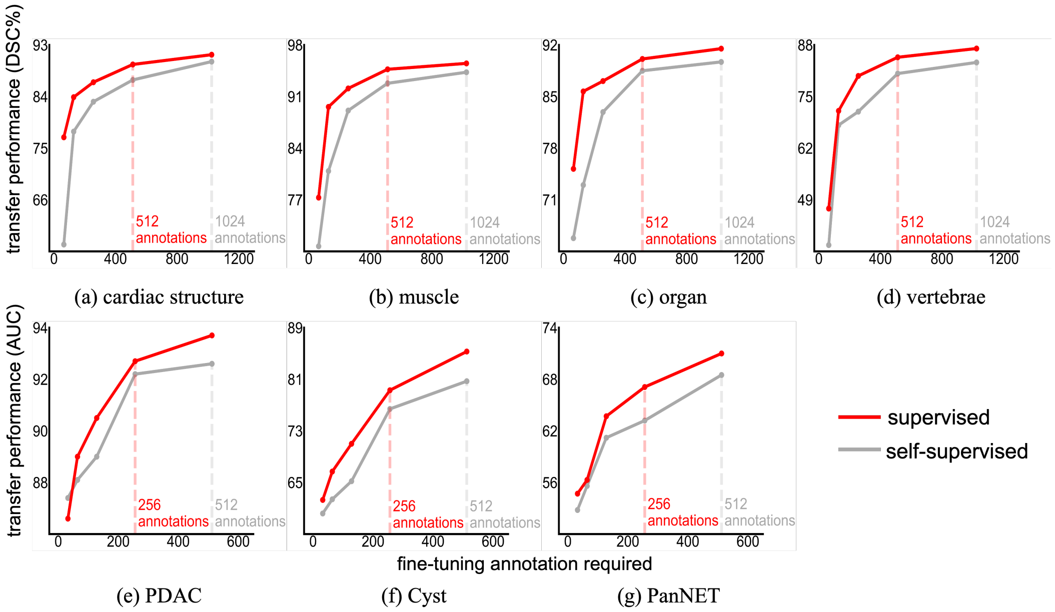
Figure 8: **A comprehensive benchmark on supervised and self-supervised models.** We present the segmentation performance achieved by fine-tuning models using the entire TotalSegmentator training set ($N = 1081$ annotated CT volumes) as target tasks. A larger circle size denotes a greater number of model parameters. Overall, for target tasks, supervised models (in red) transfer better for pre-training in comparison with self-supervised models (in gray). Detailed per-class performance can be found in Appendix Table 9.

Table 9: **Detailed results of supervised and self-supervised benchmarking on TotalSegmentator.** The performance is measured by the Dice Similarity Coefficient (DSC) and normalized surface dice (NSD) with a tolerance of 1mm.

TotalSegmentator	scratch		Tang et al. (2022)		Liu et al. (2023b)		SuPreM (ours)	
(Wasserthal et al., 2022)	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)
spleen	96.2±0.1	73.8±0.6	96.0±0.2	73.0±0.7	96.3±0.2	72.8±0.7	96.5±0.1	74.5±0.4
kidney right	95.0±0.2	76.3±0.6	95.2±0.4	75.6±0.3	94.1±0.2	74.1±0.8	94.1±0.6	71.6±0.8
kidney left	89.4±1.9	73.8±1.3	89.6±0.6	70.0±0.3	92.7±1.1	72.4±0.7	95.3±0.3	72.8±0.7
gallbladder	78.6±1.3	47.8±1.4	83.2±1.0	48.8±0.9	83.6±1.0	48.7±1.2	85.7±0.6	43.3±1.1
liver	97.6±0.0	72.3±0.3	97.6±0.1	71.9±0.4	97.7±0.1	71.1±0.6	97.7±0.2	72.0±0.4
stomach	88.4±0.2	59.5±0.8	92.5±0.5	57.7±0.6	92.8±0.3	57.1±0.5	93.1±0.5	59.0±0.4
aorta	92.7±0.7	73.1±0.4	93.2±0.1	70.0±0.4	93.2±0.5	69.1±0.4	96.1±0.1	71.9±0.4
inferior vena cava	89.3±0.3	62.8±0.6	88.0±0.4	60.0±0.3	86.4±0.2	58.5±0.5	89.4±0.2	60.5±0.7
portal & splenic vein	75.4±1.3	58.4±1.4	77.4±0.5	55.0±0.9	78.7±0.7	51.6±0.7	81.5±0.5	54.7±0.7
pancreas	86.2±0.5	58.3±1.0	85.9±0.7	56.4±0.9	86.3±0.5	56.3±0.9	88.9±0.4	56.7±0.3
adrenal gland right	82.1±0.5	69.4±0.6	82.9±1.2	66.5±0.8	82.4±0.5	61.0±0.4	87.1±0.1	64.9±0.3
adrenal gland left	75.1±1.2	59.3±1.4	74.1±0.7	56.2±0.7	77.8±0.7	52.1±0.6	79.1±0.5	55.1±0.5
lung upper lobe left	96.6±0.2	74.5±0.4	96.0±0.2	72.7±0.3	95.7±0.2	70.4±0.3	96.3±0.0	72.8±0.2
lung lower lobe left	92.3±0.7	71.5±0.2	93.1±0.4	69.0±0.6	92.6±0.6	67.1±0.5	93.5±0.5	69.3±0.4
lung upper lobe right	91.0±0.5	68.9±0.4	88.4±0.2	65.5±1.0	89.8±0.2	61.1±0.5	91.5±0.5	66.5±0.5
lung middle lobe right	90.2±0.3	60.0±0.4	90.7±0.6	56.1±0.5	89.9±0.6	54.1±0.4	91.6±0.2	56.8±0.4
lung lower lobe right	94.8±0.1	72.7±0.5	94.1±0.5	70.4±0.5	94.5±0.5	69.8±0.3	94.7±0.1	71.4±0.4
average	88.9±0.6	66.6±0.7	89.3±0.5	64.4±0.6	89.7±0.5	62.8±0.6	91.3±0.3	64.3±0.5
humerus left	88.6±1.5	69.1±1.0	93.0±0.8	78.0±0.9	92.7±0.3	73.0±0.6	93.3±0.4	72.8±1.0
humerus right	86.1±1.4	64.0±1.3	87.9±1.2	71.4±0.9	88.3±0.3	69.2±0.8	94.9±0.1	68.6±2.1
scapula left	92.4±0.2	79.4±0.9	93.5±0.6	79.8±0.2	92.4±0.2	78.0±0.7	92.7±0.5	77.1±0.5
scapula right	95.4±0.2	81.2±0.3	95.1±0.2	81.6±0.3	95.3±0.2	81.3±0.3	94.6±0.2	79.3±0.2
clavicula left	95.9±0.0	81.3±0.2	95.2±0.1	82.5±0.5	96.7±0.0	84.0±0.2	95.9±0.0	81.7±0.2
clavicula right	95.1±0.0	77.0±0.2	95.7±0.1	79.6±0.3	95.7±0.1	79.2±0.2	94.8±0.3	76.8±0.2
femur left	85.5±0.5	64.3±1.1	86.2±0.8	66.9±1.6	89.8±0.4	65.3±1.0	94.5±0.2	65.8±2.1
femur right	97.8±0.0	77.8±0.7	98.2±0.1	83.0±0.9	97.5±0.1	81.8±0.2	98.2±0.0	82.5±0.3
hip left	97.5±0.0	83.4±0.4	98.0±0.0	84.2±0.2	97.9±0.0	83.5±0.2	97.8±0.1	82.9±0.4
hip right	98.2±0.0	84.9±0.4	98.4±0.0	86.2±0.3	98.4±0.0	85.8±0.1	98.2±0.0	84.5±0.3
sacrum	96.6±0.2	80.1±0.8	96.9±0.1	81.2±0.3	97.0±0.1	81.0±0.2	96.6±0.0	78.5±0.3
gluteus maximus left	96.8±0.1	63.8±0.6	96.4±0.1	63.0±0.7	96.5±0.0	62.8±0.1	96.3±0.1	59.9±0.3
gluteus maximus right	96.9±0.1	66.2±0.6	96.9±0.1	66.9±0.7	96.9±0.0	65.1±0.2	97.0±0.0	64.8±0.2
gluteus medius left	95.6±0.1	59.2±0.4	95.3±0.1	56.9±0.5	95.1±0.1	55.6±0.3	94.9±0.1	52.8±0.3
gluteus medius right	90.8±0.3	58.8±1.4	90.9±0.3	57.8±0.7	91.2±0.2	57.2±0.4	95.4±0.1	53.9±1.0
gluteus minimus left	93.6±0.1	62.2±0.5	93.2±0.1	59.6±0.5	92.2±0.1	55.3±0.4	92.4±0.2	55.9±0.4
gluteus minimus right	88.8±1.6	64.6±0.7	93.7±0.1	63.8±0.9	93.7±0.1	61.5±0.3	93.2±0.1	59.8±0.7
autochthon left	96.3±0.1	69.5±0.2	96.2±0.0	66.3±0.4	96.4±0.0	68.0±0.2	95.8±0.0	63.3±0.2
autochthon right	96.6±0.0	69.7±0.2	96.1±0.0	65.9±0.4	96.2±0.1	66.4±0.1	95.9±0.0	63.2±0.3
iliopsoas left	79.4±1.2	59.0±0.6	84.6±0.5	60.3±0.2	86.6±0.7	62.9±0.6	85.6±0.6	54.5±0.7
iliopsoas right	87.6±0.5	64.3±0.6	87.8±0.8	64.2±0.6	89.7±0.4	65.4±0.3	89.1±0.6	60.5±0.4
average	92.9±0.4	70.5±0.6	93.8±0.3	71.4±0.6	94.1±0.2	70.6±0.4	94.6±0.2	68.5±0.6
esophagus	93.4±0.1	73.9±0.3	88.7±0.2	60.6±0.6	89.4±0.2	61.8±0.4	90.0±0.2	63.3±0.4
trachea	91.2±1.4	82.2±0.2	93.4±0.1	75.9±0.5	93.3±0.1	75.9±0.3	93.5±0.4	76.3±0.4
heart myocardium	89.7±0.2	56.7±0.5	89.0±0.1	46.0±0.5	89.5±0.1	46.9±0.4	89.9±0.2	48.9±0.3
heart atrium left	93.6±0.2	66.1±0.5	93.6±0.4	50.8±0.5	94.2±0.2	54.8±0.4	94.6±0.0	55.7±0.4
heart ventricle left	94.9±0.1	55.7±0.6	93.2±0.1	44.9±0.5	93.5±0.1	46.4±0.4	93.5±0.4	48.0±0.3
heart atrium right	90.2±0.9	55.8±0.7	92.0±0.2	42.9±0.5	92.6±0.1	44.3±0.5	92.8±0.2	46.6±0.2
heart ventricle right	87.7±0.1	54.3±0.6	90.4±0.8	43.6±0.5	93.3±0.3	44.9±0.3	93.9±0.1	47.3±0.7
pulmonary artery	93.1±0.2	62.4±0.6	88.7±0.2	51.5±0.6	91.8±0.1	52.4±0.6	92.0±0.2	52.8±0.2
brain	87.5±2.8	53.4±0.7	95.5±0.1	55.0±1.2	95.6±0.4	55.1±0.7	95.6±0.6	54.7±1.8
iliac artery left	93.1±0.1	78.9±0.4	87.1±0.4	65.4±0.6	87.1±0.4	65.5±0.5	87.7±0.4	67.7±1.0
iliac artery right	88.4±1.5	78.0±0.5	85.8±0.2	64.1±0.9	87.6±0.4	65.4±0.5	87.7±0.6	67.3±1.1
iliac vena left	94.2±0.1	75.4±0.5	87.5±0.3	58.8±0.6	88.5±0.4	60.0±0.5	90.1±0.4	63.4±0.7
iliac vena right	84.8±1.4	74.4±0.5	87.3±0.1	59.7±0.4	88.5±0.6	60.7±0.6	89.9±0.3	63.3±0.9
small bowel	81.7±1.0	53.8±1.0	84.3±0.3	48.3±0.5	84.4±0.8	48.4±0.5	86.8±0.2	50.6±0.7
duodenum	78.0±1.0	47.0±0.6	79.1±0.7	40.7±0.5	80.3±1.1	42.3±0.5	81.9±0.3	40.7±0.7
colon	91.1±0.3	55.5±0.4	87.2±0.3	45.2±0.4	87.8±0.3	45.9±0.3	88.8±0.3	48.1±0.4
urinary bladder	88.3±0.8	51.0±0.8	90.0±0.3	38.9±0.8	89.7±0.2	39.2±0.5	91.2±0.4	43.5±0.6
face	77.9±0.9	50.7±0.9	76.0±1.1	41.2±1.2	82.0±1.2	37.3±1.6	85.0±0.9	38.0±1.8
average	88.8±0.7	62.5±0.6	88.3±0.3	51.9±0.6	89.4±0.4	52.6±0.5	90.3±0.3	54.2±0.7

D Data, Annotation, and Computational Efficiency

D.1 Annotation Efficiency in Fine-tuning



(a) cardiac structure

(b) muscle

(c) organ

(d) vertebrae

(e) PDAC

(f) Cyst

(g) PanNET

Figure 9: **SuPreM detailed annotation and learning efficiency for segmenting 66 novel classes.** We assesses the annotation & learning efficiency by fine-tuning models on different number of annotated CT volumes from TotalSegmentator and the proprietary dataset for a total of 66 novel classes.

D.2 Convergence and Learning Efficiency in Fine-tuning

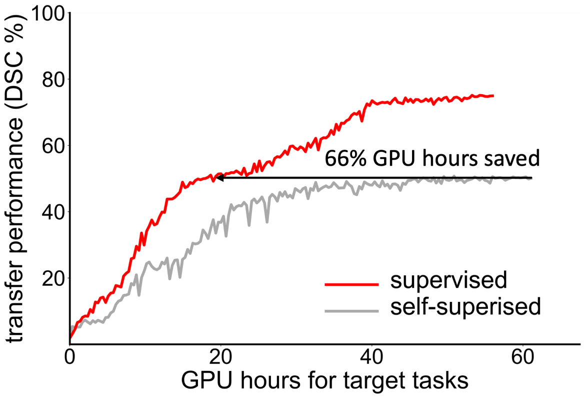


Figure 10: **Convergence & learning efficiency in fine-tuning.** We present the learning curves for fine-tuning supervised and self-supervised models for target tasks. Supervised models converge faster and achieve markedly better performance in cardiac segmentation using TotalSegmentator.

E Enhanced Features for Novel Datasets, Classes, and Tasks

E.1 Direct Inference on Three External Datasets

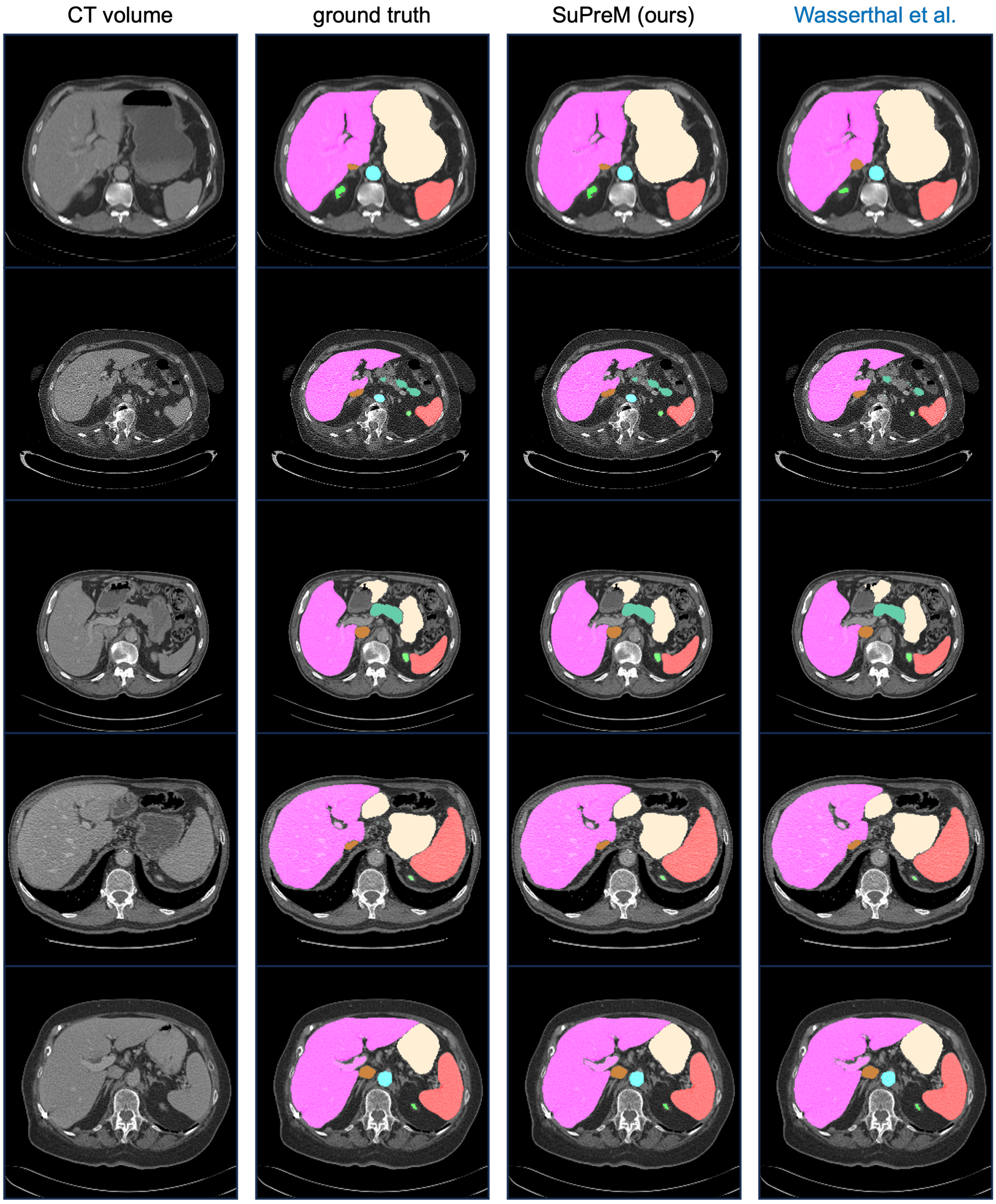


Figure 11: **Direct inference on TotalSegmentator.**

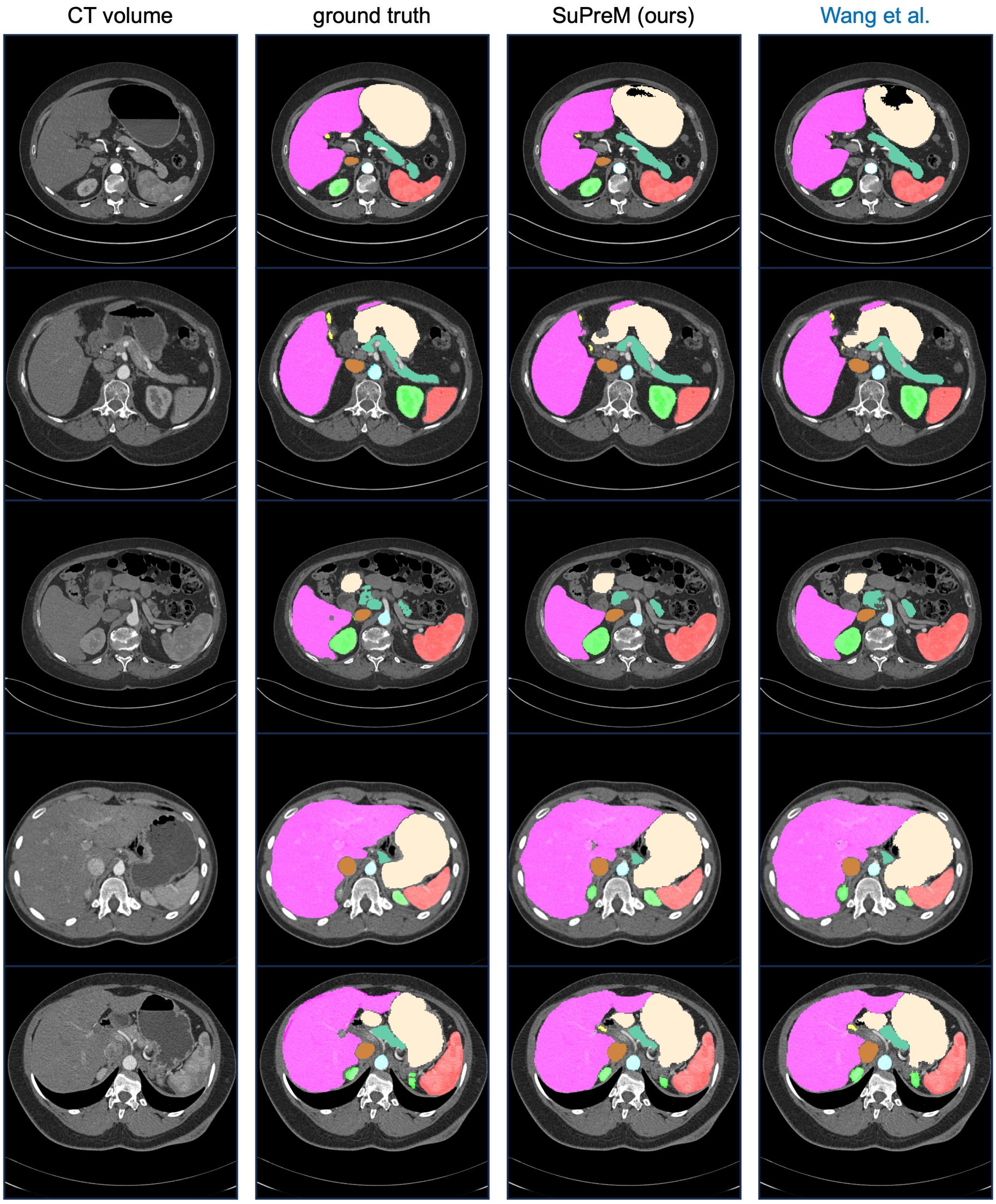


Figure 12: **Direct inference on the proprietary dataset.** We performed direct inference using the proprietary dataset, which covers 23 specific classes. These include the aorta, adrenal gland, common bile duct, celiac abdominal aorta, colon, duodenum, gallbladder, IVC, left kidney, right kidney, liver, pancreas, pancreatic duct, superior mesenteric artery, small bowel, spleen, stomach, and various veins like the left and right renal veins. It also includes PDAC, pancreatic cysts, and PanNET.

E.2 Fine-tuning SuPreM on 17 Seen Classes

Table 10: **Fine-tuning SuPreM on 17 seen classes.** The fine-tuning performance of 17 seen classes is promising, but this is expected, given that the model is exposed to more examples of these classes in both pre-training and fine-tuning phases. We perform an independent two-sample t-test between the self-supervised pre-trained model and the supervised pre-trained model. The performance gain (Δ) is statistically significant at the $P = 0.05$ level, with highlighting in a light red box.

seen class	self-super.	super.	Δ	seen class	self-super.	super.	Δ
spleen	95.9±0.1	96.5±0.2	0.6	pancreas	85.6±0.3	88.8±0.5	3.2
kidney right	95.4±0.5	93.7±0.3	-1.7	adrenal gland right	82.2±0.2	86.9±0.2	4.7
kidney left	89.5±0.7	95.1±0.1	5.6	adrenal gland left	74.0±0.4	79.2±0.7	5.3
gallbladder	82.6±0.3	85.3±0.3	2.7	femur left	86.2±0.8	94.5±0.2	8.3
liver	97.6±0.0	97.6±0.0	0.1	femur right	98.2±0.1	98.2±0.0	0.0
stomach	92.7±0.6	93.2±0.5	0.4	esophagus	88.7±0.2	90.0±0.2	1.3
aorta	93.2±0.1	96.0±0.1	2.8	duodenum	79.1±0.7	81.9±0.3	2.8
inferior vena cava	88.1±0.5	89.4±0.4	1.3	colon	87.2±0.3	88.8±0.3	1.6
portal & splenic vein	77.0±0.4	80.9±0.5	3.9				
average	87.8±0.4	90.4±0.3	2.5				

E.3 Fine-tuning SuPreM on 63 Novel Classes

Table 11: **F**ine-tuning SuPreM on 66 novel classes. Following the standard transfer learning paradigm, we fine-tune our SuPreM on several novel segmentation tasks. These tasks include segmenting 19 muscles, 15 cardiac structures, 5 organs, and 24 vertebrae from TotalSegmentator, as well as 3 pancreatic tumors from the proprietary dataset. It is important to note that these classes were not part of the pre-training of SuPreM. We observe that SuPreM, supervised pre-trained on only a few classes, can transfer better than those self-supervised pre-trained on raw, unlabeled data. In other words, it is the task of segmentation itself that can enhance the model's capability of segmenting novel-class objects. This benefit is much more straightforward and understandable than such self-supervised tasks as contextual prediction, mask image modeling, and instance discrimination in the context of transfer learning. We hypothesize that it is because the model learns to understand the concept of *objectness* in a broader sense through full supervision, as suggested by Kirillov et al. (2023), but this certainly deserves further exploration. In addition, we have further performed an independent two-sample t-test between the self-supervised pre-trained model and the supervised pre-trained model. The performance gain (Δ) is statistically significant at the $P = 0.05$ level, with highlighting in a light red box.

novel class	self-super.	super.	Δ	novel class	self-super.	super.	Δ
humerus left	92.6±0.3	93.1±0.2	0.5	lung upper lobe left	96.0±0.1	96.4±0.2	0.4
humerus right	87.3±1.0	94.9±0.1	7.6	lung lower lobe left	92.9±0.4	93.5±0.8	0.6
scapula left	93.1±0.1	92.5±0.1	-0.6	lung upper lobe right	89.1±1.2	91.5±0.3	2.5
scapula right	94.9±0.1	94.6±0.1	-0.3	lung middle lobe right	90.5±0.4	91.7±0.2	1.2
clavicula left	95.3±0.4	95.9±0.1	0.6	lung lower lobe right	94.0±0.4	94.7±0.1	0.7
clavicula right	95.6±0.1	94.7±0.1	-0.9	**average (organ)**	92.5±0.5	93.6±0.3	1.1
hip left	98.1±0.1	97.8±0.1	-0.3				
hip right	98.4±0.0	98.2±0.0	-0.2	vertebrae L5	89.6±0.7	89.0±0.7	-0.6
sacrum	96.9±0.1	96.4±0.1	-0.5	vertebrae L4	90.4±0.7	93.0±0.2	2.5
gluteus maximus left	96.4±0.1	96.0±0.1	-0.4	vertebrae L3	87.4±1.4	92.4±0.2	4.9
gluteus maximus right	96.9±0.1	96.7±0.1	-0.2	vertebrae L2	82.2±1.9	86.0±0.5	3.8
gluteus medius left	95.3±0.2	94.5±0.2	-0.8	vertebrae L1	89.0±1.1	92.3±0.3	3.3
gluteus medius right	92.3±2.4	94.8±0.3	2.5	vertebrae T12	88.9±1.1	88.6±0.3	-0.4
gluteus minimus left	93.2±0.1	92.1±0.1	-1.1	vertebrae T11	91.0±1.4	90.3±0.3	-0.7
gluteus minimus right	93.7±0.1	93.2±0.1	-0.5	vertebrae T10	91.2±1.2	90.3±0.4	-0.8
autochthon left	96.1±0.0	95.8±0.0	-0.3	vertebrae T9	87.0±1.3	89.4±0.6	2.3
autochthon right	96.1±0.0	95.9±0.0	-0.2	vertebrae T8	81.9±1.1	84.4±0.8	2.6
iliopsoas left	84.5±0.4	85.9±0.4	1.5	vertebrae T7	80.7±1.3	85.3±0.8	4.6
iliopsoas right	87.6±0.4	88.8±0.2	1.1	vertebrae T6	78.2±1.3	80.4±0.8	2.2
average (muscle)	93.9±0.3	94.3±0.1	0.4	vertebrae T5	77.2±1.8	77.8±1.5	0.6
				vertebrae T4	74.6±1.6	74.9±1.3	0.3
trachea	93.4±0.1	93.3±0.1	-0.1	vertebrae T3	82.1±1.4	81.9±1.3	-0.3
heart myocardium	88.9±0.2	89.7±0.2	0.8	vertebrae T2	85.0±0.9	86.1±1.3	1.1
heart atrium left	93.5±0.2	94.6±0.0	1.0	vertebrae T1	90.1±1.0	90.3±1.2	0.1
heart ventricle left	93.3±0.3	93.4±0.4	0.1	vertebrae C7	89.3±1.0	86.6±1.3	-2.6
heart atrium right	92.1±0.2	92.8±0.2	0.7	vertebrae C6	76.4±1.8	79.9±1.2	3.5
heart ventricle right	90.4±0.8	93.9±0.2	3.6	vertebrae C5	73.4±0.8	73.8±1.4	0.4
pulmonary artery	88.7±0.1	92.1±0.2	3.3	vertebrae C4	80.4±3.9	81.4±1.8	1.1
brain	95.6±0.4	95.5±0.4	-0.2	vertebrae C3	90.7±0.2	90.0±0.6	-0.6
iliac artery left	87.0±0.3	87.6±0.2	0.6	vertebrae C2	86.3±0.5	86.5±1.9	0.2
iliac artery right	85.8±0.5	86.8±0.6	0.9	vertebrae C1	79.5±2.3	78.9±1.1	-0.6
iliac vena left	87.5±0.7	88.9±0.5	1.4	**average (vertebrae)**	84.3±1.3	85.4±0.9	1.1
iliac vena right	87.3±0.7	88.5±0.6	1.2				
small bowel	84.5±0.9	86.9±0.6	2.4	PDAC	53.4±0.3	53.6±0.4	0.2
urinary bladder	90.1±0.9	91.2±0.5	1.1	Cyst	41.6±0.4	49.2±0.5	7.6
face	75.3±0.8	85.0±0.4	9.7	PanNet	35.4±0.8	45.7±0.8	10.2
average (cardiac)	88.9±0.5	90.7±0.3	1.8	**average (tumor)**	48.9±0.4	53.1±0.4	4.2

E.4 Fine-tuning SuPreM on Fine-grained Tumor Classification

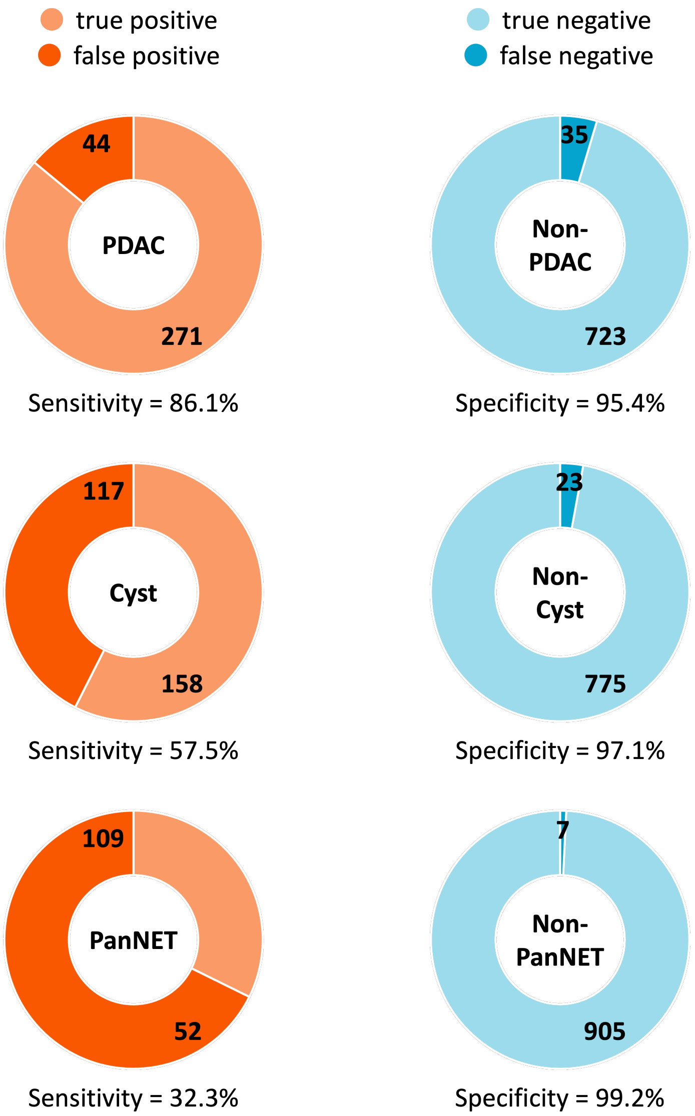


Figure 13: **Fine-grained pancreatic tumor classification.** We would like to stress the challenges in benchmarking tumor segmentation/classification, particularly due to the scarcity of annotations in publicly available datasets (often limited to hundreds of tumors). To overcome this limitation, we employed our proprietary dataset, which comprises 3,577 annotated pancreatic tumors, including detailed sub-types: 1,704 PDACs, 945 Cysts, and 928 PanNets. The proprietary dataset contains CT scans taken by a variety of vendors, e.g., Philips, Siemens, GE, and Toshiba. This extensive dataset enabled us to thoroughly assess the transfer learning ability of our pre-trained models in tumor-related tasks. Notably, the transfer learning results detailed here demonstrate a sensitivity of 86.1% and specificity of 95.4% for PDAC detection. This performance surpasses the average radiologist's performance in PDAC identification by 27.6% in sensitivity and 4.4% in specificity, as reported in Cao et al. (2023). This is one of the demonstrations of how our pre-trained models could be deployed for clinical applications.

F More Discussion

F.1 How Transferable Are Features in SuPreM?

Table 12: **How transferable are features in SuPreM?** The suite of models (to be released) provides both pre-trained encoder (▷) and decoder (◁); the encoder encodes input images into features, and the decoder decodes features back to images. We conduct an ablation study to assess how the encoder and decoder features are transferable. The white block denotes random features, and the red one denotes pre-trained encoder or decoder features. The ablation study reveals that the improved target performance is mainly attributable to the encoder features; conversely, decoder features generally do not contribute to transfer learning and often impair performance (a.k.a., *negative transfer* as reviewed in Zhang et al. (2022)). This insight is consistent across both self-supervised and supervised pre-trained models.

	pre-training	vertebrae	cardiac	muscle	organ
⬜⬜	scratch	58.6±0.9	60.4±0.7	75.9±0.5	54.7±0.8
🟥⬜	self-supervised	72.9±0.3	67.9±0.6	86.6±0.2	61.8±0.2
	supervised	73.5±0.4	78.0 ±0.2	87.1±0.1	75.3±0.2
🟥🟥	self-supervised	62.0±0.5	67.5±0.4	81.3±0.1	62.8 ±0.3
	supervised	65.6±0.3	79.8±0.5	85.2±0.1	73.3±0.4

F.2 Transfer Learning Ability Estimation

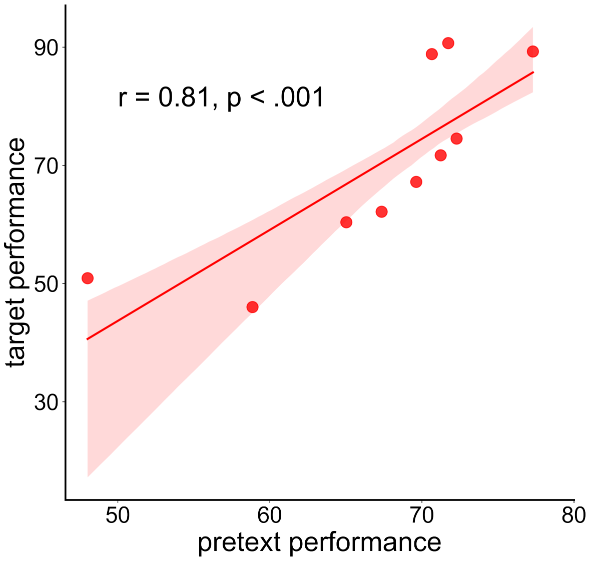


Figure 14: **Correlation analysis between pretext and target task performance.** For the pretext task, we pre-train 12 different models using AbdomenAtlas 1.1 and its per-voxel annotations. For the target task, we fine-tune these models using TotalSegmentator. The paired performance of pretext (X-axis) and target (Y-axis) tasks reveals a strong positive correlation. Specifically, the Pearson correlation coefficient (r = 0.81; p = 0.0031) suggests that the pretext performance can estimate the transfer learning ability of supervised models to some extent. This insight provides an explicit objective for effectively learning image features that are relevant to segmentation. It also offers a more precise measure for estimating model transferability than that proposed by Nguyen et al. (2020); Tan et al. (2021); You et al. (2021); Pándy et al. (2022), reducing the need for actual fine-tuning.

F.3 SUPERVISED PRE-TRAINING: IMAGE-LANGUAGE OR IMAGE-MASK PAIRS AS SUPERVISION?

We have proven that pre-training with masks as supervision transfers much better than pre-training using data only. Acquiring these masks needs extra annotations because annotating organ/tumor boundaries per voxel is not part of radiologists' workflow. In recent years, exploiting image-language pairs for pre-training has been a trending research topic because acquiring image-language pairs are easier, (e.g., from social media) than acquiring human-annotated image-mask pairs. It is also true in the medical domain—the use of radiology reports. This is because radiologists must write a report for each subject during the clinical workflow. However, we argue that image-mask pairs are expected to be more effective if the annotated datasets are already available than image-language pairs. Language is not accurate. A tumor if described in the form of radiology reports often contains information such as its rough position and size[6]. Mask is more accurate, and more expressive if incorporated with the image. In supervised pre-training, image-mask pairs, despite requiring more effort to obtain than image-language pairs, offer greater accuracy and effectiveness. For instance, a radiology report might vaguely describe a tumor's size and location, but this lacks the precision of a mask's per-voxel boundary annotations. While image-language pairs are easier to collect, especially through radiology reports, the detailed and precise information from image-mask pairs is invaluable in medical imaging, making them a more effective choice for training when annotated datasets are available.

[6] A well-circumscribed and homogeneously enhancing mass is noted at the pancreatic head, measuring 27×36×39 mm.

F.4 Broader Impact to 3D Vision Tasks?

Transfer learning across different imaging modalities, such as from CT to MRI, might be less effective compared to transfers within the same modality, primarily due to the significant differences in their imaging techniques. The discrepancies in image acquisition methods between CT and MRI result in distinct intensity values and ranges. Nonetheless, our pre-trained model could still be valuable for abdominal MRI applications. This is because the underlying anatomical structures remain consistent across both CT and MRI, allowing for the potential transfer of shared knowledge.

Given that our dataset includes detailed per-voxel annotations for 25 anatomical structures and tumors, it enables the automatic generation of 3D shape representations. These representations can be formatted as point clouds, voxel occupancy grids, meshes, and implicit surface models (e.g., signed distance functions), each catering to different algorithmic needs. We anticipate our dataset could be useful for a variety of other 3D medical vision tasks (Li et al., 2023), such as pose estimation, surface reconstruction, depth estimation, etc. Since these studies go far beyond the scope of the current manuscript and our expertise, we would like to leave the investigation as an independent work in the future.