This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

From Generalist to Specialist: Distilling a Mixture of Foundation Models for Domain-specific Medical Image Segmentation

Qing Li[#1], Yizhe Zhang[#2], Shengxiao Yang[3], Qirong Li[3], Zian Wang[3], Junhong Liu[1], Haoyang Zhang[1], Shuo Wang[4(✉)], and Chengyan Wang[1(✉)]

[1] Human Phenome Institute, Fudan University, Shanghai, China
wangcy@fudan.edu.cn
[2] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu China
[3] School of Computer Science, Fudan University, Shanghai, China
[4] Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, Shanghai, China
shuowang@fudan.edu.cn

Abstract. Segmentation foundation models (SFMs) hold promise for medical image analysis, but their direct clinical application is limited by computational cost, potentially suboptimal accuracy, and fairness concerns. In this paper, we propose a novel framework to address these challenges by distilling knowledge from a heterogeneous ensemble of pre-trained SFMs, generating specialized, high-performance models for domain-specific medical image segmentation. Unlike existing single-SFM approaches, our methodology leverages the collective intelligence of diverse SFMs to enhance accuracy, fairness, and efficiency. A key contribution is a ground-truth-free knowledge distillation strategy using the ensemble's aggregate predictions on unlabeled data to minimize reliance on manual annotation. Evaluated on a large, diverse dataset of CT and MRI scans from 702 individuals, our distilled model significantly outperforms individual SFMs and their ensemble average, achieving state-of-the-art segmentation accuracy, improved fairness across demographics (sex, age, BMI), and substantially reduced computational cost. These results offer a practical paradigm for leveraging foundation models in real-world clinical settings, mitigating key SFM limitations.

Keywords: Segmentation Foundation Models · Knowledge Distillation · Ensemble · Ground-Truth-Free

1 Introduction

Segmentation foundation models (SFMs) have emerged as a significant development in medical image analysis, demonstrating the potential to generalize

[#] Equal Contribution; [(✉)] Corresponding Author.

across diverse anatomical structures and imaging modalities, including MRI and CT [1–5]. Unlike traditional task-specific models, SFMs aim to provide a single model capable of segmenting a wide variety of objects. Prominent examples include the Segment Anything Model (SAM) and its medical variant, MedSAM, which utilize point and/or bounding-box prompts to generate class-agnostic segmentation masks [1, 2]. Other approaches, such as Segment Anything in medical scenarios, driven by Text prompt (SAT) [3] and nnU-Net-based foundation models [4, 5], operate on a predefined set of medical object classes.

Despite their promises, recent studies have revealed limitations in the accuracy, computational efficiency and fairness of SFMs in clinical applications [6–8], highlighting the need for further research to improve their practical usability. In numerous practical applications, access to a diverse set of pre-trained SFMs is readily available. However, the challenge lies in adapting these generalist models to domain-specific segmentation tasks using only unlabeled image data. Instead of focusing on developing new standalone SFMs, this paper addresses a critical, yet under-explored, challenge: *how to effectively adapt and leverage existing SFMs for specific, real-world medical segmentation tasks, particularly when ground-truth annotations are unavailable.*

We investigate a novel approach: knowledge distillation from a *heterogeneous ensemble* of SFMs to create a high-performance model for a specialized task, such as organ segmentation in CT. Our proposed approach applies the available SFMs (some requiring prompts) to generate initial segmentation predictions. These predictions, though imperfect, serve as a rich source of "soft" supervision. We hypothesize that by distilling knowledge from this *ensemble of generalist outputs*, we can train a domain-specific model that significantly outperforms any individual SFM or a naive ensemble averaging approach. This distilled model not only achieves superior segmentation accuracy but also offers substantial reductions in computational cost and demonstrates improved fairness across diverse demographic groups. The highlights of this work are summarized as follows: **(1)** We introduce a novel, probabilistic knowledge distillation framework that effectively leverages the collective knowledge of a heterogeneous ensemble of generalist SFMs to train a highly specialized and efficient segmentation model. This framework explicitly addresses the challenge of lacking ground truth labels. **(2)** The distilled specialized model demonstrably outperforms both individual SFMs and their ensemble average, achieving state-of-the-art results on the target segmentation task. Furthermore, by incorporating a fairness-aware distillation objective, we mitigate inherent biases, leading to improved segmentation equity across sensitive demographic attributes, including sex, age, and BMI. **(3)** The specialized model is intentionally designed to be significantly smaller than the SFMs, leading to substantial improvements in inference speed and reduced computational requirements.

2 Materials and Methods

Our method distills knowledge from multiple pre-trained SFMs into a specialized model for medical image segmentation. This process comprises two principal stages (Figure 1): (1) generation of diverse segmentation masks using distinct SFMs, and (2) distillation of this aggregated knowledge into a smaller, more efficient, and higher-performing student model, with an optional fairness enhancement component. We investigate and compare deterministic and probabilistic distillation strategies within this framework.

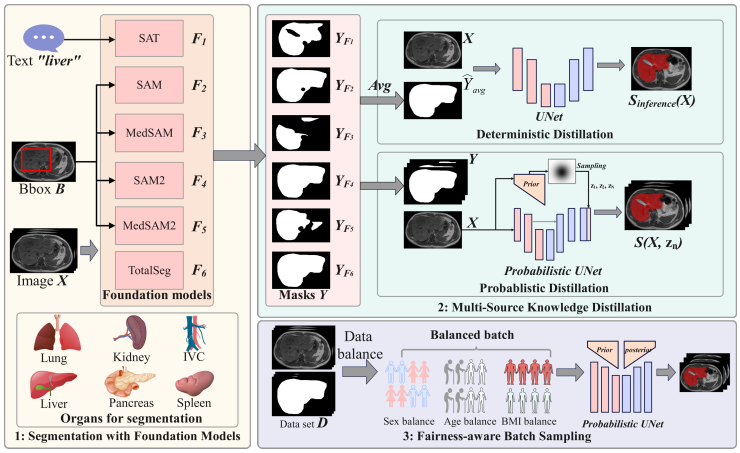

Fig. 1. Overview of the pipeline of this study.

We leverage a set of six pre-trained SFMs, denoted as $\mathcal{F} = \{F_1 : \text{SAT}, F_2 : \text{SAM}, F_3 : \text{MedSAM}, F_4 : \text{SAM2}, F_5 : \text{MedSAM2}, F_6 : \text{TotalSeg}\}$. For a given 3D medical image X_i within dataset D, and its corresponding set of 2D axial slices $\{x_{i,j}\}$, each SFM $F_k \in \mathcal{F}$ generates a segmentation mask according to its specific input requirements.. The mask generation process for each model is as follows: (1)SAT (F_1): A textual prompt, specifying the target organ "o", is provided. The 3D segmentation mask is obtained as: $Y_{i,SAT} = F_1(X_i, \text{"}o\text{"})$. (2)SAM-based Models ($F_2, F_3, F_4, F_5$): These models require a region of interest, specified as a bounding box, for each 2D slice. Let $B_{i,j,o}$ represent the bounding box delineating organ o within slice $x_{i,j}$. The 2D segmentation mask for slice $x_{i,j}$ using model F_k (where $k \in \{2, 3, 4, 5\}$) is computed as: $y_{i,j,F_k} = F_k(x_{i,j}, B_{i,j,o})$. These 2D masks are subsequently assembled to construct the 3D mask Y_{i,F_k}. (3) TotalSeg (F_6): This model operates directly on the 3D image: $Y_{i,TotalSeg} = F_6(X_i)$. The output of this stage is a set of six 3D segmentation masks, $\{Y_{i,F_1}, ..., Y_{i,F_6}\}$, for each input image X_i.

2.1 Multi-Source Knowledge Distillation

Deterministic Distillation (DD). In this baseline approach, the multiple expert masks are aggregated into a single "consensus" mask. This consensus mask serves as the pseudo-label for training a small (relative to the foundation) student model. We evaluate two representative student model architectures: UNet [9] and HSNet [10], where UNet is CNN-based model and HSNet is a Transformer-driven model (using PvT [11]).
Mask Aggregation: For each image X_i, the average of masks generated by 6 foundation model is computed: $\hat{Y}_{i,avg} = \frac{1}{6}\sum_{k=1}^{6} Y_{i,F_k}$.
Student Model Training: The student model, denoted as S, is trained to predict $Y_{i,avg}$ given the input image X_i. The loss function is a combination of Intersection over Union (IoU) loss and Binary Cross-Entropy (BCE) loss (in Eq. 1).

$$L(X_i, Y_{i,avg}) = \sum_{m=1}^{M} \left[L_{IoU}(S_m(X_i), Y_{i,avg}) + L_{BCE}(S_m(X_i), Y_{i,avg}) \right], \quad (1)$$

where $S_m(X_i)$ represents the m-th output of the student model and M represents the number of segmentation generalized. For UNet, $M = 1$. For HSNet, $M = 4$, consistent with the original HSNet publication.
Inference: For UNet inference, the final prediction is obtained by applying a sigmoid activation to the model's output: $S_{inference}(X_i) = \sigma(S(X_i))$. For HSNet, the final prediction is obtained by averaging the M outputs after a sigmoid activation: $S_{inference}(X_i) = \sigma\left(\frac{1}{M}\sum_{m=1}^{M} S_m(X_i)\right)$, where σ is the sigmoid function.
Probabilistic Distillation (PD). This strategy employs the Probabilistic U-Net (PUNet) [12], a model designed to accommodate multiple plausible segmentation masks by learning a distribution over possible outputs.
Architecture: We incorporate a latent space $\mathbb{R}^N$ ($N = 6$ in our implementation, corresponding to the number of SFMs) to represent segmentation variability. The full model consists of a prior network $P(\mathbf{z}|X_i)$, a posterior network $Q(\mathbf{z}|X_i, Y_i)$, and a U-Net backbone. *(a) Prior Network:* Generates a Gaussian distribution over the latent space, conditioned on the input image X_i: $\mathbf{z} \sim P(\cdot|X_i) = \mathcal{N}(\boldsymbol{\mu}_{prior}(X_i;\omega), \boldsymbol{\Sigma}_{prior}(X_i;\omega))$, where ω represents the parameters of the prior network. *(b) Posterior Network:* Generates a Gaussian distribution conditioned on both the input image X_i and a target mask Y_i: $\mathbf{z} \sim Q(\cdot|X_i, Y_i) = \mathcal{N}(\boldsymbol{\mu}_{post}(X_i, Y_i;\nu), \boldsymbol{\Sigma}_{post}(X_i, Y_i;\nu))$, where ν denotes the parameters of the posterior network. *(c) Segmentation Generation:* A latent vector $\mathbf{z}$, sampled from the latent space distribution, is combined with features extracted by the U-Net backbone, $f_{U-Net}(X_i;\theta)$, to produce a segmentation mask: $S(X_i, \mathbf{z}) = f_{comb}(f_{U-Net}(X_i;\theta), \mathbf{z};\psi)$ where θ are the U-Net parameters, f_{comb} represents a combining function (implemented as three sequential 1x1 convolutional layers), and ψ are the parameters of f_{comb}.
Stochastic Target Mask Selection: During training, for each image X_i, one of the six SFM-generated masks is randomly selected as the target mask. Let K be a discrete uniform random variable taking values in the set $\{1, 2, 3, 4, 5, 6\}$. The

target mask Y_i is defined as: $Y_i = \sum_{k=1}^{6} \mathbb{I}(K = k) Y_{i,F_k}$, where $\mathbb{I}(\cdot)$ denotes the indicator function.

Loss Function: The loss function encourages accurate segmentation and consistency between the prior and posterior distributions (in Eq. 2).

$$\mathcal{L}(X_i, Y_i) = \mathbb{E}_{\mathbf{z}\sim Q(\cdot|X_i,Y_i)}[-\log P_c(Y_i|S(X_i,\mathbf{z}))] + \beta \cdot D_{KL}(Q(\mathbf{z}|X_i,Y_i)||P(\mathbf{z}|X_i)), \tag{2}$$

where P_c is the likelihood of the correct segmentation, D_{KL} denotes the Kullback-Leibler divergence, and β is a weighting hyperparameter.

Inference: At inference time, given an input image X_i, we aim to generate a single, high-quality segmentation mask. This is achieved by leveraging the learned prior distribution $P(\mathbf{z}|X_i)$ to sample N(N=10 in our study) latent vectors, $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N\}$, where $\mathbf{z}_n \sim P(\cdot|X_i)$ for $n = 1, \ldots, N$. Each latent vector $\mathbf{z}_n$ is then used to generate a corresponding segmentation mask $S(X_i, \mathbf{z}_n)$. The final segmentation mask, $S_{final}(X_i)$, is obtained by averaging these individual masks: $S_{final}(X_i) = \frac{1}{N} \sum_{n=1}^{N} S(X_i, \mathbf{z}_n)$.

2.2 Fairness-aware Batch Sampling

To mitigate potential biases arising from imbalanced demographic representation, we introduce a novel, jointly-stratified batch sampling strategy for training the student model. This approach deviates from traditional stratified sampling [13], which typically focuses on balancing a single demographic attribute. Instead, our method aims to achieve a simultaneous balance across three key demographic factors: sex, age, and Body Mass Index (BMI). Below are the details of the Balanced Batch Sampling method.

Multi-Attribute Data Stratification: The training dataset D is partitioned into nine mutually exclusive queues: Sex: $D_{male} = \{x \in D \mid \text{sex}(x) = \text{male}\}$ and $D_{female} = \{x \in D \mid \text{sex}(x) = \text{female}\}$. Age (years): D_{20-30}, D_{30-40}, D_{40-50}, D_{50-60}, where $D_{a_1-a_2} = \{x \in D \mid a_1 \leq \text{age}(x) < a_2\}$. BMI (kg/m^2): $D_{BMI\leq 21}$, $D_{BMI\in(21,24)}$, $D_{BMI\geq 24}$, where, e.g., $D_{BMI\leq 21} = \{x \in D \mid \text{BMI}(x) \leq 21\}$. These queues are constructed such that $D_{male} \cup D_{female} = D$, $\bigcup_{a\in\mathcal{A}} D_a = D$ (where $\mathcal{A} = \{\text{20-30, 30-40, 40-50, 50-60}\}$), and $\bigcup_{b\in\mathcal{B}} D_b = D$ (where $\mathcal{B} = \{\text{BMI} \leq 21, \text{BMI} \in (21, 24), \text{BMI} \geq 24\}$).

Iterative Batch Construction and Refinement: For each training batch of size B, the following procedure is employed: (1) *Initial Sex Balancing:* The batch, $batch_i$, is initialized by drawing $B/2$ samples from D_{male} and $B/2$ samples from D_{female}. The remaining seven queues (D_{20-30}, D_{30-40}, D_{40-50}, D_{50-60}, $D_{BMI<21}$, D_{21-24}, $D_{BMI\geq 24}$) are updated by removing the selected samples. (2) *Age and BMI Balancing:* The batch composition is then iteratively refined to achieve balance across both age and BMI distributions, while preserving the initial sex balance. This is accomplished through a series of adjustments: (2.a) Age Balancing: If the ratio between the counts of the most and least represented age groups within $batch_i$ exceeds a predefined threshold (1.1, representing a 10% tolerance), samples are strategically removed from the over-represented age group and added from the under-represented age group. The sex information

Table 1. Segmentation Performance (DSC) of Foundation Models and Our Proposed Methods. Red bold indicates the best performance, black bold denotes the second-best, and blue represents the third-best. "Overall" refers to the average performance across six organs.

Model	Liver	Kidney	Spleen	Lung	IVC	Pancreas	Overall
SAT	0.910	0.918	0.868	0.850	0.347	0.500	0.733
MedSAM	0.784	0.882	0.818	0.886	0.723	0.462	0.760
SAM	0.878	0.892	0.838	0.955	0.774	0.585	0.821
TotalSeg	0.911	0.573	0.853	0.855	0.515	0.229	0.657
MedSAM2	0.905	0.801	0.863	0.897	0.667	0.670	0.801
SAM2	0.912	0.868	0.879	0.944	0.764	0.623	0.832
Avg. Ensemble	**0.936**	0.907	0.903	0.962	**0.814**	0.690	0.870
UNet(DD)	0.935	**0.942**	**0.909**	**0.983**	0.791	**0.704**	**0.878**
HSNet(DD)	**0.937**	**0.950**	**0.910**	**0.983**	**0.814**	**0.744**	**0.890**
PUNet(PD)	**0.939**	**0.955**	**0.918**	**0.984**	**0.832**	**0.770**	**0.900**

Table 2. Segmentation Performance (HD) of Foundation Models and our Distilled Models.

Model	Liver	Kidney	Spleen	Lung	IVC	Pancreas	Overall
SAT	8.508	5.219	7.264	15.401	17.69	16.029	11.685
MedSAM	22.087	4.025	14.213	22.441	4.505	20.108	14.563
SAM	18.485	5.062	13.423	9.449	5.917	18.508	11.807
TotalSeg	12.528	7.434	**4.108**	60.647	11.583	18.570	19.145
MedSAM2	9.500	6.841	9.932	22.066	7.012	9.606	10.826
SAM2	13.199	6.294	8.351	14.503	6.808	17.038	11.032
Avg. Ensemble	**7.551**	4.251	5.613	8.631	4.771	9.911	6.788
UNet(DD)	**7.595**	**3.751**	**3.849**	**1.797**	**4.513**	**7.947**	**4.909**
HSNet(DD)	8.199	**2.005**	4.534	**2.959**	**3.835**	**6.576**	**4.685**
PUNet(PD)	**5.898**	**2.225**	**4.017**	**1.838**	**3.814**	**5.359**	**3.859**

of removed samples is recorded, and replacement samples are drawn from the corresponding sex queue within the under-represented age group. (2.b) BMI Balancing: A similar iterative process is performed for BMI. If the ratio between the most and least represented BMI groups exceeds the threshold (1.1), samples are removed and added. In step, both the sex and age of removed samples are recorded, and replacement samples are selected to match these attributes, ensuring that the balance achieved in previous steps is maintained. After each sample addition or removal, all nine queues are updated to reflect the current state of the dataset. (3) *Termination:* The batch construction process for a given epoch terminates when any of the nine demographic queues contains an insufficient number of samples to maintain the desired balance during further refinement steps (e.g., fewer than $B/2$ samples remaining in D_{male} or D_{female}).

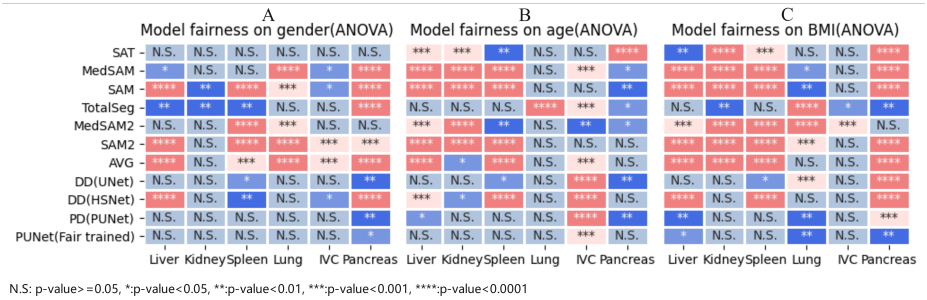

Fig. 2. Group-Level Fairness of Various Segmentation Foundation Models and Our Distilled Models (Blue and N.S. indicate strong fairness and Red indicates severe fairness issues).

Table 3. Multi-Attribute Balanced Batch Sampling Enhances Fairness. Black bold denotes the group bias (in STD) of PUNet performance is reduced via fair training, represents stronger fairness

Attribute	Models	Liver	Kidney	Spleen	Lung	IVC	Pancreas
Sex	PUNet	0.0045	0.0009	0.0033	0.0009	0.0014	0.0152
	Fair trained	**0.0036**	**0.0002**	**0.0020**	**0.0004**	0.0071	**0.0146**
Age	PUNet	0.0050	0.0044	0.0057	0.0011	0.0296	0.0174
	Fair trained	**0.0033**	**0.0040**	**0.0052**	**0.0009**	**0.0253**	**0.0148**
BMI	PUNet	0.0069	0.0024	0.0063	0.0016	0.0079	0.0201
	Fair trained	**0.0048**	**0.0022**	**0.0048**	**0.0016**	**0.0029**	**0.0131**

3 Experiments and Results

We curated a private MRI/CT dataset (702 scans; 291 male, 411 female) for foundation model evaluation, ensuring no prior exposure during model training. The dataset includes MRI scans of seven anatomical structures (liver, kidney, spleen, pancreas, IVC) and CT scans of lungs. Participants (20-60 years) were stratified by age (20-30: 43.7%, 30-40: 22.2%, 40-50: 16.7%, 50-60: 17.4%) and BMI ($<$21: 28.8%, 21-24: 39.2%, $>$24: 32.0%). Ethically approved and annotated by two experienced radiologists (each with over five years of experience), the dataset was divided into knowledge distillation set (with no ground truth) and testing set (with a 1:1 ratio), ensuring that both sets preserved the original distribution of sex, age, and BMI.

Performance in Overall Accuracy. To evaluate the segmentation performance, DSC was calculated for models among 6 organs. As shown in Table 1, the knowledge-distilled models (DD (UNet), DD (HSNet), and PD (PUNet)) consistently and significantly outperformed all foundation models and their ensemble average($\frac{1}{6}\sum_{k=1}^{6} Y_{i,F_k}$) across all six organs on DSC. Notably, PD (PUNet) achieved the highest DSC results in every organ, demonstrating the substantial performance gains achievable through the proposed multi-source knowledge distillation. This highlights the effectiveness of distilling knowledge from multiple,

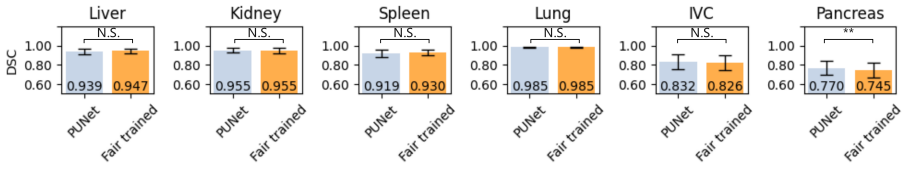

Fig. 3. No (statistically) significant segmentation performance drop when applying fairness-enhanced training in PD (PUNet).

Table 4. Model Size and Inference Cost for 512x512 Input.

Models	SAT	SAM	MedSAM	SAM2	MedSAM2	TotalSeg	UNet	HSNet	PUNet
Param(M)	878.59	635.63	90.48	216.90	31.42	49.08	2.26	29.85	5.00
FLOPs(G)	1090.18	2736.62	371.99	2813.98	106.72	70.41	55.65	23.15	96.95

diverse foundation models into specialized, high-performing segmentation models. The superiority of our method was further validated on HD (Table 2), where the knowledge-distilled models performs better than other foundation models and PD (PUNet) achieved the lowest HD values across most organs. This dual dominance in both DSC (structural overlap) and HD (boundary accuracy) underscores the robustness of multi-source distillation in preserving anatomical fidelity while minimizing segmentation errors.

Performance in Group-level Fairness and Model Efficiency. To quantify segmentation performance disparities among groups, group-wise DSC was calculated based on sex, age, and BMI attributes. One-way ANOVA [14, 15] was applied to DSC distributions, with $p < 0.001$ (***) indicating significant disparities (unfairness). As shown in Fig. 2, our experiments focusing on six different organs revealed significant fairness differences in AI-driven medical image segmentation across gender, age, and BMI. Models like SAT and TotalSeg, while exhibiting lower overall segmentation accuracy, demonstrated better fairness. The multi-model averaging ensemble (AVG) enhanced segmentation performance but did not lead to clear improvements in fairness. In contrast, significant improvements were achieved using our multi-source knowledge distillation method, resulting in models (DD: UNet, HSNet; and PD: PUNet) that surpassed the AVG ensemble in fairness metrics across all attributes, with PD (PUNet) showing the most prominent gains in achieving greater fairness. In addition, we show that employing multi-attribute balanced batch sampling during PUNet's training resulted in further fairness enhancements across all six organs and all three attributes (gender, age, and BMI), as further confirmed by the reduced standard deviation (STD) [7, 8, 13] in the mean DSC for each group presented in Table 3. The proposed fairness-enhancing training significantly reduced inter-group performance differences and, as illustrated in Fig. 3, maintained overall segmentation accuracy. Finally, as shown in Table 4, the distilled models are generally smaller in size and more efficient to run than the segmentation foundation models.

4 Conclusion

In conclusion, we proposed a multi-source distillation framework that adapts segmentation foundation models (SFMs) for medical tasks without ground-truth labels by leveraging heterogeneous SFM ensembles. Our results demonstrate that this approach not only significantly improves performance on multi-organ segmentation compared to individual SFMs and their ensemble average, but also significantly enhances fairness across demographic attributes. Specifically, the distilled model incorporating multi-attribute balanced batch sampling exhibited superior performance and fairness improving computational efficiency. The demonstrated ability to create specialized, high-performing, and fair models from readily available generalist SFMs, without relying on ground truth labels, represents a significant advancement towards realizing the clinical potential of foundation models, with future work focusing on broader applicability and theoretical underpinnings of fairness improvements.

Acknowledgments. The computations in this research were performed using the CFFF platform of Fudan University. This work was supported in part by the Shanghai Municipal Science and Technology Major Project (Grant No.2023SHZD2X02A05), the National Natural Science Foundation of China (Grant No.62331021, 62201263), the Natural Science Foundation of Jiangsu Province(Grant No.BK20220949) and the Shanghai Sailing Program(No.20YF1402400, 22YF1409300).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alexander Kirillov, Eric Mintun, Piotr Dollár, Ross Xiong, Ananya Moitra, Lei Jin, Jonathon Rae, Mircea Lavin, Ruslan Salakhutdinov, and Piotr Mirowski. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.
2. Jun Ma, Li Zhang, Yang Bai, Wen Wang, and Wei Han. Medical sam: Segment anything model for medical image analysis? *arXiv preprint arXiv:2304.12306*, 2023.
3. 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.
4. Fabian Isensee, Jens Petersen, Alexander Klein, Felix Zimmer, Niki Christodoulou, Hubert Handels, Pascal Kieslich, Jens Segsa, Alexander M Böhm, et al. nnu-net: A self-adapting framework for u-net-based medical image segmentation. *Nature methods*, 18(2):203–211, 2021.
5. Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023.
6. Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024.

7. Qing Li, Yizhe Zhang, Yan Li, Jun Lyu, Meng Liu, Longyu Sun, Mengting Sun, Qirong Li, Wenyue Mao, Xinran Wu, et al. An empirical study on the fairness of foundation models for multi-organ image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 432–442. Springer, 2024.
8. Zikang Xu, Jun Li, Qingsong Yao, Han Li, and S Kevin Zhou. Fairness in medical image analysis and healthcare: A literature survey. *Authorea Preprints*, 2023.
9. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.
10. Wenchao Zhang, Chong Fu, Yu Zheng, Fangyuan Zhang, Yanli Zhao, and Chiu-Wing Sham. Hsnet: A hybrid semantic network for polyp segmentation. *Computers in biology and medicine*, 150:106173, 2022.
11. Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.
12. Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018.
13. Esther Puyol-Antón, Bram Ruijsink, Stefan K Piechnik, Stefan Neubauer, Steffen E Petersen, Reza Razavi, and Andrew P King. Fairness in cardiac mr image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 413–423. Springer, 2021.
14. Jie M Zhang and Mark Harman. " ignorance and prejudice" in software fairness. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 1436–1447. IEEE, 2021.
15. Alessa Angerschmid, Jianlong Zhou, Kevin Theuermann, Fang Chen, and Andreas Holzinger. Fairness and explanation in ai-informed decision making. *Machine Learning and Knowledge Extraction*, 4(2):556–579, 2022.