One-Prompt to Segment All Medical Images

Supplementary Material

1. One-Prompt Data

1.1. Data Source

We provided the details of our data source on our website.

1.2. Data Preprocessing

The One-Prompt Model primarily focuses on processing 2D inputs with a concentration on single-target segmentation. In the case of 3D images, we extract the 2D mid-slice from the 3D volume along all axes. For tasks involving multiple segmentation targets, each target is treated as an individual task for predicting a binary segmentation mask. During the inference stage for predicting multiple targets, such as in the 'segment everything' setting, we predict a soft segmentation mask with a fixed threshold (averaging 0.5) to filter out uncertain predictions.

The original 3D datasets contain a variety of CT and MRI images stored in DICOM, nrrd, or mhd formats. For ensuring uniformity and compatibility, all images, regardless of modality, were converted to the widely used NifTI format. This conversion included grayscale images such as X-Ray and Ultrasound. RGB images depicting endoscopy, dermoscopy, fundus, and pathology, all converted into the PNG format.

Notably, image intensities varied significantly across modalities. For example, CT images ranged from -2000 to 2000, MRI values ranged from 0 to 800, endoscopy/ultrasound images from 0 to 255, and some modalities were already in the 0 to 1 range. To make all this data work together, we did some normalization steps for each type. Intensity normalization was systematically conducted to establish a shared intensity range. The default setting for all modalities, during training and inference, each image is normalized independently by first subtracting its mean and then dividing by its standard deviation. For MRI, X-Ray, ultrasound, mammography, and Optical Coherence Tomography (OCT) images, we trimmed intensity values to fall between the 0.95th and 99.5th percentiles before the normalization. If cropping leads to a 25% or more average size decrease, a mask for central non-zero voxels is generated, and normalization is confined to this mask, disregarding surrounding zero voxels. For CT images, we first normalized Hounsfield units using window width and level values before the standard normalization. In addition, since the intensity values are quantitative and reflect physical properties of the tissue in CT images. we conduct a global normalization scheme that is applied to all images. In specific, we use the 0.5th and 99.5th percentiles of the foreground voxels for clipping as well as the global foreground mean and standard deviation for the normalization.

To standardize the image size, we first crops the provided samples to their non-zero region, then uniformly resize them to 256 x 256. During the resizing process, we used bi-cubic interpolation for images and nearest-neighbor interpolation for masks, ensuring smooth standardization and compatibility across all images. On 3D images, we generally operate on the two axes with the highest resolution. If all three axes are isotropic, the two trailing axes are used for slice extraction. The channel is replicated threefold for the consistency processing. For slice-based processing, no resampling along the out-of-plane axis is required. In our experiments, we observed that maintaining a higher resolution for the image and processing it using a sliding window approach (splitting a large image into smaller patches, processing each patch individually, and then combining the results) can lead to quantitative improvements. However, we did not apply it in our comparison experiments as our primary focus in these experiments was to showcase the algorithmic contributions. Various engineering tricks can be easily applied over the algorithm to enhance results in practical applications after then.

Masks with multiple classes are processed to individual masks for each class. Masks with multiple connected components were dissected. We keep the original masks in situations where masks have only one component. Additionally, we exclude masks where the target area is less than 0.153% of the total image, equivalent to areas smaller than 100 pixels in a resized 256×256 resolution. This deliberate decision ensures that the dataset only includes significant and well-defined target areas. This standardized preprocessing pipeline is consistently applied across all compared methods to ensure a fair and unbiased comparison.

1.3. Data Augmentation

During training, we employ a range of data augmentation techniques, all dynamically computed on the CPU. Spatial augmentations are applied to 2D images or slices, encompassing rotations, scaling, Gaussian noise, Gaussian blur, adjustments of intensities and contrast, simulation of low resolution, gamma correction, and flipping. For enhanced image variability, most augmentations involve varying parameters randomly from predefined ranges. The application of these augmentations is stochastic, adhering to predefined probabilities. We maintain consistent parameters across different datasets. Importantly, each augmentation is individually applied to the template sample and the query sample.

Details of the augmentation are shown below:

1. **Rotation:** Rotation is applied with a probability of 0.15

to all images. The angle of rotation is uniformly sampled from [-25, 25].

2. **Scaling:** Scaling is implemented via multiplying coordinates with a scaling factor. Thus, scale factors smaller than one result in a "zoom out" effect, while values larger than one result in a "zoom in" effect. The scaling factor is uniformly sampled from [0.7, 1.4]. Scaling is applied with a probability of 0.15.
3. **Gaussian Noise:** Zero-centered additive Gaussian noise is independently added to each sample. This augmentation has a probability of 0.15. The variance of the noise is drawn from [0, 0.1], considering that intensities in all samples are close to zero mean and unit variance due to intensity normalization.
4. **Gaussian Blur:** Blurring is applied with a probability of 0.15 per sample. On each task, blurring occurs with a probability of 0.5 for each modality. The size of the Gaussian kernel is uniformly sampled from [0.5, 1.5] for each modality.
5. **Intensities:** Image intensities are either multiplied by a factor uniformly sampled from [0.65, 1.2] with a probability of 0.15 or flipped using $1 - image$ for each. Notably, the intensity augmentation is not applied to the labels. After the multiplication, the values will be clipped to their original value range.
6. **Low Resolution:** Applied with a probability of 0.25 per sample and 0.5 per associated modality, this augmentation downsamples triggered modalities by a factor uniformly sampled from [1, 2]. We use nearest neighbor interpolation for downsampling and then sample them back up to their original size with cubic interpolation.
7. **Gamma Augmentation:** Gamma Augmentation is applied With a probability of 0.15. First, image intensities are scaled to a range of 0 to 1. Then, a nonlinear intensity transformation is implemented as $x_{\text{new}} = x_{\text{old}}^{\gamma}$, with γ uniformly sampled from [0.7, 1.5]. The intensities are subsequently scaled back to their original value range. With a probability of 0.15, this augmentation is applied following the flip intensities as described above.
8. **Spacial Flip:** All samples are flipped with a probability of 0.5 along all axes.

2. Discussion of Related Tasks

2.1. Task-tailored Medical Image Segmentation Models

Unlike natural image segmentation, medical image segmentation has historically relied on task-tailored models, which are still widely used in both academia and clinical practice [2, 3, 26]. Task-tailored models offer clear advantages, because of the model optimization based on the unique characteristics of each specific task, resulting in higher performance. The main rationale behind this lies in the significant differences between medical images compared to natural images. Different medical images, like colorful fundus images and abdominal MRI images exhibit distinct features. Fundus images contain ambiguous structures like the optic cup with smooth contours, while abdominal MRI images depict organs like the pancreas, which have clear boundaries but complex structures. As a result, task-tailored models can achieve higher performance by specifically addressing the unique challenges of each task. For instance, uncertain-aware modules have been utilized to effectively handle the ambiguity associated with the optic cup in fundus images [12]. Additionally, dynamic convolution kernels have been proposed to adapt to the distorted structure of the pancreas in abdominal MRI images [9]. It is worth noting that the process of designing a unique model for each dataset in task-tailored models can be highly labor-intensive. Additionally, some tasks may have limited annotated data available for training well-performing models.

2.2. Interactive Segmentation on Medical Images

In medical image segmentation, an interactive system that allows clinicians to prioritize areas of interest provides a more immersive and personalized experience. In a single fundus image, for example, there are often complex and overlapping structures such as vessels, optic disc, optic cup, and melanoma. Interactive segmentation can greatly assist clinicians in efficiently differentiating target tissues from these intricate structures. However, most previous methods use task-tailored models. These models take a task-specific query image and a user-provided prompt to predict the target on the image based on the given prompt. A task-tailored interactive model is essentially unnecessary in medical image segmentation. Since the organs on a specific kind of image are often limited, it suffices to use a task-tailored model to segment all the organs and select the target based on user-provided prompts. This can be achieved through simple post-processing and does not require model-level interaction.

The recent success of the Segment Anything Model (SAM) [14] in interactive nature image segmentation has sparked renewed interest in the field. Many methods [4, 7, 22, 27] are being developed to fine-tune SAM for medical image segmentation, adapting its interactive approach to this domain. Few of these methods are proposed as universal interactive models [4, 16], that are able to adapt to the unseen tasks in the inference. However, they still acquire the uses to provide the prompt for each given query image, which is time and effort consuming in the practical usage.

2.3. Few/Zero-shot Image Segmentation

Few and zero methods often refer to the methods that can adapt to new tasks from few training examples, by

fine-tuning pretrained networks. Though its wide application in both natural image[19, 23] and medical image [8, 15, 20, 25], they often perform on the specific pretrained model on a specific task, and then generalize to new classes in a particular subdomain, like abdominal CT or MRI scans. In addition, they contain the re-training procedure that needs to update the model parameters again based on the given samples. Recent progress in medical image segmentation[1] extend the method by training a foundation model on varies tasks, then adapt to the unseen tasks during the inference without training.

However, this approach poses challenges in real clinical practice as it relies solely on segmentation labels to specify the task. Unfortunately, segmentation labels are unavailable in many cases. Additionally, it demands multiple samples as support to achieve satisfactory performance. Our model distinguishes itself by requiring only a single prompted sample as a template to adapt to unseen tasks. Users can provide flexible sparse prompts beyond segmentation to address a variety of clinical tasks.

3. Prompt Details

We offer four distinct prompt types which can deal with various clinical tasks. An illustration of our prompts is presented in Fig. 1.

3.0.1 Doodle Prompt

Doodle accepts a sequence of user-drawn doodles. The doodles can be categorized to one positive doodle and one negative doodle. The positive doodles are supposed to be drawn over the anatomies, while the negative doodles are drawn outside of the anatomies. We represent the doodles by a series of sampled points. Overlapping points are only included once. All of the points are concatenated into a vector. Each point is associated with a prompt label that indicates whether it is positive or negative. Then we use positional encoding to embed the positive and negative point coordinates to prompt embedding p^1 and p^2 respectively.

3.0.2 SegLab Prompt

SegLab allows the user to provide a segmentation mask as the prompt. In order to abstract the segmentation prompt into an embedding, we train an autoencoder-based structure learning to restore the given segmentation, and use its encoded embedding as the prompt.

Specifically, we concatenate the supportive image and the given segmentation mask. Then we tokenize the input as a sequence of discrete tokens obtained by an encoder. We find using this way to represent the binary segmentation map will be more efficient than representing it in the continuous space. Following [21], we use the image tokenizer learned by discrete variational autoencoder (dVAE). There are two modules during visual token learning, namely, tokenizer and decoder. The tokenizer maps image pixels into discrete tokens according to a visual codebook. The decoder learns to reconstruct the input segmentation based on the visual tokens. In order to train the discrete latent visual tokens, we use Gumbel-softmax relaxation [11] to train the model parameters. We tokenize each image to a 14×14 grid of visual tokens, and set the vocabulary size as 8192.

This autoencoder is trained on our training set and four extra unlabeled datasets. These are the RadImageNet dataset[18], a large-scale collection containing 1.35 million radiology images (CT, MRI, US) covering a wide range of organs such as the ankle/foot, brain, hip, knee, shoulder, spine, abdomen, pelvis, chest, pelvis, and thyroid. The EyePACS dataset[5], containing 88,702 color fundus images captured under various conditions by various devices at multiple primary care sites. The BCN-20000 dataset[6] and the HAM-10000[24] dataset, containing approximately 30,000 dermoscopic images with melanoma or nevus on the images.

3.0.3 Click Prompt

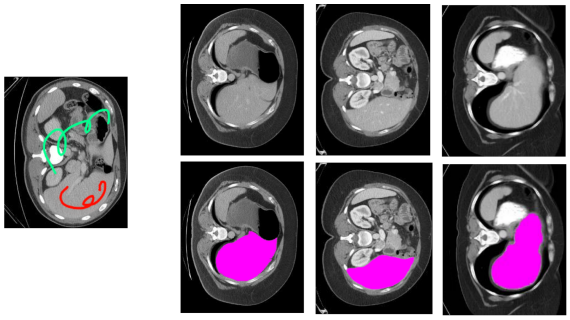
Click allows the user to click some regions over the image as the prompts. The point prompts can be categorized into two types: positive points indicate foreground regions, and negative points indicate background regions. The users need to provide at least one positive point when prompting. In this case, no positional encoding will be added to the learnable embedding which represent the background (i.e. p^2).

3.0.4 BBox Prompt

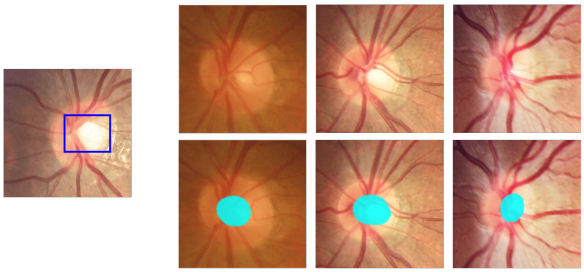
BBox allows the user to select some regions with bounding box as the prompt. The users are supposed to ensure that the major part of the target is covered by the BBox, but do not need to ensure that the whole object is in the BBox. Following SAM [14], we represent one BBox by its left-top and right-bottom corner points, and use them as two prompt embedding respectively.

3.1. Prompt Simulation

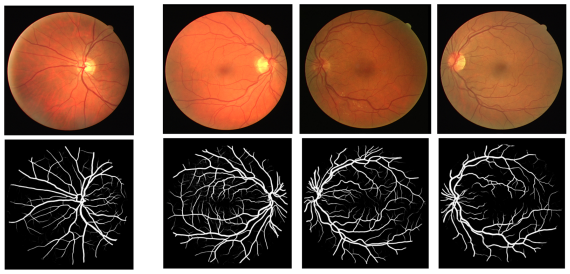
In this manuscript, we simulate prompts to compare them with interactive segmentation models and test the model performance under different prompt qualities. In order to simulate the prompts with different qualities, we first generate *Oracle* prompts from the ground truth. On *Click* and *BBox* prompts, these prompts are typically the center point of the object or the smallest box covering the object. In cases where we aim to segment a larger tissue (e.g., optic disc) that encompasses a smaller inner tissue (e.g., optic cup), we generate the point with the largest sum of distances between the boundary of the outer tissue and the boundary



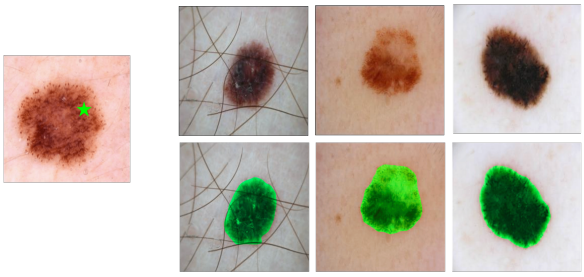
Liver Segmentation with Doodle Prompt



Optic Cup Segmentation with BBox Prompt



Vessel Segmentation with SegLab Prompt



Melanoma Segmentation with Click Prompt

Figure 1. Typical user-cases employing four prompt types to address diverse medical segmentation tasks. The prompts flexibility enables the easy adaption for various clinical practices.

of the inner tissue to prompt segmentation of the larger tissue. Then we jet the points (top left and bottom right points for *BBox*) on both axis with standard deviation equal to 6%, 15%, 25% of the target side length for simulating the *High*, *Medium*, and *Low* prompts respectively. On *Doodle* prompt, we use [13] to simulate the scribble on ground-truth with 0%, 4%, 10%, and 20% variance as *Oracle*, *High*, *Medium*, and *Low* prompts. On the *SegLab* prompt, we use nnUnet [10] with early-stop training to predict the segmentation about 85%, 70%, and 55% dice score as *High*, *Medium*, and *Low* prompts. Finally, we compare these setting with *Human*, which indicates the human prompts or the annotations. The *Oracle* and *Human* are set as the same on *SegLab* prompt.

To simulate multiple *Click* or *Doodle* prompts on a single image, we employ an iterative sampling strategy. Starting with an initial prompt generated through the mentioned procedure, we incorporate one or more new prompts on the image using an iterative sampling process. This strategy mimics real user interaction by placing each new click/scribble in the error-prone region of a prediction generated by the network using the set of previous prompts. We follow [17] for simulating the iterative sampling process.

3.2. Prompts Annotation Details

In the training set annotation process, a team of clinicians was given access to four user-friendly prompts for annotating the target in medical image segmentation. These prompts, designed to cover a wide range of cases, proved convenient for clinicians without technical backgrounds. Among them, *Click*, *BBox*, and *Doodle* are designed for human interaction, and we strongly recommend clinicians to use them if available. *SegLab* annotations are derived directly from the ground truth, but annotators are instructed not to choose it unless they've attempted prompting the image multiple times and are convinced that the target is not promptable.

Analyzing the use conditions of these prompts reaffirms their effectiveness in medical segmentation interaction. Our statistical results indicate that clinicians take an average of 0.8 seconds for a single-click prompt, 1.2 seconds for a BBox prompt, and 1.8 seconds for a Doodle prompt. This quick interaction demonstrates that clinicians can easily engage with the model using the provided prompts. Importantly, we did not provide any prior training to clinicians on how to interact with the model. They learned on their own through our user inference design. This further em-

phasizes that the given prompts can be easily and naturally understood by humans in the context of human-computer interaction.

In the test set annotation, regular individuals and junior clinicians were involved to simulate real-world user scenarios. This simulation offers a reliable evaluation for cases such as education for junior clinicians or rough annotations by regular individuals. In this scenario, users take an average of 1.4 seconds for a single-click prompt, 2.9 seconds for a BBox prompt, and 3.8 seconds for a Doodle prompt. While users in this case take a bit more time compared to senior clinicians, the duration remains reasonable, considering our model adapts to an entirely unseen task with just one prompt.

4. Implementation

We train and test the algorithm on Pytorch platform. The training process is distributed across 64 NVIDIA A100 GPUs. Our optimization is carried out using the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a linear learning rate warmup and a cosine decay for learning rate adjustments. The batch size is set at 128 images. To ensure a diverse range of tasks and prompts during training, we adopt a strategy that avoids equal sampling across all tasks. Given that certain image modalities, tasks, or prompt types are more prevalent than others, our goal is to prevent overfitting to these dominant elements. To achieve this, we uniformly select tasks and sequential states from all possible choices. For instance, we start by randomly selecting a task from the available options. Subsequently, we narrow down the selection pool to include only data associated with that chosen task. Following this, we randomly pick an image modality from all the modalities available for the selected task. We continue this process until we get a homogeneous filtered pool, and then we can randomly select an individual sample from it.

5. Analysis

5.1. Distribution differences between pre-training and testing data

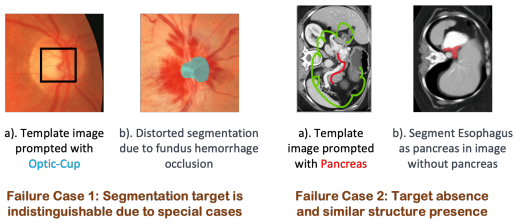
We've further assessed the distribution differences between our pre-training and test sets. We employed the Fréchet Inception Distance (FID) to compare pre-training and test sets similarities in Table 1. FID measures the difference in Fréchet Distance between feature vectors extracted from the Inception-v3 model. Recognizing its potential limitations with medical images, we also computed a 'Mededical-FID' using our pre-trained VAE model. We calculated FID scores between the pre-training and test sets, as well as between two random halves of the pre-training set, as detailed in the table below: The significant variance in average FID scores between training/test and half-train/half-train sets suggests major differences between training and test distributions.

Table 1. FID between pre-training and test sets

	Train v.s. Test	Half Train v.s. Half Train
FID	18.46	3.01
Medical-FID	33.87	7.62

5.2. Failure cases and limitations.

Our method may produce errors in two specific cases. Firstly, if the segmentation target is indistinguishable due to unique conditions, as shown in Failure Case 1, where hemorrhage obscures the optic cup. Secondly, there's a risk of false positives, such as in Failure Case 2, where the model mistakenly identifies the esophagus as the pancreas when the pancreas isn't present in the image. We have included detailed discussions and examples of failures in our revision.



References

[1] 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. 3

[2] Bingzhi Chen, Yishu Liu, Zheng Zhang, Guangming Lu, and Adams Wai Kin Kong. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023. 2

[3] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[4] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023. 2

[5] Mohamed Chetoui and Moulay A Akhloufi. Explainable end-to-end deep learning for diabetic retinopathy detection across multiple datasets. *Journal of Medical Imaging*, 7(4): 044503–044503, 2020. 3

[6] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019. 3

[7] Guoyao Deng, Ke Zou, Kai Ren, Meng Wang, Xuedong Yuan, Sancong Ying, and Huazhu Fu. Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image. *arXiv preprint arXiv:2307.04973*, 2023. 2

[8] Hao Ding, Changchang Sun, Hao Tang, Dawen Cai, and Yan Yan. Few-shot medical image segmentation with cycle-resemblance attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2488–2497, 2023. 3

[9] 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*, pages 272–284. Springer, 2022. 2

[10] 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. 4

[11] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 3

[12] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12341–12351, 2021. 2

[13] Bingjie Jiang, Tongwei Ren, and Jia Bei. Automatic scribble simulation for interactive image segmentation evaluation. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part I 22*, pages 596–608. Springer, 2016. 4

[14] 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. 2, 3

[15] Yiwen Li, Yunguan Fu, Iani JMB Gayo, Qianye Yang, Zhe Min, Shaheer U Saeed, Wen Yan, Yipei Wang, J Alison Noble, Mark Emberton, et al. Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration. *Medical Image Analysis*, 90:102935, 2023. 3

[16] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 2

[17] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Iteratively trained interactive segmentation. *arXiv preprint arXiv:1805.04398*, 2018. 4

[18] 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. 3

[19] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 622–631, 2019. 3

[20] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 762–780. Springer, 2020. 3

[21] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[22] Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H Maier-Hein. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. *arXiv preprint arXiv:2304.05396*, 2023. 2

[23] Jun Seo, Young-Hyun Park, Sung Whan Yoon, and Jaekyun Moon. Task-adaptive feature transformer with semantic enrichment for few-shot segmentation. *arXiv preprint arXiv:2202.06498*, 2022. 3

[24] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 3

[25] Junde Wu, Shuang Yu, Wenting Chen, Kai Ma, Rao Fu, Hanruo Liu, Xiaoguang Di, and Yefeng Zheng. Leveraging undiagnosed data for glaucoma classification with teacher-student learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 731–740. Springer, 2020. 3

[26] Junde Wu, Huihui Fang, Yu Zhang, Yehui Yang, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. *arXiv preprint arXiv:2211.00611*, 2022. 2

[27] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 2