One-Prompt to Segment All Medical Images

Junde Wu
National University of Singapore
MBZUAI
University of Oxford

Min Xu
Carnegie Mellon University
MBZUAI

Abstract

*Large foundation models, known for their strong zero-shot generalization, have excelled in visual and language applications. However, applying them to medical image segmentation, a domain with diverse imaging types and target labels, remains an open challenge. Current approaches, such as adapting interactive segmentation models like Segment Anything Model (SAM), require user prompts for each sample during inference. Alternatively, transfer learning methods like few/one-shot models demand labeled samples, leading to high costs. This paper introduces a new paradigm toward the universal medical image segmentation, termed 'One-Prompt Segmentation.' One-Prompt Segmentation combines the strengths of one-shot and interactive methods. In the inference stage, with just **one prompted sample**, it can adeptly handle the unseen task in a single forward pass. We train One-Prompt Model on 64 open-source medical datasets, accompanied by the collection of over 3,000 clinician-labeled prompts. Tested on 14 previously unseen datasets, the One-Prompt Model showcases superior zero-shot segmentation capabilities, outperforming a wide range of related methods. The code and data is released as* https://github.com/KidsWithTokens/one-prompt.

1. Introduction

Large foundation models pre-trained on large-scale datasets, are transforming the landscape with powerful zero-shot capabilities [27, 54–56]. These foundational models showcase an impressive ability in adapting to the tasks not seen during training. A standout example is the Segment Anything Model (SAM) [27], which has gained great success for the zero-shot image segmentation. The strength of SAM lies in its interactive segmentation paradigm: the model segments the target following the user-given prompts, such as a point, a bounding box (BBox), or free text-like descriptions.

Medical image segmentation, as a unique component of the image segmentation, plays a vital role in real-world clinical practices, including disease diagnosis and image-guided surgery. Many efforts have been made on bring this interactive foundation model to the medical image segmentation through fine-tuning [13, 36, 58]. However, most of them still need to re-training the model for each new task, leading to actually a loss of zero-shot generalization. Additionally, in these interactive models, the users have to provide prompts for each image, which is time-consuming and inapplicable for building the automatic pipeline.

Another way towards the universal medical image segmentation is few/one-shot learning [8, 41, 42, 51]. In this setting, a pre-trained foundational model needs one or few **labeled** samples as the 'supportive examples', to grasp a new specific task. However, securing labels for new tasks is not always feasible. Furthermore, the success of these methods heavily depends on the number of supportive examples provided. For example, UniverSeg[8] achieves competitive performance with 64 supportive samples. But obtaining such amount of data can be challenging in real clinical practice.

In this paper, we introduce a new paradigm for the universal medical image segmentation, called *One-Prompt Medical Image Segmentation*. This method combines the strengths of both one-shot and interactive models to meet the real clinical requirements. Specifically, given an unseen task, the user only needs to provide *one prompted sample* to the trained model, then it can perform well at this new task without any retraining or fine-tuning, even for tasks significantly different from those encountered during training. An illustration is shown in Fig. 1.

The success of the One-Prompt Model is driven by three key factors. First, we propose novel One-Prompt Former modules as the model decoder. Such design helps to efficiently integrate the prompted template feature into the process of query image segmentation. Secondly, we gather a large-scale data collection comprising 78 open-source datasets covering diverse biomedical domains. Our model is trained on 64 datasets, with clinicians prompting a part of the data. Moreover, to further enhance the clinical utility,

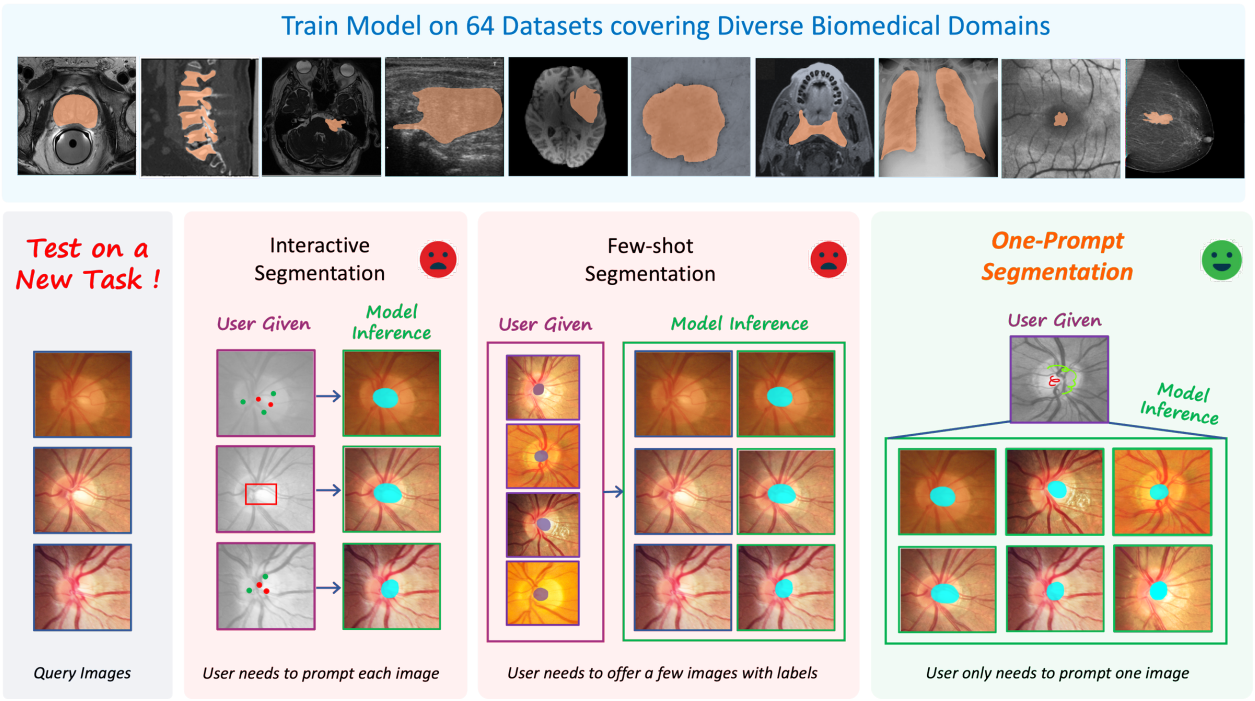

Figure 1. Medical segmentation involves a wide range of different organs, tissues and anatomies. One-Prompt Segmentation is a novel paradigm to building a foundation model that can generalize to unseen tasks. Given an unseen task, One-Prompt Model only needs the users to *prompt one image* to grasp the task, which is notably cost-effective comparing with interactive and few-shot segmentation.

we offer four different prompt types, which are *Click*, *BBox*, *Doodle*, and *SegLab*. The *Click* and *BBox* work the same as those in SAM [27]. *Doodle* allows the users to freely draw on the image, helpful for prompting irregular organs like the pancreas or lymph glands. *SegLab* allows the user to provide the segmentation mask as the prompt for representing more detailed tissues like vessels.

In sum, our contributions can be summarized as follows:

- We introduce novel One-Prompt Segmentation, which is a strong but low-cost paradigm for the universal medical image segmentation.
- We propose a model with unique One-Prompt Former to fuse the prompted template feature with the query feature in the multiple feature scales.
- We set four different prompt types for better prompting the special medical targets, thus to meet the various clinical practices .
- We gather a large-scale collection comprising 78 open-source datasets to train and test the model, and also annotated over 3000 of the samples with clinician-given prompts.

2. Method

Consider a set D containing all medical image segmentation tasks. Each task d consists of image-label pairs x^d, y^d. In conventional fully-supervised segmentation methods, a function $y^d = f_\theta^d(x^d)$ is typically learned to estimate a segmentation map y^d based on an input image x^d. However, this function f_θ^d is tailored exclusively for the specific task d. In the case of few-shot strategies, the target is to learn a universal function $y^d = f_\theta(x^d, S^d)$ performing on any task d guided by the task-specific support set $S^d = \{(x_j^d, y_j^d)\}_{j=1}^n$, comprising example image & label pairs available for task d.

In contrast, our One-Prompt Segmentation learns a more general function $y = f_\theta(x_j^d, k^d)$ performing on any task d, where $k^d = \{x_c^d, p_c^d\}$ comprising one fixed template image x_c^d and a paired prompt p_c^d available for task d. This prompt can be freely chosen by the users from different types like *Click*, *BBox*, *Doodle*, and *SegLabel*. This learning paradigm is more user-friendly in the clinical practice in that users only need to provide a single sample with prompts, and the model can adapt to any new task in one forward pass. This makes it easy for clinicians without a computer science background to use the system without the complexities of training or fine-tuning.

It is worth noting that interactive and one-shot models can be seen as special cases of One-prompt Segmentation. In specific, when $x_c^d = x^d$, it works as an interactive segmentation model, and when p_c^d is a segmentation label, it aligns with a one-shot model.

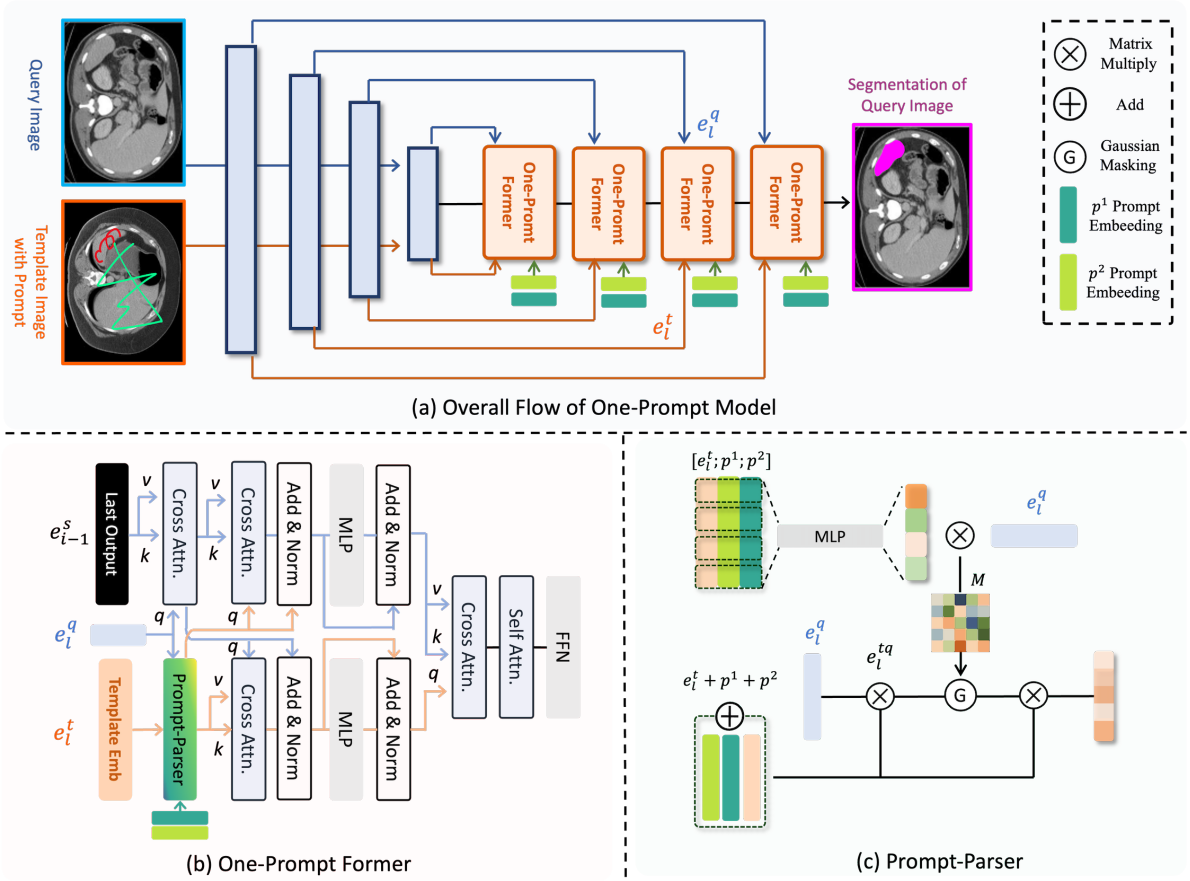

Figure 2. An illustration of One-Prompt Model, which starts from (a) an overview of the pipeline, and continues with zoomed-in diagrams of individual Models, including (b) One-Prompt Former, and (c) Prompt-Parser.

2.1. Prompts

Our model supports four types of prompts. Going beyond the usual *Click* and *BBox*, users can also use segmentation labels (*SegLab*) and freehand doodles (*Doodle*) as prompts. Each prompt type is best suited for its specific situations. For example, *Click* work well for obvious lesions like melanoma. *BBox* are effective for lesions with blurry boundaries but can be refined well by the boxes, such as the optic cup. *SegLab* are ideal for scenarios with detailed features, like complex vessels. *Doodle* are handy for organs with varies and unusual structures, like the pancreas and mandible. An illustration of it is shown in Section C. in the supplementary.

All prompts are represented using two embeddings, denoted as p^1 and p^2. For *Click* and *Doodle*, the two embeddings are used to denote the foreground and background. *BBox* use them to denote the left-top and right-bottom corner points. For these three kinds of prompts, we use positional encoding to compress the coordinate information of the prompts, then add them to the learnable embeddings. These embeddings learn themselves the concepts they to represent. *SegLab* is converted into an embedding using a pre-trained autoencoder. Two prompt embeddings share the same parameters in this case.

2.2. Model

The One-Prompt Model comprises an image encoder and a sequence of One-Prompt Former as the decoder, illustrated in Fig. 2 (a). The model takes three inputs: the query image x_q, the template image x_t, and the prompt of the template image p_t, and subsequently predicts the segmentation of the query denoted as y_q. The multi-scale features of the encoder and decoder are skip-connected.

Encoder The image encoder can be CNN based [24] or ViT based [12]. We show a CNN based encoder in the figure for the simple illustration. The query sample x_q and the template sample x_t will both go through the same encoder to get the feature f_q and f_t.

One-prompt Former We then decode the down-sampled query feature by incorporating the prompt embeddings, multi-scale template features, and multi-scale query fea-

tures together through a sequence of our proposed One-Prompt Former. All feature maps are pachlized, flattened and projected to the embedding $e \in \mathcal{R}^{N\times L}$ for further processing. The One-Prompt Former primarily consists of attention blocks, and its structure involves two parallel branches for processing query and template features, as depicted in Fig. 2 (b).

In each One-Prompt Former block (considering the i^{th} block), the Cross Attention [10] in the query branch initially takes the l^{th} level skip-connected query embedding e_l^q as the *query* and the last output embedding e_{i-1}^s serves as both the *key* and *value*, followed by another Cross Attention to incorporate the template feature symmetrically. Simultaneously, the template branch employs a proposed Prompt-Parser to integrate the prompts p with e_l^q and e_l^t, followed by another Cross Attention to integrate the query feature. In the end, a Cross Attention integrates the two branches by transferring the prompted template segmentation to the query domain. Then a self-attention followed by Feedforward Neural Network (FNN) are employed to project the embedding to the desired length.

Prompt-Parser In the template branch, we propose a simple Prompt-Parser to mix the prompt, query and template feature in an effective way. We show an illustration of the Prompt-Parser in Fig. 2 (c). The high-level idea is to produce an adaptive attentive mask M to activate the prompted target on the query-template-integrated embedding e_l^{tq}:

$$e_l^{tq} = e_l^t(p^1 + p^2 + e_l^q). \tag{1}$$

We then divide Prompt-Parser to a Prompting Step and a Masking Step. In the Prompting Step, we build a mask M adaptive to the different feature scale by mixing prompts with the given e_l^t and e_l^q. Specifically, we fist apply a MLP layer on the stacked embedding $[f^t; p^1; p^2] \in \mathcal{R}^{3N\times L}$. An MLP layer with weight $w \in \mathcal{R}^{N\times 3N}$ is applied along N to mix three different embeddings and reduce its dimension back to N. e_l^q is then matrix multiply on it to transfer its activation to the query domain. The process can be formalized as:

$$M = w[e_l^t; p^1; p^2](e_l^q)^T. \tag{2}$$

Then in the Masking Step, we apply $M \in \mathcal{R}^{N\times N}$ to e_l^{tq} through a proposed *Gaussian Masking* operation:

$$e^G = Max(e_l^{tq} * k_G[Conv(M)], e_l^{tq}), \tag{3}$$

where k_G is Gaussian kernel, * denotes general convolution operation. *Gaussian Masking* first projects M to a 2-channel feature map by convolution layer. Then we generate k_G by taking two channels as mean and variance respectively. k_G then multiply with e_l^{tq} in a pixel-wise manner to enlarge the prompted space but with uncertainty. Finally, we select the maximum value between the original feature and the smoothed one, preserving the highest activation and eliminating low-confidence uncertainty regions. The output is obtained by finally multiplying with e_l^t.

2.3. Training and Loss

We divided our One-Prompt dataset into 64 datasets for model training and 14 datasets for testing. Each training dataset is further split into three parts: a prompted template split, a training split, and a validation split. Similarly, each test dataset has a test split and a prompted template split. Human users prompt each sample in the template split, for both training and testing. Model training is performed on the template and training splits across all datasets. In each iteration, we randomly pick one prompted template from the template set of the same dataset with the query image. Training is collectively conducted across the 64 datasets. Our final loss is a simple sum of cross-entropy loss and dice loss.

During the inference stage, we randomly choose a prompted template from the template split and run the model over the test/validation split for the evaluation. Unless otherwise specified, we run the model 50 times and ensemble the predictions to mitigate variance.

3. One-Prompt Data

3.1. Data Source

In order to construct a foundation model with high generalization on the unseen tasks, we train our model on large-scale and diverse medical images consisted by online open-access datasets. Our data source is constructed from 78 datasets encompassing diverse medical domains and imaging modalities. The dataset covers a wide array of organs, such as lungs [48–50], eyes [16, 23, 39, 40], brain [5, 17, 22, 28, 29], and abdominal [6, 20, 25, 26, 30–33, 35, 38, 44, 50]. A detailed list of One-Prompt datasets is released with our code.

3.2. Prompt Annotation

A team of clinicians prompt over 3000 samples across over all collected dataset. These samples are meticulously selected by experienced annotators to ensure diversity and comprehensiveness. The clinicians involved in prompting come from diverse backgrounds, including cardiologists, dermatologists, gastroenterologists, neurologists, oncologists, pulmonologists, rheumatologists, endocrinologists, and ophthalmologists. They are encouraged to choose datasets aligned with their expertise during the prompting process.

In this process, all four prompt types are available for each sample, and clinicians are encouraged to use the most convenient prompt tool for the given targets. The clinicians employ a browser-based interactive segmentation tool to

prompt images. Upon prompting, ground-truth masks immediately appear on the images based on the given prompts. Clinicians have the flexibility to refine their prompts, but adjustments are suggested only if they feel their initial prompt was incorrect. Our prompt-based segmentation operates in real-time directly within a browser. Notably, we do not set strict constraints on prompt quality. Clinicians are encouraged to prompt images in their most natural way, with a suggested time limit of no more than 5 seconds for each image. The prompting details on the test set can be found in Section 4.1.

4. Experiments

In this section, we present task generalization experiments with One-Prompt Model. We divide out 14 tasks in our available datasets, which differ significantly in term of image modalities and target structures, as our held-out test set. This set comprises 8 MICCAI2023 Challenge tasks, encompassing various anatomies including kidney tumor [21], liver tumor [43], breast cancer [1], nasopharynx cancer [4], vestibular schwannoma [2], mediastinal lymph node [45], cerebral artery [11], and inferior alveolar nerve [7]. The other 6 tasks including the segmentation of white blood cell [60], optic cup [16], mandible [3], coronary artery [52], pancreas [46], and retinal blood vessel [23]. We assess the model performance on each test dataset using a specific prompt type, informed by the observation that users tend to favor specific prompts for particular tasks. We provide our implementation and data processing details in the appendix.

4.1. Human-User Prompted Evaluation

For the evaluation, we involved human users to simulate real-world interactions for prompt-based segmentation. We assigned 15 users to prompt about 10% of the test images. The users comprised 5 regular individuals with a clear understanding of the task but no clinical background, 7 junior clinicians, and 3 senior clinicians. This aims to simulate real-world prompting scenarios such as clinical education or semi-automatic annotation.

4.2. One-Prompt Transfer Capability

Task We first validate the One-Prompt transfer capability by comparing it with few/one-shot learning baseline using varies prompts. Our main objective is to assess the generalization of One-Prompt Model in solving unseen tasks. We compare with various few-shot models: PANet[51], ALPNet[41], SENet[47], UniverSeg[8], all provide with the same template to run. Few-shot methods are all given only one template in the testing for the fair comparison. Additionally, we compare with one-shot models: DAT[59], ProbOne [15], HyperSegNas[42], and LT-Net [53]. All these models are trained on the same dataset as ours, and are all given segmentation labels as the 'prompt' as they could not accept sparse prompts.

Data In this comparison, we conduct tests on the held-out test set with 14 different tasks. Among them, KiTS23[21], ATLAS23[43], TDSC[1], and WBC[60] datasets using the one *Click* prompt. For the SegRap [4], CrossM23 [2], and REFUGE [16] datasets, we employ the *BBox* Prompt. The *Doodle* prompt is applied to the Pendal [3], Pancreas-CT [46], LNQ23 [45], and CAS23 [11] datasets, while the *SegLab* prompt is used for CadVidSet [52], STAR [23], and ToothFairy [7] datasets

Results Fig. 3 illustrates the average Dice score per task for each method, and Fig. 5 provides the comparison on visualized results. It is worth noting that the compared few/one-shot models all necessitate the segmentation label as the 'prompt.' Therefore, they have a comparative advantage in contrast to our model. Despite this, our model consistently outperforms the competitors by significant margins, showcasing its robust generation ability across various tasks. In a fair comparison where all methods are provided segmentation labels (Fig. 3 SegLab), our model demonstrates more substantial leads, which averagely outperforms the second 11.2%, which is the most among all the prompt settings.

4.3. Interactive Segmentation Capability

Task

Interactive segmentation models achieve zero-shot generalization by prompting each of the test sample. When we offer the One-Prompt Model with the same query image and prompted template image, the model degrades to a standard interactive segmentation model. We compare this setting with other interactive segmentation models, including vanilla SAM [27], SAM-U [14], VMN [61], iSegFormer[34], MedSAM [37], MSA [58], and SAM-Med2D [13]. Except vanilla SAM, all models are trained on the same dataset as ours. Since most of these models only accept *Click* and *BBox* prompts, *Doodle* and *SegLab* prompt settings are not included in this comparison. Since all these models need the prompt on each input image, we simulate the *oracle* prompts (details in Section 4.5: Effect of prompt quality & types in the inference) over the images if needed. It is worth noting that we did not re-train One-Prompt Model on the simulated prompts: we use the same trained One-Prompt Model as that in the last section, but only offered the simulated prompts in testing for the possible of comparison.

Data We conduct the comparison on all 14 held-out test datasets.

Results. We present a quantitative comparison with the interactive segmentation methods in Fig. 4. We can see in the figure that One-Prompt Model outperforms all other interactive competitors by a significant margin. These results demonstrate that One-Prompt Model could perform as well

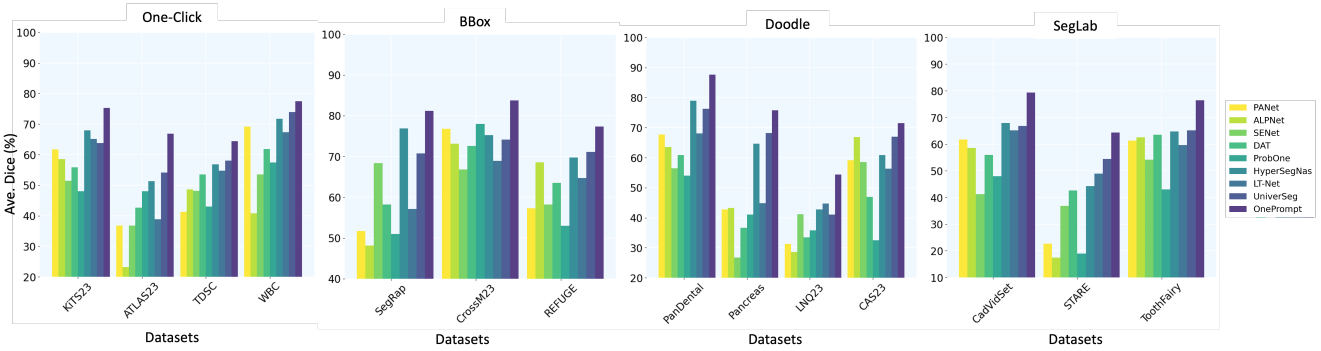

Figure 3. One-Prompt Model v.s. Few/One-shot Models on 14 held-out test datasets with 4 different prompts.

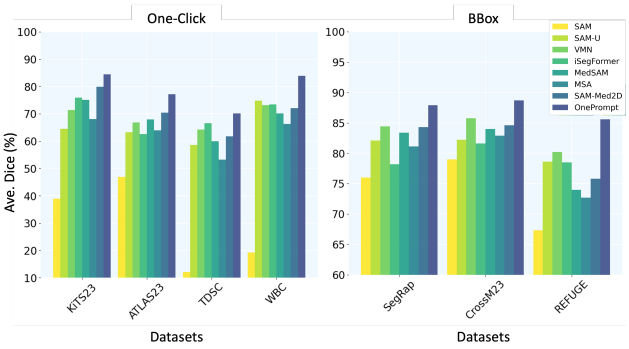

Figure 4. One-Prompt Model v.s. Interactive Segmentation Models on 7 held-out datasets with *One-Click* and *BBox* prompts.

when the query image itself is prompted, despite not being intentionally trained for this specific setting. By training under our more challenging setting, the One-Prompt Model demonstrates superior capability compared to interactive models.

4.4. Zero-shot Capability

Task Our model can automatically segment all salient targets following the similar 'segment everything' setting in SAM. In this setup, we prompt the template image with a regular grid of foreground points, generating an average of approximately 50 masks per image. We compare our model under this setting with conventional fully-supervised models that are not promptable [9, 12, 18, 19, 24, 57], and also SAM-based methods [13, 36, 58] under 'segment everything' setting.

Data We use 11 unseen datasets in the held-out test set to verify the zero-shot transfer ability of the models. Detailed datasets are shown in Table. 1

Results We present a quantitative comparison of zero-shot segmentation results in Table 1. It shows the challenge faced by fully-supervised segmentation methods in generalizing to unseen tasks, as they may struggle to understand the task, such as 'what to segment,' without the human interaction. When compared to SAM-based models under the 'segment everything' setting, the One-Prompt Model consistently outperforms them across all tasks, achieving the highest average performance of 64.0%, which surpasses the second-highest by a substantial 10.7%. It again highlights the value of setting a challenging learning task with a comparable model for enhancing the generalization.

4.5. Ablation Study and Analysis

Ablation on Prompt-Parser In the design of Prompt-Parser, we experimented with various combinations of strategies in both the Masking and Prompting steps. The comparative results are shown in Fig. 6. For the Prompting step, we explored simply adding or concatenating three embeddings and then projecting them to the desired length using a MLP. In the Masking step, we tested different approaches, such as directly adding the mask to the feature, binary thresholding the mask (setting negatives to zero and positives to one) then do element-wise multiplication with the feature (*Binary Masking*), or normalizing the mask and then element-wise multiplying it with the feature (*Norm Masking*). We can see the combination of the proposed Stack MLP + Gaussian Masking achieves the highest score on the held-out test dataset.

Variance of offering different Templates in the inference To assess the variance of using different templates during inference, we conducted 100 repetitions using different random templates across 8 test tasks. The results are shown in Fig. 7. We can see given the same type of prompts for different tasks, the larger variances are shown in tasks with more diverse and uncertain target structures, such as *Optic-Cup* (REFUGE) and *Pancreas* segmentation. This is because the template may significantly differ from the query samples in these tasks. The variance also varies depending on the type of prompts, such as the notably smaller variance is observed when using fine-grained *SegLab* prompts. Overall, we observed variance consistently staying below 13%. This inherent stability of the model suggests a robust

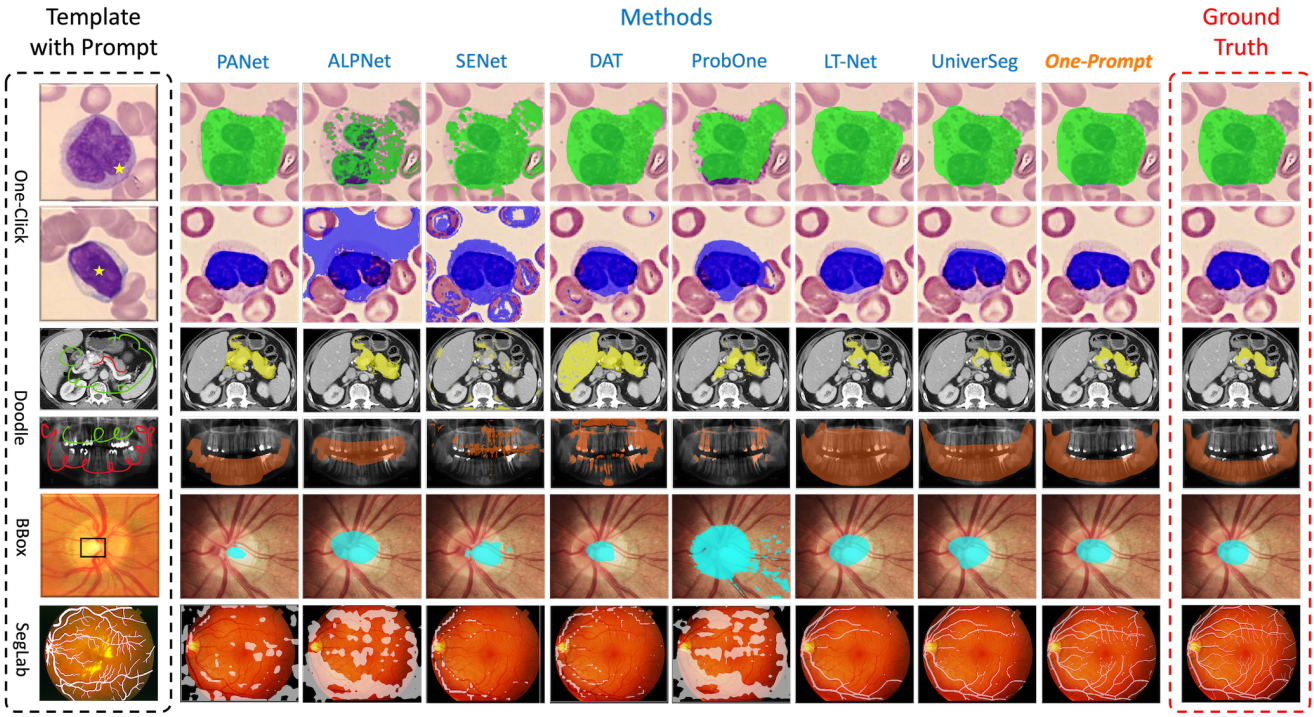

Figure 5. Visualized comparison of One-Prompt Model and few/zero-shot models. One-Prompt Model is given templates with prompts for the prediction. Few/zero-shot models are given templates with segmentation labels for the prediction.

Table 1. The zero-shot comparison between One-Prompt Model, full-supervised models and SAM-based models under 'segment everything' setting. Evaluated on 11 unseen tasks by Dice Score (%).

Methods	KiTS	ATLAS	WBC	SegRap	CrossM	REFUGE	Pendal	LQN	CAS	CadVidSet	ToothFairy	Ave
TransUNet	38.2	34.5	49.1	25.5	37.7	36.3	31.2	23.3	24.5	31.6	37.9	33.6
Swin-UNetr	37.2	26.5	32.1	25.6	29.7	28.9	31.4	17.2	20.5	22.6	32.1	28.5
nnUNet	39.8	30.3	40.4	26.8	35.0	34.9	42.9	18.9	37.4	41.8	35.3	34.9
MedSegDiff	40.1	30.5	42.9	34.7	37.7	31.9	42.6	21.1	38.3	34.7	33.5	35.3
MSA	54.6	48.9	55.9	47.3	51.7	49.2	54.2	41.0	48.9	53.5	47.6	50.3
MedSAM	62.4	53.1	67.8	52.3	59.3	54.5	58.7	42.5	41.5	45.7	56.2	53.9
SAM-Med2D	56.3	51.4	52.6	43.5	47.2	52.0	50.8	47.4	44.3	49.0	55.1	50.0
One-Prompt	67.3	63.8	72.5	62.2	65.8	58.4	72.6	49.5	64.5	66.3	61.4	64.0

zero-shot generalization capability.

Effect of prompt quality & types in the inference To verify the effect of prompt quality and types in the inference, we categorize five different levels of the prompt quality, from the lowest to the highest quality, respectively denoted as: *Low*, *Medium*, *High*, *Oracle*, and *Human*, on each of the prompt type. The detailed prompt simulation process is provided in the supplementary. We provide the model the same template with different prompt qualities each time in the comparison. The model is tested under One-Prompt setting on REFUGE and WBC datasets.

In Fig. 8, we observe a gradual improvement in model performance as prompt quality increases, highlighting the significant impact of prompt quality on the final model performance. The choice of prompt types also has a notable effect on the results. For both tasks, fine-grained *SegLab* exhibit the highest performance. *BBox* and *Doodle* demonstrate comparable performance after convergence and generally outperform the quicker and simpler *Click* prompt. This underscores a trade-off between user prompting and model performance: achieving better performance typically requires more detailed and high-quality prompts.

Model Efficiency In Table 2, we compare the efficiency of our model with one/few-shot learning models across 14 test tasks. Additionally, we train 14 TransUNet on 14 test datasets to establish an upper-bound for the performance. Unlike current one/few-shot models that require the fully-labeled images, One-Prompt only needs the users to simply prompt the image, significantly reducing the user-cost time. On average, it takes a user 27.47 seconds to anno-

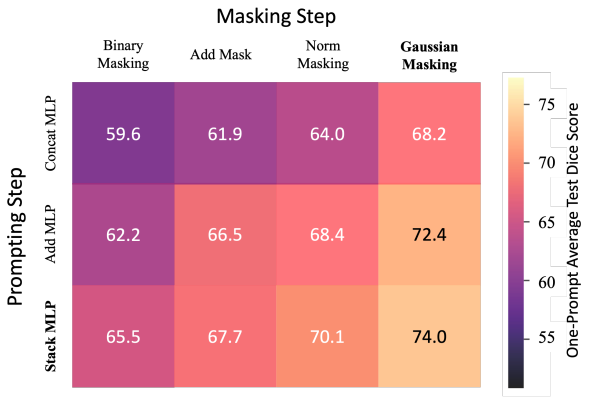

Figure 6. Ablation study on Prompt-Parser. We cross-validate the combination of different methods in Masking and Prompting steps, and show the average dice score under one-prompt segmentation setting on the held-out test set.

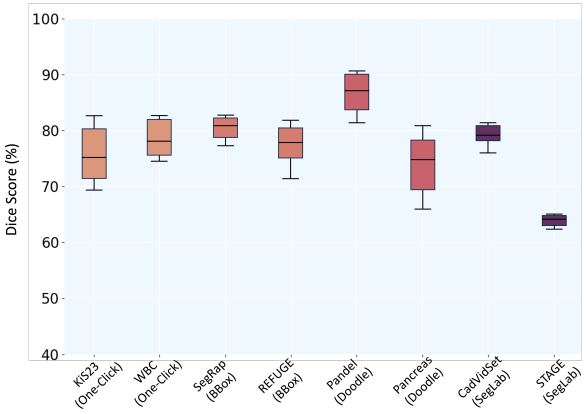

Figure 7. Variance of offering different templates to One-Prompt Model in the inference. Validated on 8 held-out test task giving different prompts.

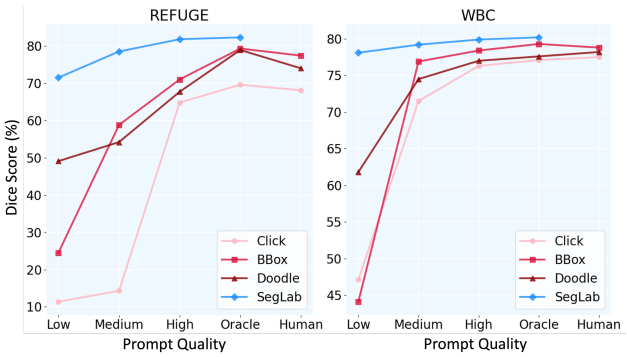

Figure 8. The variance of model performance given different prompts with different qualities on REFUGE and WBC test datasets.

tate one image across 14 datasets, while prompting for one image only takes an average of 2.28 seconds. Moreover, the prompting process requires the users much less clinical background and domain knowledge, making it more practical. The One-Prompt Model also exhibits superior scale-up capability, showing a significant improvement of about 10% compared to smaller models and only a 3.23% decrease compared to the TransUNet upper-bound. Comparing with the fully-supervised upper-bound, One-Prompt only needs to train one time for all downstream tasks, which saves significant parameters, training run time, and user-cost time for the annotation.

Table 2. **Model efficiency comparison with few/one-shot transfer learning models.** To establish an upper bound for performance, we individually train 14 task-specific TransUNet models for 14 held-out datasets. The run-time is its cumulative training time. The user-cost time is denoted as ∞ since the user must annotate all training samples in using.

Models	params (M)	Run Time (ms)	User-Cost Time (s)	Dice
ALPNet	14.7	240	27.47	52.96
PANet	43.0	528	27.47	50.11
HyperSegNas	1321.0	2154	27.47	63.86
UniverSeg	1.2	142	27.47	64.66
OnePrompt	192.0	741	2.28	73.98
TransUNet (sup.)	14×10^3	$14 \times 5.7 \cdot 10^7$	∞	77.21

5. Conclusion

In this paper, we introduce "One-Prompt Medical Image Segmentation", a new paradigm for building foundation model to handle diverse medical segmentation tasks. The model competed many related methods with just one prompted sample. With user-friendly prompt options for clinicians and remarkable results, the model holds significant promise for practical applications in clinical settings.

6. Acknowledgements

We sincerely apologize to **Jiayuan Zhu** and **Yueming Jin** for their absence from the authorship list, despite their **significant contributions** to this project. Unfortunately, due to missing the CVPR registration deadline, we were unable to include them. We deeply appreciate their dedication, expertise, and insights, which have been instrumental in bringing this work to fruition. Additionally, our gratitude extends to MBZUAI and Junde Wu for their generous support in funding this project.

References

[1] Tumor detection, segmentation and classification challenge on automated 3d breast ultrasound (abus) 2023. *https://tdsc-abus2023.grand-challenge.org*, 20203. 5

[2] Cross-modality domain adaptation for medical image segmentation. *https://crossmoda-challenge.ml*, 2023. 5

[3] Amir Hossein Abdi, Shohreh Kasaei, and Mojdeh Mehdizadeh. Automatic segmentation of mandible in panoramic x-ray. *Journal of Medical Imaging*, 2(4):044003–044003, 2015. 5

[4] Mehdi Astaraki, Simone Bendazzoli, and Iuliana Toma-Dasu. Fully automatic segmentation of gross target volume and organs-at-risk for radiotherapy planning of nasopharyngeal carcinoma. *arXiv preprint arXiv:2310.02972*, 2023. 5

[5] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 4

[6] Nicholas Bloch, Anant Madabhushi, Henkjan Huisman, John Freymann, Justin Kirby, Michael Grauer, Andinet Enquobahrie, Carl Jaffe, Larry Clarke, and Keyvan Farahani. Nci-isbi 2013 challenge: automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370(6):5, 2015. 4

[7] Federico Bolelli. Tooth fairy: A cone-beam computed tomography segmentation challenge. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2023*, 2023. 5

[8] 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. 1, 5

[9] 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. 6

[10] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 4

[11] Huijun Chen. Cerebral artery segmentation challenge. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2023.*, 2023. 5

[12] 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. 3, 6

[13] 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. 1, 5, 6

[14] 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. 5

[15] Yuhang Ding, Xin Yu, and Yi Yang. Modeling the probabilistic distribution of unlabeled data for one-shot medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1246–1254, 2021. 5

[16] Huihui Fang, Fei Li, Huazhu Fu, Xu Sun, Xingxing Cao, Jaemin Son, Shuang Yu, Menglu Zhang, Chenglang Yuan, Cheng Bian, et al. Refuge2 challenge: Treasure for multi-domain learning in glaucoma assessment. *arXiv preprint arXiv:2202.08994*, 2022. 4, 5

[17] Randy L Gollub, Jody M Shoemaker, Margaret D King, Tonya White, Stefan Ehrlich, Scott R Sponheim, Vincent P Clark, Jessica A Turner, Bryon A Mueller, Vince Magnotta, et al. The mcic collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics*, 11:367–388, 2013. 4

[18] 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. 6

[19] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022. 6

[20] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021. 4

[21] Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, Yoel Shoshan, Flora Gilboa-Solomon, Yasmeen George, Xi Yang, Jianpeng Zhang, Jing Zhang, Yong Xia, Mengran Wu, Zhiyang Liu, Ed Walczak, Sean McSweeney, Ranveer Vasdev, Chris Hornung, Rafat Solaiman, Jamee Schoephoerster, Bailey Abernathy, David Wu, Safa Abdulkadir, Ben Byun, Justice Spriggs, Griffin Struyk, Alexandra Austin, Ben Simpson, Michael Hagstrom, Sierra Virnig, John French, Nitin Venkatesh, Sarah Chan, Keenan Moore, Anna Jacobsen, Susan Austin, Mark Austin, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct, 2023. 5

[22] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1): 762, 2022. 4

[23] AD Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise

threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3):203–210, 2000. 4, 5

[24] 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. 3, 6

[25] 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. 4

[26] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 4

[27] 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. 1, 2, 5

[28] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11): 2556–2568, 2019. 4

[29] Maria Kuklisova-Murgasova, Paul Aljabar, Latha Srinivasan, Serena J Counsell, Valentina Doria, Ahmed Serag, Ioannis S Gousias, James P Boardman, Mary A Rutherford, A David Edwards, et al. A dynamic 4d probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763, 2011. 4

[30] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: Segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2020. 4

[31] 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*, page 12, 2015.

[32] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in biology and medicine*, 60:8–31, 2015.

[33] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014. 4

[34] Qin Liu, Zhenlin Xu, Yining Jiao, and Marc Niethammer. isegformer: interactive segmentation via transformers with application to 3d knee mr images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 464–474. Springer, 2022. 5

[35] 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*, 1(2):13, 2021. 4

[36] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 1, 6

[37] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 5

[38] 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*, 44(10):6695–6714, 2021. 4

[39] Yuhui Ma, Huaying Hao, Jianyang Xie, Huazhu Fu, Jiong Zhang, Jianlong Yang, Zhen Wang, Jiang Liu, Yalin Zheng, and Yitian Zhao. Rose: a retinal oct-angiography vessel segmentation dataset and new model. *IEEE transactions on medical imaging*, 40(3):928–939, 2020. 4

[40] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020. 4

[41] 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. 1, 5

[42] Cheng Peng, Andriy Myronenko, Ali Hatamizadeh, Vishwesh Nath, Md Mahfuzur Rahman Siddiquee, Yufan He, Daguang Xu, Rama Chellappa, and Dong Yang. Hypersegnas: Bridging one-shot neural architecture search with 3d medical image segmentation using hypernet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20741–20751, 2022. 1, 5

[43] Félix Quinton, Romain Popoff, Benoît Presles, Sarah Leclerc, Fabrice Meriaudeau, Guillaume Nodari, Olivier Lopez, Julie Pellegrinelli, Olivier Chevallier, Dominique Ginhac, et al. A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma. *Data*, 8(5):79, 2023. 5

[44] Perry Radau, Yingli Lu, Kim Connelly, Gideon Paul, Alexander J Dick, and Graham A Wright. Evaluation framework for algorithms segmenting short axis cardiac mri. *The MIDAS Journal*, 2009. 4

[45] Erik Ziegler Ron Kikinis, Steve Pieper. Mediastinal lymph node quantification (lnq): Segmentation of heterogeneous ct data. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2023. 5

[46] Holger R Roth, Amal Farag, E Turkbey, Le Lu, Jiamin Liu, and Ronald M Summers. Data from pancreas-ct. the cancer imaging archive. *IEEE Transactions on Image Processing*, 2016. 5

[47] Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. 'squeeze & excite'guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587, 2020. 5

[48] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, SQ Truong, CD Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, AY Ng, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *MedRxiv*, 2021. 4

[49] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

[50] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 4

[51] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019. 1, 5

[52] Lu Wang, Dongxue Liang, Xiaolei Yin, Jing Qiu, Zhiyun Yang, Junhui Xing, Jianzeng Dong, and Zhaoyuan Ma. Coronary artery segmentation in angiographic videos utilizing spatial-temporal information. *BMC medical imaging*, 20: 1–10, 2020. 5

[53] Shuxin Wang, Shilei Cao, Dong Wei, Renzhen Wang, Kai Ma, Liansheng Wang, Deyu Meng, and Yefeng Zheng. Lt-net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9162–9171, 2020. 5

[54] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 1

[55] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023.

[56] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 1

[57] 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. 6

[58] 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. 1, 5, 6

[59] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8543–8553, 2019. 5

[60] Xin Zheng, Yong Wang, Guoyou Wang, and Jianguo Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018. 5

[61] Tianfei Zhou, Liulei Li, Gustav Bredell, Jianwu Li, Jan Unkelbach, and Ender Konukoglu. Volumetric memory network for interactive medical image segmentation. *Medical Image Analysis*, 83:102599, 2023. 5