SegPrompt: Using Segmentation Map as a Better Prompt to Finetune Deep Models for Kidney Stone Classification

Editors: Under Review for MIDL 2023

Abstract

Recently, deep learning has produced encouraging results for kidney stone classification using endoscope images. However, the shortage of annotated training data poses a severe problem to improving the performance and generalization ability of the trained model. It is thus crucial to fully exploit the limited data at hand. In this paper, we propose SegPrompt to alleviate the data shortage problems by exploiting segmentation maps from two aspects. First, SegPrompt integrates segmentation maps to facilitate classification training, so that the classification model is aware of the regions of interest. The proposed method allows the image and segmentation tokens to interact with each other to fully utilize the segmentation map information. Second, we use the segmentation maps as prompts to tune the pretrained deep model, resulting in much fewer trainable parameters than vanilla finetuning. We perform extensive experiments on the collected kidney stone dataset. The results show that SegPrompt can achieve an advantageous balance between the model fitting ability and the generalization ability, eventually leading to an effective model with limited training data.

Keywords: Kidney Stone Classification, Prompt Tuning, Parameter Efficient Finetuning

1. Introduction

Nephrolithiasis, commonly referred to as kidney stones, are crystal concretions formed in the kidney and leave the body through the urethra without discomfort. However, larger stones typically cause severe pain and require surgical intervention (Nojaba L, 2022). In the United States, it is estimated that about 5% of women and 12% of men will develop kidney stones at some point in their lifetime and the prevalence of kidney stone development has been rising in both genders (Coe et al., 2005). One study has shown that the relapse rate of kidney stones after 2, 5, 10, and 15 years was 11%, 20%, 31%, and 39%, respectively (Rule et al., 2014). Therefore, understanding the type of kidney stones will help determine the cause and may give physicians clues on reducing the risks of relapse through personalized treatments (Scales Jr et al., 2012). The standard way of classifying kidney stone types involves kidney stone fragment extraction and subsequent visual examination through microscopes. This approach typically takes 1-2 months to get the classification result back to the physician, even though the patients can be in a critical condition and suffer from great pain (Ochoa-Ruiz et al., 2022).

Recently, deep learning-based methods have been developed to perform efficient diagnosis using endoscope images, and these methods often directly fine-tune the whole model (Ochoa-Ruiz et al., 2022; Estrade et al., 2022). However, similar to most medically related tasks (Zhu et al., 2020), the limited training data makes it hard to obtain a robust deep model that could generalize to unseen cases with vanilla finetuning (Zang et al., 2022). In this paper, inspired by the recent progress on Visual Prompt Tuning (VPT) (Jia

© 2023 .

et al., 2022), we propose SegPrompt for kidney stone classification by taking segmentation maps as prompts to tune the pretrained model. On one hand, SegPrompt integrates the segmentation map into the training process to make the model aware of the regions of interest, which intuitively benefits the classification training process (Khan et al., 2019). The segmentation map is obtained with a pretrained Unet (Ronneberger et al., 2015). On the other hand, as a prompt tuning-based method, SegPrompt does not update the backbone model and thus has much fewer trainable parameters than finetuning. In this way, we can avoid the overfitting problem that often comes with small-scale training data. Moreover, our model allows the image and segmentation tokens to mutually interact with each other so that the model can make full use of the segmentation map.

We highlight our contributions as follows:

1. We present SegPrompt in this paper. SegPrompt regards segmentation maps as prompts to tune the pretrained deep model for kidney stone classification with limited training data.

2. SegPrompt incorporates the segmentation maps to facilitate the classification training process and only prompt-tunes a small part of the model, thus alleviating the overfitting problem and improving the classification performance.

3. We conduct thorough experiments on the collected dataset to validate the effectiveness of SegPrompt.

2. Related Work

2.1. Kidney Stone Classification

Both traditional machine learning and deep learning approaches have been used for kidney stone classification. Serrat *et al.* adopt random forest to classify kidney stone images with hand-crafted texture and color feature vectors (Serrat et al., 2017). Motivated by the encouraging results of deep medical image analysis (Ronneberger et al., 2015), Amado and Alejandro (Torrell Amado, 2018) exploit deep metric learning approaches such as Siamese Networks and Triplet Networks to learn the embedding for kidney stone images and use the k-nearest neighbor algorithm to classify testing images. Black *et al.* (Black et al., 2020) incorporate ex-vivo kidney stones images to finetune a pre-trained deep neural network and obtain reasonable results. Estrade *et al.* leverage transfer learning to classify mixed stones using surface and cross-section images of kidney stones (Estrade et al., 2022). In Manoj *et al.*'s work (Manoj et al., 2022), they present the visualization analysis of the well-trained kidney stone classifier with Grad-CAM. Finally, Ochoa-Ruiz *et al.* (Ochoa-Ruiz et al., 2022) use deep neural networks to classify kidney stones with in-vivo images from medical endoscopes. However, most of these methods finetune the whole model, potentially leading to an overfitting problem. In contrast, SegPrompt has fewer trainable parameters to enhance the generalization ability and incorporates the segmentation map to facilitate the training.

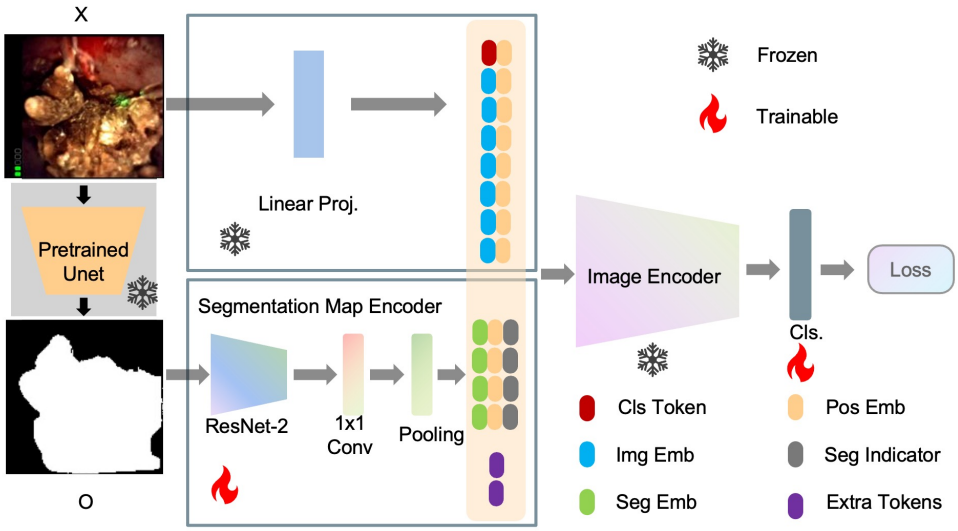

Figure 1: Block diagram of SegPrompt. We first extract the segmentation map with a pretrained Unet. The segmentation map is encoded into segmentation embeddings by the first two blocks of a pretrained ResNet18. We add the position embedding and segmentation indicator to the segmentation embedding to obtain segmentation tokens. Finally, we concatenate the image tokens, segmentation tokens, and extra learnable tokens and feed all tokens to the transformer backbone. We only update the segmentation map encoder and the last classifier during training.

2.2. Visual Prompt Tuning

Visual Prompt Tuning (VPT) was recently proposed to adjust a pretrained vision transformer for specific tasks with few trainable parameters and shows advantages in generalization ability over vanilla finetuning particularly with limited training data (Jia et al., 2022). VPT simply adds learnable tokens to the input of vision transformers (Jia et al., 2022). Visual Prompting pads the original images with learnable pixels (Bahng et al., 2022). NOAH performs a neural architecture search to learn optimal prompt structure (Zhang et al., 2022). Unified vision and language prompt tuning are proposed to jointly tune the VL models (Zang et al., 2022). S-Prompt is proposed to handle domain incremental learning with prompt tuning (Wang et al., 2022). The prompts used by these methods are simply trainable parameters. In contrast, SegPrompt learns to generate prompts based on the segmentation map, and achieve a better balance between fitting and generalization ability.

3. Methodology

In this section, we present the proposed segmentation map-based prompt tuning framework for kidney stone classification. Iy is designed toimprove the model's performance and

generalization ability by better exploiting the knowledge of segmentation maps with fewer trainable parameters.

3.1. Overview

Our method contains a frozen ViT backbone (Dosovitskiy et al., 2020), a segmentation map encoder, and a linear classifier. Similar to most medical image tasks, we also suffer from the scarcity of annotated kidney stone images, and it is expensive to collect more samples (Zhu et al., 2020). The shortage of training data makes it hard to obtain an effective model which could generalize to unseen cases. To alleviate the problem, on one hand, we involve the segmentation map in the classification training so that the model is aware of the regions of interest. On the other hand, we tokenize the segmentation map into prompts to finetune the backbone model and only update the segmentation map encoder and the last linear classifier. Consequently, the much fewer trainable parameters empower our model with better generalization abilities to avoid the overfitting problem (Jia et al., 2022). Moreover, since we perform self-attention on both image and segmentation tokens (Vaswani et al., 2017), our method allows the model to exploit the knowledge of the segmentation map more comprehensively and flexibly. We show the block diagram of our method in Fig. (1).

3.2. Tokenize the Segmentation Map

We first show how to convert the segmentation map into prompt tokens with the proposed segmentation map encoder h. The encoder h consists of the first two blocks of an ImageNet pre-trained ResNet18 followed by a projector, where the projector is composed of a 1x1 convolutional layer and an adaptive pooling layer (He et al., 2016). The convolutional layer of the projector is used to match the dimension of Resnet output to that of ViT while the pooling layer reduces the segmentation tokens to a desirable length. Moreover, the encoder h also contains learnable tokens P_s, Z_e, and r which will be introduced later. Given a training image X, we first obtain the binarized pixel-wise segmentation map $O \in \{-1, 1\}$ from a pretrained Unet (Ronneberger et al., 2015), where 1 denotes foreground regions with kidney stones, and -1 denotes background regions, the segmentation map embeddings $M = \{m^i\}_{i=1}^{l_m} \in R^{l_m \times d}$ can be obtained by flattening the output of segmentation map encoder as

$$M = flatten(h(O)), \tag{1}$$

where d is the dimension of the backbone model. Then, to convert the embedding M to tokens, we first add the learnable position embedding $P_s = \{p_s^i\}_{i=1}^{l_m} \in R^{l_m \times d}$ to retain the position information and then add a learnable indicator token $r \in R^d$ to enable the model to distinguish the segmentation tokens from the image tokens. Specifically, we obtain the segmentation tokens $Z_s = \{z_s^i\}_{i=1}^{l_m} \in R^{l_m \times d}$ by

$$z_s^i = m^i + p_s^i + r \;\; for \;\; i = 1, \ldots, l_m. \tag{2}$$

3.3. Prompt Tuning with Segmentation Tokens

To enable the model to better interact with and exploit the segmentation map, we propose to prompt-tune the model with the segmentation tokens (Jia et al., 2022). In particular, we

concatenate the classification token z_{cls}, image tokens $Z_x = \{z_x^i\}_{i=1}^{l} \in R^{l \times d}$, the segmentation tokens Z_s, and some extra learnable tokens $Z_e = \{z_e^i\}_{i=1}^{l_e} \in R^{l_e \times d}$ as input Z to the transformer backbone

$$Z = [z_{cls}, z_x^0, \ldots, z_x^l, z_s^0, \ldots, z_x^{l_s}, z_e^0, \ldots, z_e^{l_e}] \tag{3}$$

The extra learnable tokens make the pretrained model better adapt to kidney stone classification. The classification token z_{cls} and image tokens Z_x are frozen during training (Dosovitskiy et al., 2020). We perform multi-head self-attention on the input tokens Z and take the z_{cls} from the last layer as the output, which will be further processed by the classifier to get the final prediction. We adopt cross-entropy loss to train the model. During training, we keep the transformer backbone frozen and only update the segmentation map encoder (including the corresponding position embedding P_s and the indicator token r), the last classifier, and the introduced extra tokens Z_e.

We discuss several important features of SegPrompt as follows. Existing tuning methods strive to balance the generalization and fitting abilities to adapt a pretrained model for small-scale tasks. For example, finetuning suffers from the overfitting problem with over-abundant learnable parameters, while VPT may underfit the target dataset which deviates significantly from the pretrained dataset (Wortsman et al., 2022). Moreover, it is not trivial to integrate additional knowledge (e.g., segmentation map) into the training process for these methods. We empirically find that SegPrompt leads to a powerful model without severe overfitting and can effectively utilize extra knowledge. Specifically, compared with vanilla fine-tuning, the much fewer trainable parameters of SegPrompt significantly alleviate the overfitting problem. Compared with VPT, SegPrompt, equipped with the learnable segmentation map encoder, has a more powerful learning capacity and also does not severely distort the pretrained model because the number of newly introduced tokens (default set to 51) is much fewer than that of the original tokens (197 for ViT-B/16) (Dosovitskiy et al., 2020). To make use of additional knowledge, i.e., the segmentation map, SegPrompt allows image tokens and segmentation tokens to interact with and extract information from each other to improve classification performance. Last but not least, one can simply extend SegPrompt to integrate other kinds of knowledge, such as patient demographic information (Daneshjou et al.), device/scanner information (Ji et al., 2022), text descriptions of the symptoms (Qin et al., 2022), and medical record. We leave these as future directions.

4. Experiments

4.1. Dataset

We collect 1496 kidney stone images from 5 different videos. We filter out low-quality and background images and obtain 867 images from 3 videos with COM (calcium oxalate monohydrate) stones and 629 images from 2 videos with CAP(calcium phosphate) stones. We split the dataset into training and validation sets in a video-wise fashion to prevent any data leaks.

4.2. Implementation of Segmentation Model

The kidney stone regions are labeled for all images, and we train a Unet (Ronneberger et al., 2015) implemented by the MMsegmentation framework (Contributors, 2020) to perform

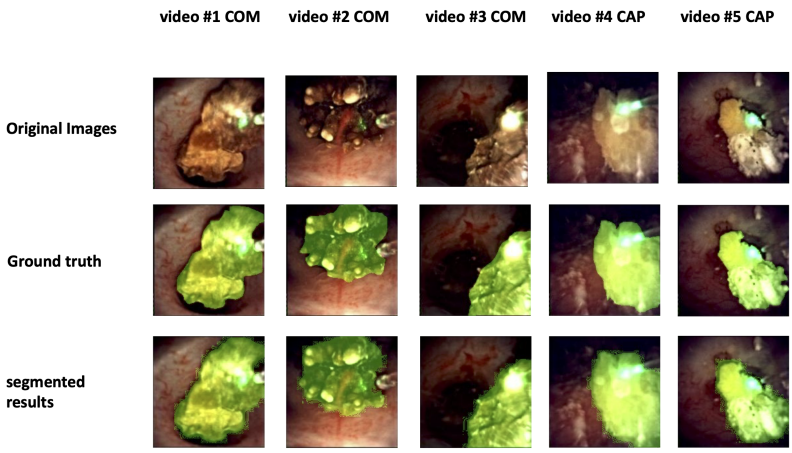

Figure 2: Illustration of segmentation results. Each column represents one image sample. The images come from the validation set of different folds.

the segmentation. Besides the training data from different folds, we also leverage some extra images without kidney stone labels to facilitate the segmentation model training. In total, there are eight additional kidney stone videos without classification labels, which provide 1860 additional images. Finally, we obtain segmentation models with the pixel-level accuracy averaged over different folds as 96.7% and the dice score as 92.93%. The outputs are binarized to obtain final segmentation maps. We visualize segmentation results from the validation set in Fig. (2).

4.3. Baseline Methods and Implementation Details

Three finetuning-based models are included as the baselines to validate the superiority of the proposed SegPrompt for the kidney stone classification task as FT (Finetuning), FT-crop (Finetuning-crop), FT-concat (Finetuning-concat). FT uses raw images without segmentation maps. FT-crop takes the cropped regions of interest as input. As for FT-concat, we channel-wisely concatenate the segmentation maps to the images as the input. The concatenated images then passed through a 1x1 convolution layer to reduce their channel size to 3 before feeding them into the ViT backbone. Moreover, we also compare our method with Visual Prompt Tuning (VPT) and VPT-Deep (Jia et al., 2022). VPT adds the learnable tokens to the input, while VPT-deep adds different tokens for each layer. Similar to VPT and VPT-deep, we also implement two variants of SegPrompt as SegPrompt and SegPrompt-Deep.

For all methods, the standard image preprocessing steps, such as resizing and normalization, have been applied to the training images. The batch size, number of epochs, and learning rate are 16, 20, and 0.001, respectively. We adopt an ImageNet pretrained ViT-B/16 as the backbone which contains 196 image tokens and 1 classification token (Dosovitskiy et al., 2020). As for the proposed SegPrompt (Deep), we set the number of segmentation tokens

Table 1: Kidney Stone Classification Results averaged over 6 Folds (corresponding to all possible combinations of videos). The best results are highlighted in bold. (%)

Methods	Accuracy	Precision	Recall	F1
FT	95.07 ± 3.2	94.21 ± 4.7	95.46 ± 2.7	93.73 ± 5.9
FT-crop	94.34 ± 4.0	94.14 ± 4.2	94.19 ± 3.9	92.96 ± 5.6
FT-concat	95.18 ± 2.5	94.91 ± 2.5	95.13 ± 2.3	94.53 ± 3.0
VPT	96.87 ± 1.1	96.60 ± 1.7	96.65 ± 1.3	96.07 ± 2.8
VPT-Deep	95.85 ± 0.8	95.49 ± 1.2	95.60 ± 1.2	95.13 ± 2.4
SegPrompt	$\mathbf{99.56 \pm 0.3}$	$\mathbf{99.45 \pm 0.4}$	$\mathbf{99.60 \pm 0.3}$	$\mathbf{99.45 \pm 0.5}$
SegPrompt-Deep	99.19 ± 0.3	99.06 ± 0.3	99.24 ± 0.3	99.26 ± 0.2

Table 2: Ablation Studies. (%)

Methods	Accuracy	Precision	Recall	F1
SegPrompt	99.56 ± 0.3	99.45 ± 0.4	99.60 ± 0.3	99.45 ± 0.5
SegPrompt w/o r	99.07 ± 0.6	98.90 ± 0.7	99.11 ± 0.5	98.87 ± 0.8
SegPrompt w/o z_e	99.38 ± 0.5	99.34 ± 0.5	99.36 ± 0.8	99.30 ± 0.3

to $l_s = 49$, and the extra learnable tokens to $l_e = 2$. The number of learnable tokens for VPT (Deep) is searched from $\{8, 16, 32, 51\}$. We perform 6-fold cross-validation, and the averaged accuracy, precision, recall, and F1 scores with standard deviation are reported.

4.4. Experimental Results

We present the experimental results in Table 1, and draw several interesting conclusions according to the results. First, FT outperforms FT-crop, suggesting that the cropped clean kidney stone images do not benefit the classification performance, which in turn suggests that the surrounding background regions contain critical information. Second, FT-concat utilizes the segmentation map without removing the background regions and it slightly outperforms FT for 0.8% in terms of F1 score. This shows that the classification performance could be boosted by properly exploiting the information of the segmentation map. Third, the overfitting problem is crucial in adapting the pre-trained models to the kidney stone classification with limited training data. VPT and VPT-deep outperform all FT-based methods with much fewer trainable parameters (Jia et al., 2022). In particular, VPT obtains a 1.54% improvement over FT-concat in terms of the F1 score and also surpasses its direct counterpart VPT-deep which has much more trainable parameters. Finally, the proposed SegPrompt makes better use of the segmentation map with few trainable parameters and achieves the best overall performance. For example, SegPrompt improves the F1 score from 96.07% to 99.45% compared with the second-best method VPT. We also note that SegPrompt-deep only slightly degrades the performance compared with SegPrompt, which shows the possibility of applying our method to handle large-scale tasks that require more learnable parameters to fit the training set.

Table 3: Performance with the different number of segmentation tokens. (%)

# l_s	Accuracy	Precision	Recall	F1
25	98.81 ± 1.0	98.75 ± 0.9	98.73 ± 1.0	98.69 ± 0.9
36	99.13 ± 0.6	99.16 ± 0.4	98.87 ± 1.2	98.78 ± 1.4
49	$\mathbf{99.56 \pm 0.3}$	$\mathbf{99.45 \pm 0.4}$	$\mathbf{99.60 \pm 0.3}$	$\mathbf{99.45 \pm 0.5}$
64	99.28 ± 0.2	99.32 ± 0.1	99.22 ± 0.3	99.18 ± 0.4
81	99.22 ± 0.6	99.14 ± 0.7	99.25 ± 0.6	99.13 ± 0.7

4.5. Ablation Studies

Extensive ablation studies have been performed to verify the effectiveness of different components of SegPrompts. All variants are developed based on SegPrompt instead of SegPrompt-Deep due to its superior performance and simplicity.

We first study the importance of indicator token r and the extra learnable tokens z_e. The indicator token r enables the model to distinguish the segmentation map tokens from other tokens, and the extra learnable tokens could make the pretrained model better adapt to our task. The two variants are denoted as SegPrompt w/o r and SegPrompt w/o z_e, respectively. The experimental results are shown in Table 2. According to the results, we find that the indicator token is essential for SegPrompt while the extra learnable tokens further slightly improve the performance.

We also conduct experiments to study the influence of different numbers of segmentation tokens l_s, and select $l_s \in \{25, 26, 49, 64, 81\}$. Based on the results shown in Table 3, the increasing number of segmentation tokens benefits the final performance, while over-large values may lead to a slight decrease. Considering the computational cost and the performance, we default set $l_s = 49$.

5. Conclusions and Future Work

In this paper, we present a novel segmentation map-based prompt tuning method for kidney stone classification with limited data, named SegPrompt. We first employ a well-trained Unet to extract the segmentation maps, which are further converted to segmentation tokens by a segmentation map encoder. Then SegPrompt takes the concatenation of image, segmentation, and some extra tokens as input to the transformer. During training, we only prompt-tune the segmentation map encoder and the linear classifier. SegPrompt can better exploit the knowledge of segmentation maps with few trainable parameters and significantly outperforms existing methods for kidney stone classification. The main limitation of our work is the small-scale training dataset, and we will collect more data to improve our model for more types of kidney stones. Moreover, we plan to extend our work to other medical tasks to further validate the effectiveness of SegPrompt (Daneshjou et al.; Qin et al., 2022).

References

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*,

2022.

Kristian M Black, Hei Law, Ali Aldoukhi, Jia Deng, and Khurshid R Ghani. Deep learning computer vision algorithm for detecting kidney stone composition. *BJU international*, 125(6):920–924, 2020.

Fredric L Coe, Andrew Evan, Elaine Worcester, et al. Kidney stone disease. *The Journal of clinical investigation*, 115(10):2598–2608, 2005.

MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto A Novoa, and James Zou. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Vincent Estrade, Michel Daudon, Emmanuel Richard, Jean-christophe Bernhard, Franck Bladou, Grégoire Robert, and Baudouin Denis de Senneville. Towards automatic recognition of pure and mixed stones using intra-operative endoscopic digital images. *BJU international*, 129(2):234–242, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

Muhammad Attique Khan, Tallha Akram, Muhammad Sharif, Tanzila Saba, Kashif Javed, Ikram Ullah Lali, Urcun John Tanik, and Amjad Rehman. Construction of saliency map and hybrid set of features for efficient segmentation and classification of skin lesion. *Microscopy research and technique*, 82(6):741–763, 2019.

B Manoj, Neethu Mohan, Sachin Kumar, et al. Automated detection of kidney stone using deep learning models. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pages 1–5. IEEE, 2022.

Guzman N Nojaba L. Nephrolithiasis, 2022. URL https://www.ncbi.nlm.nih.gov/books/NBK559227/.

Gilberto Ochoa-Ruiz, Vincent Estrade, Francisco Lopez, Daniel Flores-Araiza, Jonathan El Beze, Dinh-Hoan Trinh, Miguel Gonzalez-Mendoza, Pascal Eschwège, Jacques Hubert, and Christian Daul. On the in vivo recognition of kidney stones using machine learning. *arXiv preprint arXiv:2201.08865*, 2022.

Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. Medical image understanding with pretrained vision language models: A comprehensive study. *arXiv preprint arXiv:2209.15517*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Andrew D Rule, John C Lieske, Xujian Li, L Joseph Melton, Amy E Krambeck, and Eric J Bergstralh. The roks nomogram for predicting a second symptomatic stone episode. *Journal of the American Society of Nephrology*, 25(12):2878–2886, 2014.

Charles D Scales Jr, Alexandria C Smith, Janet M Hanley, Christopher S Saigal, Urologic Diseases in America Project, et al. Prevalence of kidney stones in the united states. *European urology*, 62(1):160–165, 2012.

Joan Serrat, Felipe Lumbreras, Francisco Blanco, Manuel Valiente, and Montserrat López-Mesas. mystone: A system for automatic kidney stone classification. *Expert Systems with Applications*, 89:41–51, 2017.

Alejandro Torrell Amado. Metric learning for kidney stone classification. 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *arXiv preprint arXiv:2207.12819*, 2022.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022.

Wei Zhu, Haofu Liao, Wenbin Li, Weijian Li, and Jiebo Luo. Alleviating the incompatibility between cross entropy loss and episode training for few-shot skin disease classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 330–339. Springer, 2020.