# VLSM-Adapter: Finetuning Vision-Language Segmentation Efficiently with Lightweight Blocks

Manish Dhakal[0000−0001−5019−2205], Rabin Adhikari[0000−0002−0101−5592], Safal Thapaliya[0000−0002−4463−6700], and Bishesh Khanal[0000−0002−2775−4748]

Nepal Applied Mathematics and Informatics Institute (NAAMII)
{manish.dhakal,rabin.adhikari,safal.thapaliya,bishesh.khanal}@naamii.org.np

**Abstract.** Foundation Vision-Language Models (VLMs) trained using large-scale open-domain images and text pairs have recently been adapted to develop Vision-Language Segmentation Models (VLSMs) that allow providing text prompts during inference to guide image segmentation. If robust and powerful VLSMs can be built for medical images, it could aid medical professionals in many clinical tasks where they must spend substantial time delineating the target structure of interest. VLSMs for medical images resort to fine-tuning base VLM or VLSM pretrained on open-domain natural image datasets due to fewer annotated medical image datasets; this fine-tuning is resource-consuming and expensive as it usually requires updating all or a significant fraction of the pretrained parameters. Recently, lightweight blocks called adapters have been proposed in VLMs that keep the pretrained model frozen and only train adapters during fine-tuning, substantially reducing the computing resources required. We introduce a novel adapter, *VLSM-Adapter*, that can fine-tune pretrained vision-language segmentation models using transformer encoders. Our experiments in widely used CLIP-based segmentation models show that with only 3 million trainable parameters, the VLSM-Adapter outperforms state-of-the-art and is comparable to the upper bound end-to-end fine-tuning. The source code is available at: https://github.com/naamiinepal/vlsm-adapter.

**Keywords:** Vision-Language Segmentation · Transfer Learning · Parameter Efficient Fine-tuning · Multimodal Adapters · Medical Imaging

## 1 Introduction

The early 2010s saw the initial success of Deep Learning in single-domain tasks such as image classification or language translation when deep neural networks could learn powerful representation using large-scale images [4,9] or texts [6]. As openly available large-scale annotated data lacked medical images, transfer learning was widely used where networks are initialized using weights obtained from pretraining in natural images such as ImageNet [4] and are further fine-tuned in domain-specific smaller datasets [28]. Recently introduced foundation Vision-Language Models (VLMs) can learn powerful joint representation from large-scale images-text pairs and can be adapted to a wide range of tasks, including dense prediction tasks to develop Vision-Language Segmentation Models (VLSMs), which allow providing text prompts during inference to guide image segmentation. VLSMs are attractive in the medical domain because robust and powerful VLSMs could aid medical professionals in many clinical tasks requiring tedious and time-consuming delineation of the target structure of interest.

VLSMs have separate vision and language encoders or joint vision-language encoders followed by a decoder or a mask-generating network that is trained end-to-end (E2E) [26], or separately using frozen encoder parameters obtained from VLM pretraining [18]. The most popular VLM, widely adapted to create different VLSMs, is Contrastive Language-Image Pretraining (CLIP) [21], which uses separate vision and language encoders. It learns joint vision-language representation by projecting images and texts into a shared embedding space through learnable parameters that bring semantically similar image-text pairs close while pushing dissimilar pairs further apart. Various VLSMs [18,26,27] have leveraged this multimodal semantic information captured by CLIP to train a segmentation model for an open-vocabulary segmentation task. Yu et al. [29] used a pretrained self-supervised mask proposal network and CLIP to realize the zero-shot referring image segmentation on the open domain without additional training.

Although open-domain VLMs show impressive zero-shot or few-shot performances in downstream tasks, adapting them to medical image segmentation requires further fine-tuning [1,20]. This fine-tuning usually requires updating all [11] or a significant fraction (usually last layers) of the pretrained parameters [16], which is expensive because VLMs are much larger than popular image-only models (a few to several hundred million parameters). Different methods have been proposed to efficiently fine-tune these foundation models, often called Parameter Efficient Fine-Tuning (PEFT) techniques. The two most popular PEFT techniques are LoRA [12] and Adapters [10] — both of them adjust the intermediate representations of the pretrained models often using lightweight networks parallel to the pretrained ones with only slight differences. Since adapters have been explored more in vision-language settings [7,23] compared to LoRA, we focus on adapters as a method for PEFT VLSMs for the scope of this paper.

Adapters are small networks with much fewer parameters that can be plugged into existing pretrained architectures, and then only adapters are trained while keeping pretrained weights frozen during fine-tuning. VL-Adapter [24] reused the pretrained VLMs for vision-text tasks like image captioning and visual questioning-answering. Although a few methods have been proposed for VLM-based classification tasks, no adapters are studied for E2E-trained VLSMs for further fine-tuning. Side-Adapter Network (SAN) [27] introduced ViT [5] as an adapter network, parallel to the CLIP's encoders, that generates segmentation masks for image-text inputs. This paper proposes learnable adapter networks to fine-tune already trained VLSMs, as *VLSM-Adapter*, which adapts the intermediate learned representations for domain-specific datasets while preserving the already learned weights from large-scale pretraining. We add learnable adapter modules to a variant of VLSM, CLIPSeg [18], introducing 3 million trainable parameters, which perform on par with the same model's E2E fine-tuning despite having almost 50 times fewer trainable parameters.

The main contributions of this paper are:

- We introduce novel adapter modules to efficiently fine-tune pretrained VLSMs to domain-specific smaller datasets using only a few learnable parameters.
- Our experiments and results on medical datasets with diverse modalities indicate that fine-tuning only the adapter modules for small datasets is better than E2E fine-tuning for VLSMs.
- We provide an ablation study on the positioning of adapter modules and show that introducing adapters deeper into the intermediate representations of the pretrained models results in better performance.

## 2 Adapters for CLIP-based VLSMs

### 2.1 Problem Statement

An encoder-decoder architecture-based pretrained model for vision-language segmentation model is frozen, while adapter modules with a much smaller number of parameters compared to the original frozen network are introduced to fine-tuning in a smaller training set comprising of the triplets: $D = \{(v_i, l_i, m_i)\}_{i=1}^{S}$. Here, $S$ is the number of training samples, $v_i$, $l_i$, and $m_i$ represent the image input, text prompt, and target mask of the $i^{th}$ data point, respectively. The input images are RGB images and targets are their corresponding binary masks, i.e., $v_i \in \mathbb{R}^{H \times W \times 3}$, and $m_i \in \{0,1\}^{H \times W}$, respectively.

### 2.2 Adapter Formulation

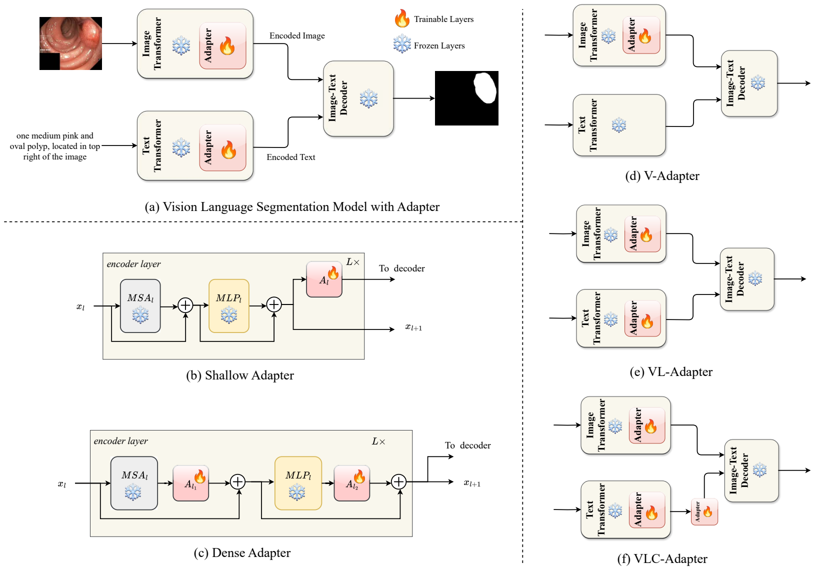
Adapter modules [10] are the non-linear projection blocks that adapt the representations of the pretrained models to a downstream task without changing their parameters, enabling the representations learned by the pretrained models to be used for other tasks. Eq. (1) represents the basic block of an adapter network.

$$f' = Adapter(f) = f + \sigma(\psi(f \cdot W_1) \cdot W_2) \tag{1}$$

Here, $f$ is the representation learned by the pretrained model, $f'$ is the adapted features, and $W_1$ and $W_2$ are learnable adapter weights. $\psi$ and $\sigma$ are non-linear activation functions, which, in most cases, are the same type. The adapter weights are initialized as $W_1 \in \mathbb{R}^{d \times d'}$, $W_2 \in \mathbb{R}^{d' \times d}$, where $d' \leq d$. The size of the input tensor must not change while exiting the adapters because they have to be used by the subsequent pretrained layers, i.e., $\{f, f'\} \in \mathbb{R}^{\cdots \times d}$.

### 2.3 Proposed VLSM-Adapter

As displayed in Fig. 1, we introduce adapters to the encoder segments while keeping the decoder static to VLSM-Adapter. The positional variation to introduce



(a) Vision Language Segmentation Model with Adapter

(b) Shallow Adapter

(c) Dense Adapter

(d) V-Adapter

(e) VL-Adapter

(f) VLC-Adapter

**Fig. 1. Overall architecture of the proposed VLSM-Adapter module.** $MSA_l$, $MLP_l$, and $A_{l_*}$ stand for multi-head self-attention block, multi-layer perceptron, and adapter layers, respectively, for the $l^{th}$ transformer layer. **(a)** Adapters are connected to each transformer block in the text and image encoders. **(b)** The shallow adapter adds learnable layers to each transformer block output. **(c)** The dense adapter employs the learnable layers before each residual addition in each transformer block. **(d,e,f)** Adapters have been configured with different positioning for text and image encoders.

transformer block adapters provides three incremental VLSM-Adapter variants. **(1) V-Adapter** has adapters in the image encoder layers. **(2) VL-Adapter** adds adapters for text encoder layers. **(3) VLC-Adapter** adds an extra layer to adapt text conditioning at the bottleneck layer. Since CLIPSeg [18] provides transformer encoders and a pretrained segmentation mask decoder, we have used it as a candidate model for our experiments to validate VLSM-Adapter. Two variants of VLSM-Adapter in CLIPSeg networks have been implemented for fine-tuning: CLIPSeg *Shallow Adapter (SA)* and CLIPSeg *Dense Adapter (DA)*.

**CLIPSeg Shallow Adapter** The shallow adapter (SA) in CLIPSeg [18] learns to project the pretrained encoder representations before feeding them to the decoder network (Fig. 1b). The adapters are introduced at the skip connections of the CLIPSeg's encoders used by the vision-language decoder to predict segmentation masks. Since the original CLIPSeg model [18] used skip connections from $L_t \in \{3, 6, 9\}$ transformer [25] layers in the image encoder, three adapter layers are introduced at these connections. A similar strategy adds a skip connection with adapter modules for $L_t$ layers in the text encoder. CLIPSeg SA introduces

$d' = 512$ as the hidden dimension of the adapter block, resulting in 4.2 million trainable parameters.

**CLIPSeg Dense Adapter** The dense adapter (DA) in CLIPSeg learns to adjust the representation of the successive layers of the encoders before feeding to the decoder network (Fig. 1c). Following Houlsby et al. [10], we apply adapters before the two residual connections in each attention layer; two adapter blocks are used in each self-attention layer. We use adapter block up to $\max(L_t) = L_T = 9$ attention layers of the image encoder because, beyond the $L_T$ layer, the intermediate representations remain unused by the decoder. Similarly, DA implements the same pattern of adapters for the text encoder. DA also uses an adapter in CLIPSeg's text conditioning embeddings [18] to ensure consistency with SA. The hidden dimension of the block is $d' = 64$, which introduces only 3 million trainable parameters. The empirical results of Table 1 exhibit that DA surpasses SA in performance despite having fewer parameters.

The principal difference between SA and DA is that DA adapts the activations of each encoder block before feeding to the next one. In contrast, SA adapts the extracted internal activations fed to the decoder.

## 3 Experiments

### 3.1 Datasets

Recently, Poudel et al. [20] proposed a wide range of automatic prompt generation methods and benchmark fine-tuning of different CLIP-based VLSMs in eight medical imaging datasets from diverse modalities, including five non-radiology and three radiology datasets. Following the convention of that work, we use their text prompts and the same splits of datasets. Their proposed method generated multiple text prompts for an image-mask pair; for our empirical analysis, we randomly sample a text prompt among many to generate an image-mask-text triplet while iterating through the datasets.

Among the non-radiology datasets, three of them are endoscopic images with the polyp segmentation task (Kvasir-SEG [13], ClinicDB [3], and BKAI [19]), one with diabetic foot ulcer segmentation task (DFU [14]), and the last one has for skin-lesion segmentation (ISIC-16 [8]). Three different radiology images include segmentation of breast ultrasound (BUSI [2]), 2D-echocardiography (CAMUS [15]), and chest X-ray (CheXlocalize [22]).

### 3.2 Baseline Methods

We benchmark five models for our experimental analysis — two (CLIPSeg [18] and CRIS [26]) are trained with E2E fine-tuning, and three (SAN [27], CLIPSeg SA, and CLIPSeg DA) with adapter fine-tuning. SAN can generate segmentation masks from image-text inputs by training a ViT block [5] along frozen CLIP [21]. We are the first to use adapters for the pretrained encoder-decoder model for vision-language segmentation tasks, such as CLIPSeg DA and CLIPSeg SA. Since the adapter module proposed by Houlsby et al. [10] is incompatible with convolutional encoders, they are not practiced with the CRIS [26]. We use dice-score (DSC (%)), intersection-over-union (IoU (%)), and Hausdorff distance at the $95^{th}$ percentile (HD95) as metrics — all averaged over a dataset — to evaluate the overall performance of the methods.

### 3.3 Implementation Details

The training and inference of the baseline and proposed methods are performed in an NVIDIA RTX 3090. We use floating-point-16 mixed-precision training with a batch size of 32. The initial learning rates for the DA and SA are $1e-3$ and $3e-4$, respectively, with a scheduler that scales them by a factor of 0.3 if no decrease in validation loss is observed for 5 consecutive epochs. If no progress in the validation DSC (%) is observed for the 20 consecutive epochs, then the training is stopped; thus, there is no fixed number of training epochs. The models are optimized with AdamW [17] with a weight decay of $1e-3$. Also, each experiment is subjected to three different seed values to test the consistencies of the methods and account for the randomness in sampling the prompts. We combined dice and binary cross-entropy losses for the objective function, as shown by Eq. (2).

$$\mathcal{L} = \lambda_d \cdot \mathcal{L}_{Dice} + \lambda_{ce} \cdot \mathcal{L}_{BCE} \tag{2}$$

Here, $\lambda_d$ and $\lambda_{ce}$ are hyperparameters; we chose their values for our experiments as $\lambda_d = 1.5$ and $\lambda_{ce} = 1$.

## 4 Results and Discussions

**Variants of VLSM-Adapter.** In Fig. 2, we present the results of three different positioning of adapters in VLSMs as defined in Section 2.3. The results show that VL-Adapter performs best in most of the datasets — so, we have kept the performance of only this configuration in Table 1. VLC-Adapter displays the optimal performance in the ClinicDB [3] dataset. V-Adapter exhibits the best score for Kvasir-SEG [13], even superior to the upper bound set by CRIS [26] as indicated in Table 1. Since the adapters are sensitive to their placements in encoder branches for generalizing domain-specific distribution of the datasets, one should evaluate different placements of adapters before selecting one variant. (see **??** in the supplementary section for more metrics)

**CLIPSeg Adapter outperforms E2E fine-tuning.** With E2E fine-tuning for all radiology and non-radiology datasets, CLIPSeg-with-adapter shows superior performance for almost all the metrics compared to its counterpart, with no adapter module (see Table 1). CLIPSeg with adapter modules, despite having 47 times fewer trainable parameters than in an E2E setting, performing better than the latter shows the benefit of introducing learnable adapter modules with few

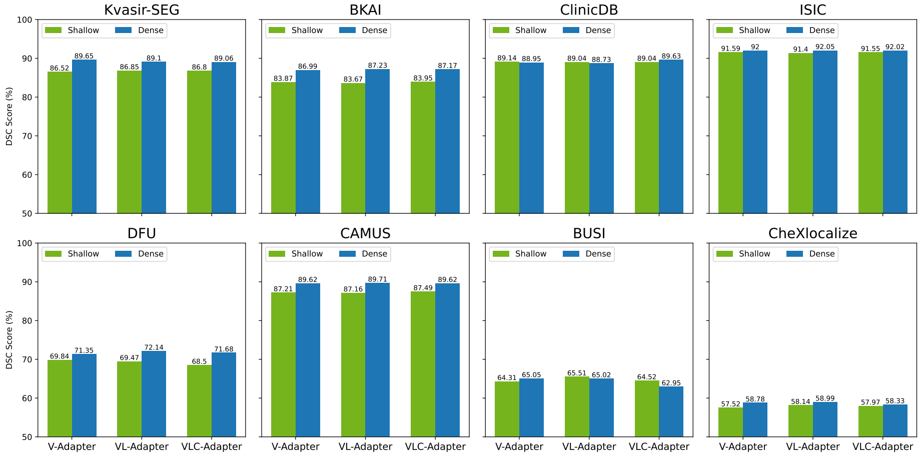


**Fig. 2. Dice Score (%) of variants of VLSM-Adapters across medical image datasets.** The dense adapter is better than the shallow adapter for almost all the datasets.

trainable parameters in intermediate layers of CLIPSeg, rather than fine-tuning the whole model for small datasets. Also, their performance is comparable to that of the state-of-the-art vision-only models for the individual datasets as reported by Poudel et al. [20].

**Parameter-Metric Trade-off.** In Table 1, the proposed CLIPSeg DA model performs better than the SAN [27] model despite having 2.6 times fewer learnable parameters. In the E2E fine-tuning cases, the CRIS [26] model has performed better than CLIPSeg [18] in almost all the datasets. The performance of CLIPSeg DA is on par with the CRIS model, even better in the ISIC-16 [8] dataset, regardless of having 46 times fewer parameters than CRIS. Also, this drop in metrics may not be significant in some scenarios with a high computation constraint, which is precisely where our proposed adapter models shine.

**SA vs. DA.** The DA network performs better than the SA network for most datasets, except for ClinicDB [3] and BUSI [2]; even in those two datasets, the metrics of the DA network are on par with that of SA, as in Table 1 and Fig. 3. Since there are more adapter layers in DA, we suspect the layers can adjust the internal representations of the pretrained models more finely compared to SA. Also, although the SA network has a broader adapter dimension of 512, it cannot outperform the DA network, which has only a 64 adapter dimension. This signifies that deeper adapter networks can capture complex representations despite having smaller projection dimensions.

**Table 1. Evaluation of models across diverse medical image datasets.** $*M$ represents the number of trainable parameters in millions. **Bold** shows the best score among adapter fine-tuned models. Gray depicts the performance from E2E fine-tuning.

| | | **Upper Bound** | | **Adapter Fine-tune** | | |
|---|---|---|---|---|---|---|
| **Datasets** | **Metrics** | CLIPSeg | CRIS | SAN | CLIPSeg SA (**Ours**) | CLIPSeg DA (**Ours**) |
| | | 150$M$ | 147$M$ | 8.4$M$ | 4.2$M$ | **3M** |
| Kvasir-SEG | DSC (%) ↑ | 87.69 | 89.43 | 69.58 | 86.85 | **89.10** |
| | IoU (%) ↑ | 81.72 | 83.37 | 58.05 | 79.26 | **82.39** |
| | HD95 ↓ | 54.02 | 55.23 | 130.75 | 52.18 | **47.79** |
| BKAI | DSC (%) ↑ | 85.59 | 92.62 | 66.26 | 83.67 | **87.23** |
| | IoU (%) ↑ | 77.52 | 88.30 | 54.58 | 75.02 | **79.81** |
| | HD95 ↓ | 87.91 | 49.80 | 224.37 | 87.79 | **70.02** |
| ClinicDB | DSC (%) ↑ | 88.58 | 93.63 | 81.36 | **89.04** | 88.73 |
| | IoU (%) ↑ | 81.51 | 88.74 | 72.61 | **81.93** | 81.84 |
| | HD95 ↓ | 19.30 | 12.36 | 38.42 | **18.03** | 18.76 |
| ISIC-16 | DSC (%) ↑ | 91.88 | 91.49 | 90.39 | 91.40 | **92.05** |
| | IoU (%) ↑ | 85.76 | 85.41 | 83.61 | 85.05 | **85.98** |
| | HD95 ↓ | 60.93 | 64.39 | 87.25 | 60.29 | **54.38** |
| DFU | DSC (%) ↑ | 72.12 | 74.01 | 63.38 | 69.47 | **72.14** |
| | IoU (%) ↑ | 61.61 | 64.31 | 51.63 | 58.27 | **61.42** |
| | HD95 ↓ | 38.24 | 41.92 | 60.10 | **38.75** | 38.79 |
| CAMUS | DSC (%) ↑ | 88.93 | 91.29 | 46.42 | 87.16 | **89.71** |
| | IoU (%) ↑ | 80.69 | 84.42 | 31.81 | 78.01 | **81.85** |
| | HD95 ↓ | 16.69 | 12.33 | 175.81 | 19.14 | **14.16** |
| BUSI | DSC (%) ↑ | 62.91 | 67.50 | 45.61 | **65.51** | 65.02 |
| | IoU (%) ↑ | 55.52 | 60.90 | 35.27 | **58.19** | 57.20 |
| | HD95 ↓ | 72.98 | 50.63 | 152.10 | **63.36** | 64.37 |
| CheXlocalize | DSC (%) ↑ | 58.51 | 60.76 | 44.37 | 58.14 | **58.99** |
| | IoU (%) ↑ | 45.45 | 47.99 | 31.97 | 44.84 | **46.01** |
| | HD95 ↓ | 537.57 | 519.21 | 724.55 | **533.04** | 535.97 |

## 5 Conclusion and Future Direction

We present a VLSM-Adapter module that adjusts to the downstream segmentation tasks without changing the parameters of the pretrained encoder-decoder architecture. We show that updating only the adapter parameters achieves on par performance to E2E fine-tuning, and is even better than the latter for some datasets. The dense adapter variant performing better in most cases than the shallow adapter one, despite having fewer parameters, shows that tweaking the internal representations of the pretrained models finely in smaller dimensions — dense adapter — is more crucial than coarsely adapting the representations in a higher dimensional space — shallow adapter. Also, one should be open to experimenting with positioning the adapters in vision or text encoder branches.

In this work, we have only used a VLSM adapter for semantic segmentation in the medical domain. The performance of adapter modules on other segmentation tasks for different domains is yet to be explored. Additionally, this paper does not benchmark the performance of the models with adapters in the language-only branch because of the higher influence of the image encoder in the decoder of

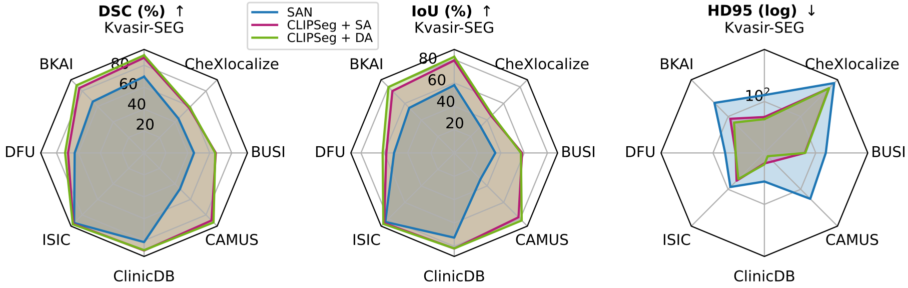


**Fig. 3. Evaluation of Adapter-finetuned models.** Our methods perform better than SAN, with Dense Adapter (DA) performing the best across diverse datasets.

CLIPSeg; the decoder has skip connections from intermediate representations of the image encoder. Future works can study the performance of models with adapters in the language encoder comparing it with the ones we demonstrated.

VLSM-Adapter also opens the pathways to continual learning and multi-task learning machines for VLSMs as specialized adapters could be trained for new data or tasks while keeping the core architecture frozen to prevent forgetting. These adapters allow efficient fine-tuning of large pretrained VLSMs for medical image segmentation where there are often datasets with small sizes.

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

## References

1. Adhikari, R., Dhakal, M., Thapaliya, S., Poudel, K., Bhandari, P., Khanal, B.: Synthetic Boost: Leveraging synthetic data for enhanced vision-language segmentation in echocardiography. In: International Workshop on Advances in Simplifying Medical Ultrasound. pp. 89–99. Springer (2023)
2. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
3. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics **43**, 99–111 (2015)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
6. Fellbaum, C.: Wordnet. In: Theory and Applications of Ontology: Computer Applications, pp. 231–243. Springer, Dordrecht (2010)

7. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: CLIP-Adapter: Better vision-language models with feature adapters. International Journal of Computer Vision pp. 1–15 (2023)
8. Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin Lesion Analysis toward Melanoma Detection: A Challenge at ISBI 2016, hosted by ISIC. arXiv preprint arXiv:1605.01397 (2016)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
10. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
11. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Association for Computational Linguistics. vol. 1, pp. 328–339 (2018)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
13. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-SEG: A segmented polyp dataset. In: MultiMedia Modeling. pp. 451–462. Springer (2020)
14. Kendrick, C., Cassidy, B., Pappachan, J.M., O'Shea, C., Fernandez, C.J., Chacko, E., Jacob, K., Reeves, N.D., Yap, M.H.: Translating clinical delineation of diabetic foot ulcers into machine-interpretable segmentation. arXiv preprint arXiv:2204.11618 (2022)
15. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on medical imaging **38**(9), 2198–2210 (2019)
16. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International conference on machine learning. pp. 97–105. PMLR (2015)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
18. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7086–7096 (2022)
19. Ngoc Lan, P., An, N.S., Hang, D.V., Long, D.V., Trung, T.Q., Thuy, N.T., Sang, D.V.: NeoUNet: Towards accurate colon polyp segmentation and neoplasm detection. In: Advances in Visual Computing. pp. 15–28. Springer (2021)
20. Poudel, K., Dhakal, M., Bhandari, P., Adhikari, R., Thapaliya, S., Khanal, B.: Exploring transfer learning in medical image segmentation using vision-language models. arXiv preprint arXiv:2308.07706 (2023)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
22. Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S.Q., Nguyen, C.D., Ngo, V.D., Seekins, J., Blankenberg, F.G., Ng, A.Y., et al.: Benchmarking saliency methods for chest x-ray interpretation. Nature Machine Intelligence **4**(10), 867–878 (2022)

23. Song, L., Xue, R., Wang, H., Sun, H., Ge, Y., Shan, Y., et al.: Meta-Adapter: An online few-shot learner for vision-language model. In: Advances in Neural Information Processing Systems. vol. 36, pp. 55361–55374 (2023)
24. Sung, Y.L., Cho, J., Bansal, M.: VL-Adapter: Parameter-efficient transfer learning for vision-and-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5227–5237 (2022)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. vol. 30, pp. 5998–6008 (2017)
26. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022)
27. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2945–2954 (2023)
28. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in neural information processing systems. vol. 27, pp. 3320–3328 (2014)
29. Yu, S., Seo, P.H., Son, J.: Zero-shot referring image segmentation with global-local context features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19456–19465 (2023)