Structure-Aware Cross-Modal Prompt Tuning for Autonomous Bronchoscopic Navigation

Hao Fang[1,2], Zhuo Zeng[3], Jianwei Yang[2], Wenkang Fan[2], and Xiongbiao Luo[1,2,4,*]

[1] National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen 361102, China
[2] Department of Computer Science and Engineering, Xiamen University, Xiamen 361102, China
[3] Xiamen University Tan Kah Kee College, Zhangzhou 363105, China
[4] Discipline of Intelligent Instrument and Equipment, Xiamen University, Xiamen 361102, China
xiongbiao.luo@gmail.com, Asterisk indicates the corresponding author

Abstract. Autonomous bronchoscopic navigation is vital for pulmonary disease diagnosis and treatment but still suffers from subtle anatomical variations and open-set bronchial variants. Current vision-language foundation models enable open-set recognition but get trapped in capturing fine-grained spatial features and disentangling class-specific attributes. We propose a structure-aware cross-modal prompt tuning framework that combines the contrastive language-image pre-training (CLIP) model and the efficient segment anything model (EfficientSAM) to address these limitations. Specifically, EfficientSAM extracts structure-aware features for learnable textual prompts via cross-modal attention to enrich visual embeddings in CLIP, while a base-unknown decoupled head disentangles shared anatomical knowledge and class-specific features in the latent space, enhancing separability for both base and open-set classes. Moreover, unified optimization aligns multi-modal distributions using image-text matching loss and base-unknown decoupled loss. We evaluate our method on clinical bronchoscopic data, with the experimental results showing that our method outperforms state-of-the-art approaches and improves recognition and open-set identification (88.94%, 87.00%).

Keywords: Surgical navigation · Foundation models · Prompt tuning · Object recognition and classification · Endoscopy.

1 Introduction

Automatic recognition of bronchial bifurcation plays a vital role in clinical bronchoscopy, encompassing applications in clinical diagnosis [9] and bronchoscopic navigation [21]. Both human operators and robotic systems depend on specific bronchial bifurcation landmarks for accurate bronchoscope localization [2]. The subtle morphological variations among bronchial bifurcations can confuse surgeons bringing potential risks and precise position recognition is also essential for accurate pose adjustment in bronchoscopic navigation [13].

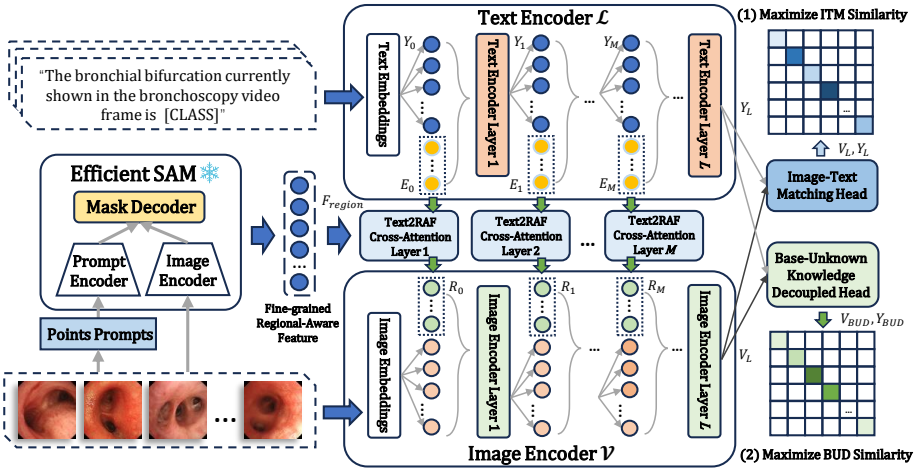


Fig. 1: The framework of our proposed foundation model with regional-aware cross-modal prompt tuning for fine-grained bronchial bifurcation recognition.

Current deep visual models have demonstrated remarkable capabilities in medical image recognition [17]. However, these models require a predetermined number of classes which is the significant limitation for recognizing diverse bronchial bifurcation variants. The emergence of advanced vision-language (V-L) foundation models, particularly the contrastive language-image pretraining (CLIP) [11], has shown promising results across various medical vision tasks[18, 15]. The reason ability about open-set visual concepts [12] of CLIP offers a potential solution for recognizing numerous bronchial bifurcation variants.

Prompt tuning is an efficient approach to transfer pre-trained knowledge of CLIP to downstream tasks [19]. Common text prompt is handcrafted in the template of "[Text Description] + [CLASS]" (e.g. "a photo of [CLASS]") where the quality of textual prompts significantly influences the performance [20]. Building upon CoOp [20], recent methods have employed context optimization with trainable parameters to automatically learn optimal contextual prompts. Jia et al.[4] advanced this field by introducing visual prompt tuning through trainable parameters in the visual branch's input space. The latest multi-modal prompt tuning approaches[5] align visual and textual prompts within encoder layers, leveraging the synergistic benefits of both modalities.

Although CLIP demonstrates remarkable capability in capturing high-level semantics, it shows limitations in representing low-level features such as textures and edge information [10]. This fundamental weakness indicates that depending solely on a single foundation model may be insufficient for accurately encoding the intricate features of bronchial bifurcations. Moreover, well-separation of class-specific features in the feature space is essential for effective classification, particularly when distinguishing between base and open-set classes which require the decoupling of shared and distinctive knowledge [19]. Current prompt tuning

methods often fall short in achieving optimal feature disentanglement [6], which can lead to ambiguity when differentiating between similar bronchial patterns. This challenge is further complicated by the need to maintain a balance between preserving base anatomical knowledge shared across all variants and capturing the unique discriminative features specific to each class.

To address these challenges, we propose structure-aware cross-modal prompt tuning (SCPT) to integrate CLIP and EfficientSAM [16] into a comprehensive foundation model for fine-grained bronchial bifurcation recognition. Our key technical contributions are threefold. First, we leverage EfficientSAM to extract fine-grained regional-aware bronchial bifurcation features capitalizing on its superior ability to capture low-level spatial details through segmentation-based feature extraction [14]. Second, we introduce a cross-attention mechanism [1] that fuses these regional-aware features with learnable textual prompts and incorporates them into the visual branch as cross-modal visual prompts which enhances feature representation of CLIP through cross-modal prompt tuning. Third, we design a Base-Unknown Knowledge Decoupled (BUKD) head that simultaneously decouples and compresses feature representations of base and open-set classes in the feature space. By optimizing both image-text matching loss and BUKD loss, our approach minimizes the disparity between image and text feature distributions while enhancing the discriminative alignment capability for bronchial bifurcation classification. Extensive experiments on a clinical bronchoscopy report dataset demonstrate that SCPT achieves state-of-the-art performance in recognizing both base and open-set bronchial bifurcation classes.

2 Structure-Aware Cross-Modal Prompt Tuning

In Fig. 1 Our proposed framework. Sect. 2.1 introduces the regional-aware features extraction pipeline through EfficientSAM. Sect. 2.2 illustrates SCPT design in foundation models. Finally, the functionality of BUKD head and the learning objective of the framework are demonstrated in Sect. 2.3.

2.1 Structure-Aware Enhancement

To capture low-level spatial features, we utilize the frozen EfficientSAM to extract region-aware features for fine-grained bronchial bifurcation classification. We investigate the morphological characteristics of bronchial bifurcation regions and discover airways (holes) and bronchial walls are naturally with distinct contrast (See Fig. 2). Therefore, we propose a automatic point prompts generation pipeline and Fig. 2 shows intermediate image outputs of each stage. We first adjust the brightness and intrinsic contrast between these two regions and obtain high-contrast bronchial bifurcation images. To further extract refined orifice edges, we convert RGB into HSV color space and set threshold to get orifice masks. Finally, we select N_p points from contours and center regions as $PointPrompts \in \mathbb{R}^{N_p \times 2}$ for EfficientSAM:

$$F_{region} = EfficientSAM(PointPrompts) \tag{1}$$

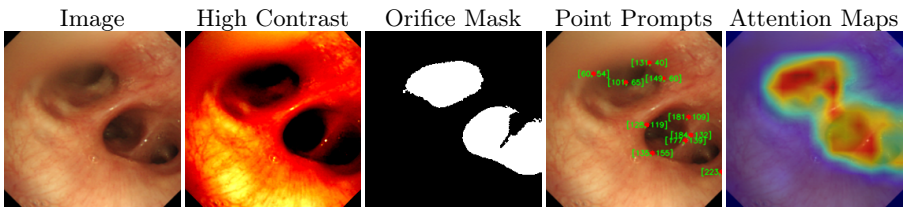


Fig. 2: Visualizations of original bronchial bifurcation image, high contrast image, rough orifice masks, extracted point prompts and attention maps outputted by Mask Decoder of EfficientSAM [16].

Note that we adopt the output of MaskDecoder in EfficientSAM as region-aware features $F_{region} \in \mathbb{R}^{N_r \times d_r}$ where N_r is the number of mask tokens in MaskDecoder and d_r denotes feature dimensionality. The attention map generated by MaskDecoder shown in Fig. 2 illustrates the effectiveness of extracting region-aware features.

2.2 Cross-modal Text Prompt Tuning

Adaptive Text Prompting. Like previous methods [19, 20], we replace "[Text Description]" with learnable tokens: $E_j = \{E_{i,j} \in \mathbb{R}^{d_l}\}_{i=1}^{N_T}$ for totally M text encoder layers to adaptively learn textual context prompts. The input adaptive word embedding of textual branch is represented as $[E_j; Y_j] \in \mathbb{R}^{(N_T+N_W)\times d_l}$ and the input and output representations of j-th text encoder layer are as:

$$Y_j = TextEncoderLayer_j([E_{j-1}; Y_{j-1}]), j = 1...M \ (M <= L) \tag{2}$$

where Y_0 corresponds to fixed input tokens and $Y_{j>0}$ is output embeddings. M is the number of layer using adaptive text prompt and L is the number of layers. d_l is the embedding dimensionality and $[*;*]$ is the concatenation operation.

Cross-Modal Text Visual Prompts We first fuse language semantic features with nuanced regional details and build cross-modal visual prompt for each visual encoder stage. The fusion of E_* with subtle details of bronchial bifurcations in F_{region} explicitly align textual features with fine-grained bifurcation features, leveraging the guiding role of textual representations to influence visual encoding. We devise a cross-modal Text2RAF (T2RAF) layer adopting cross-attention mechanisms [1] to achieve the interactions between adaptive context prompt tokens E_j and region-aware feature F_{region} from EfficientSAM. T2RAF layer absorbs E_j as query and F_{region} as key and value and the generation of cross-modal text-regional visual prompt R_j is as follows:

$$R_j = softmax(\frac{Q_{E_j} K_F^T}{\sqrt{d_l}}) V_F \tag{3}$$

$$Q_{E_j} = E_j W_{q,j}, K_F = F W_{k,j}, Q_E = F W_{v,j} \tag{4}$$

where Q_{E_j} is obtained by projecting E_j through $W_{q,j}$, while K_F and V_F are derived from F_{region} through projections $W_{k,j}$ and $W_{v,j}$ respectively. R_j retains text prompt guiding fine-grained orifice features and the concatenation of R_j and image embedding V_j denoted by $[R_j; V_j]$ is fed into image encoder layer j:

$$V_j = ImageEncoderLayer_j([R_{j-1}; V_{j-1}]), j = 1...M \ (M <= L) \tag{5}$$

where V_0 is the original image embedding.

2.3 Multimodal Text Alignment and Prompt Coupling

Image-Text Matching We follow CLIP to maximize the similarity between visual embedding V_L based on input image I and textual embedding Y_{L,c_i} based on textual prompts of class c_i. Suppose the label is $c_I \in \{1, 2...C\}$ with C classes, the common prediction probability is considered as:

$$P_{ITM}(c_I|I) = \frac{exp(sim < \Omega(V_L), \Omega(Y_{L,c_I}) >)/\tau_1}{\sum_{i=1}^{C} exp(sim < \Omega(V_L), \Omega(Y_{L,c_i}) >)} \tag{6}$$

where $sim < *, * >$ is the cosine similarity; $\Omega(\cdot)$ denotes the linear projection and τ_1 is the temperature.

Base-Unknown Knowledge Decouple We defined base and unknown classes for representing common and endless variant bronchial bifurcations. Our recognition task aims to determine whether the current bifurcation belongs to one of the base classes or represents an open-set unknown class. Due to the limited discrimination between similar bifurcation morphology and long-tail training dataset, we build Base-Unknown Knowledge Decoupled (BUKD) head to preserve base bifurcation class feature space away from open-set unknown class feature space. We apply cascade feature-wise projection and unknown feature compression [7] on visual features V_L to enlarge difference of base and unknown classes in high-denominational space:

$$V_{BUKD} = \phi_1(V_L) * (1 + \gamma_1 * (1 - c_i/(C+1)) * \mu(c_i)) \tag{7}$$

$$Y_{BUKD} = \phi_2(Y_L) * (1 + \gamma_2 * (1 - c_i/(C+1)) * \mu(c_i)) \tag{8}$$

$$\mu(c_i) = \begin{cases} 0, c_i \text{ is unknown class} \\ 1, c_i \text{ is base class} \end{cases} \tag{9}$$

where $\phi_*(\cdot)$ denotes linear projection, γ_1 and γ_2 are hyperparameters. When c_i (Class index) belongs to base class, the feature cluster is compressed in high denominational space which creates wider decision boundary. The bronchial bifurcation prediction probability of BUKD head is:

$$P_{BUKD}(c_I|I) = \frac{exp(sim < V_{BUKD}, Y_{BUKD,c_I} >)/\tau_2}{\sum_{i=1}^{C} exp(sim < V_{BUKD}, Y_{BUKD,c_i} >)} \tag{10}$$

For each image-text pair, the corresponding one-hot class label is represented by y_i. Cross-entropy loss is adopted to semantically align pairs while diverging

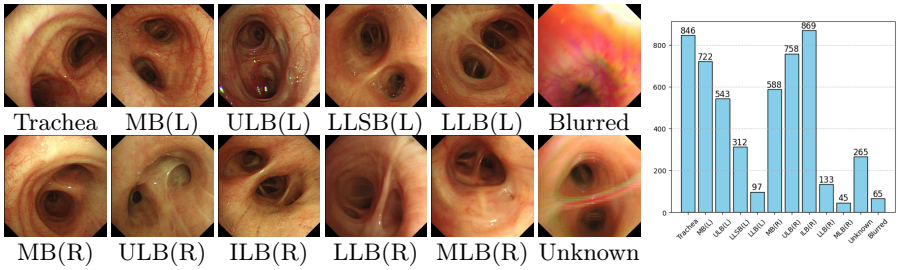


Fig. 3: Visualizations of bronchial bifurcation dataset. **Left:** (L) and (R) denote left and right bronchi (MB: Main Bronchi, ULB: Upper Lobe Bronchus, LLSB: Lower Lobe Superior Bronchus, LLB: Lower Lobe Bronchus, ILB: Intermediate Lobe Bronchus, MLB: Middle Lobe Bronchus). **Right:** Data distribution.

dissimilar image-text mappings in the feature space. Finally, the loss function for optimization is formulated by:

$$\begin{aligned}\mathcal{L}_{CE} &= \alpha\mathcal{L}_{ITM} + (1-\alpha)\mathcal{L}_{BUKD} \\ &= -(\alpha\sum_{i} y_i log P_{ITM}(c_i|I) + (1-\alpha)\sum_{i} y_i log P_{BUKD}(c_i|I))\end{aligned} \tag{11}$$

where α modulation parameter to balance the importance of the two losses.

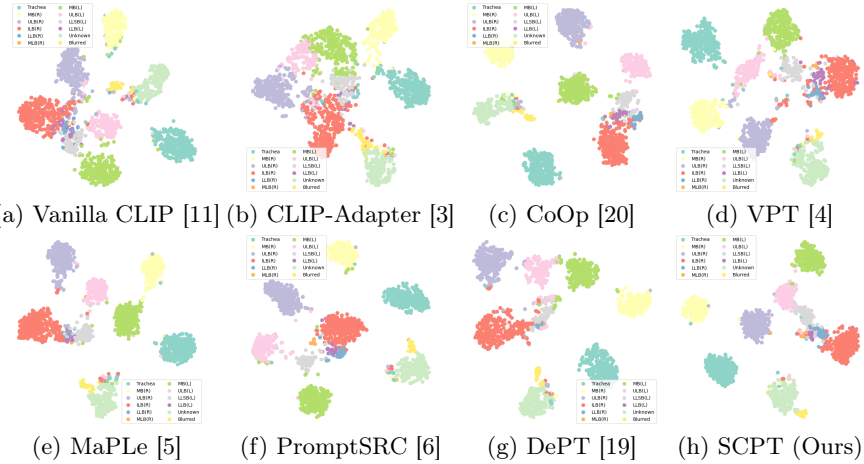
3 Experiments

We analyzed 1356 clinical bronchoscopy reports containing bronchial bifurcation images, yielding a total dataset of 5243 distinct bronchial images. The distribution of these images across different bronchial sites is illustrated in Fig. 3. We identified the ten most frequently examined locations and designated them as base classes. The remaining images were categorized as unknown open-set classes including blurred images and other less common bronchial bifurcations.

We conducted comprehensive comparisons with several state-of-the-art methods including Vanilla CLIP [11], CLIP-Adapter [3], CoOp [20], VPT [4], MaPLe [5], PromptSRC [6], and DePT [19]. For ablation studies, we evaluated our method by removing key components: EfficientSAM, T2RAF, and BUKD. In the variant w/o T2RAF, image embeddings were directly concatenated with F_{region}. We also investigate the performance of SCPT with distinct α ranging from 0.1 to 0.9 in $\mathcal{L}_{CE}$ where $\alpha = 1$ denotes $\mathcal{L}_{CE} = \mathcal{L}_{ITM}$. To evaluate performance, we employed metrics: classification accuracy, Macro F1-Score, and Macro Recall for base classes. Identification Rate (IDR) for open-set unknown classes while IDR calculates the proportion of unknown bronchial bifurcations that were successfully identified. We utilized a 7:3 train-test split ratio and implemented five-fold cross-validation on the bronchial bifurcation dataset to ensure robust evaluation of the recognition performance.

Table 1: Performance comparison of baselines for bronchial bifurcation recognition using Five-fold cross-validation.

Methods	Accuracy	F1-score	Recall	IDR	Extra Params
Vanilla CLIP [11]	74.94	55.31	56.56	78.00	0M
CLIP-Adapter [3]	79.40	62.52	63.08	76.40	0.52M
CoOp [20]	86.28	75.38	75.86	79.20	0.002M
VPT [4]	87.10	76.36	77.06	82.60	0.074M
MaPLe [5]	87.20	75.47	76.42	81.80	3.56M
PromptSRC [6]	86.43	74.65	74.40	80.80	0.046M
DePT [19]	87.35	75.52	76.11	83.10	3.57M
SCPT(Ours)	**88.94**	**79.57**	**80.71**	**87.00**	3.28M



(a) Vanilla CLIP [11] (b) CLIP-Adapter [3] (c) CoOp [20] (d) VPT [4]

(e) MaPLe [5] (f) PromptSRC [6] (g) DePT [19] (h) SCPT (Ours)

Fig. 4: t-SNE visualization of image embeddings on the bronchial bifurcation testing set (Fold 3) across eight different methods.

We apply frozen EfficientSAM-S [16] to extract regional-aware feature and the number of point prompts is set 10 and N_r for MaskDecoder is set 4. The visual encoder adopted is ViT-B/32 for CLIP and the number of adaptive textual token is 4 while M is set 3. We set γ_1 and γ_2 as 0.5 in BUKD. Note that visual and textual encoders are unfrozen for fine-tuning. We use AdamW [8] optimizer and the learning rate is $1e^{-5}$. The batch size is 48 and training epoch is 50. We use the same experimental setup for all baselines. All models are implemented on PyTorch and trained on 1× 24GB NVIDIA GTX 3090Ti GPU.

3.1 Results and Discussion

Recognition Table. 1 shows the average recognition performance and the size extra parameters of comparison methods on the bronchial bifurcation dataset. SCPT demonstrates superior performance in all evaluation metrics which under-

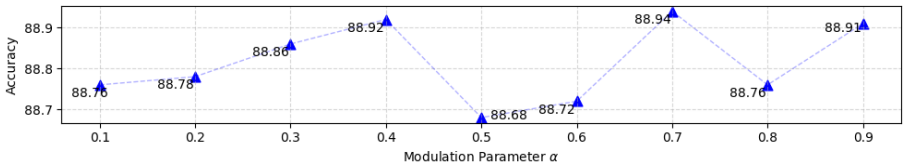


Fig. 5: Accuracy of SCPT with distinct modulation parameters α in $\mathcal{L}_{CE}$.

Table 2: Ablation Study of SCPT.

Methods	Accuracy	F1-score	Recall	IDR
w/o EfficientSAM	87.44	75.48	75.60	85.20
w/o T2RAF	88.71	78.99	80.16	83.40
w/o BUKD	88.58	78.87	80.19	85.80
SCPT(Ours)	**88.94**	**79.57**	**80.71**	**87.00**

scores its improved capability for fine-grained bronchial bifurcation classification. SCPT demonstrates significant improvements across multiple metrics: Accuracy increased by 1.59%, F1-score by 3.21%, and Recall by 3.65% compared to baseline methods. Furthermore, SCPT achieved +3.9% improvement in IDR showing its capability to better identify unknown classes. We employ t-SNE to visualize the image embeddings of different bronchial bifurcations generated by comparisons in Fig. 4. The feature point distribution of SCPT is more distinguishable and unknown classes are far away from the base classes qualitatively illustrating the better performance of SCPT. We also investigate the recognition accuracy of SCPT with different values of α where $\alpha = 0.7$ produces the best recognition performance in Fig. 5.

Ablation Study Table. 2 shows ablation studies on the key components of our method with bronchial bifurcation dataset. Compared to w/o EfficientSAM, the performance of SCPT significantly increases by +1.5% Accuracy, +4.09% F1-score, +5.11% Recall and +1.8% IDR. The fine-grained features extracted by EfficientSAM boost SCPT to focus more on low-level details of bronchial bifurcation to achieve better recognition. Without T2RAF component, IDR decreases by 3.6% which highlights the significance of the aggregation between textual and fine-grained features in open-set recognition. Moreover, the improved metrics with BUKD, +0.36% Accuracy, +0.7% F1-score, +0.52% Recall and +1.2% IDR, also demonstrate the contribution of BUKD in SCPT.

Discussion In comparison experiments, we discover than VPT achieves greater performance than CoOp (e.g. +0.82% in Accuracy) which reflects visual prompts provides more details for recognition in bronchial bifurcation recognition. Naturally, multi-model prompt tuning enhance the synergy of V-L semantic understanding and arise recognition accuracy as illustrated in the experimental results. Fig. 5 shows that the recognition performance of SCPT with distinct α values outperforms SCPT without BUKD (88.58%), demonstrating the ef-

fectiveness of BUKD. With additional fine-grained visual details, cross-modal prompt tuning and base-unknown knowledge decoupled, our proposed SCPT outperforms others in the experiments. Although our method outperforms others, it introduces a relative large number of additional parameters compared to several baseline approaches. Nevertheless, this increase remains acceptable for fine-tuning the foundation model.

In conclusion, the extensive experimental results demonstrate both the effectiveness of SCPT and highlight the significant potential of foundation models in addressing fine-grained recognition tasks.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 82272133, in part by the High-Quality Development Science and Technology Major Project of Xiamen Health Commission under Grant 2024GZL-ZD03, and in part by Ningbo 2035 Key Research and Development Program under Grant 2024Z127.

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

References

1. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 347–356 (2021). https://doi.org/10.1109/ICCV48922.2021.00041
2. Fried, I., Hoelscher, J., Akulian, J.A., Pizer, S., Alterovitz, R.: Landmark based bronchoscope localization for needle insertion under respiratory deformation. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6593–6600 (2023)
3. 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 **132**(2), 581–595 (Feb 2024)
4. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 709–727. Springer Nature Switzerland, Cham (2022)
5. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19113–19122 (2023)
6. Khattak, M.U., Wasim, S.T., Naseer, M., Khan, S., Yang, M.H., Khan, F.S.: Self-regulating prompts: Foundational model adaptation without forgetting. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15144–15154 (2023)
7. Li, J., Meng, Z., Shi, D., Song, R., Diao, X., Wang, J., Xu, H.: Fcc: Feature clusters compression for long-tailed visual recognition. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24080–24089 (2023). https://doi.org/10.1109/CVPR52729.2023.02306

8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
9. Maetani, T., Tanabe, N., Tanizawa, K., Sakamoto, R., Shiraishi, Y., Hayashi, Y., Uyama, M., Matsunashi, A., Sato, S., Suzuki, K., Masuda, I., Fukui, M., Kaji, S., Handa, T., Hirai, T.: Computed tomography morphological assessments of central airways in interstitial lung abnormalities and idiopathic pulmonary fibrosis. Respiratory Research **25**(1), 404 (Nov 2024)
10. Park, N., Kim, W., Heo, B., Kim, T., Yun, S.: What do self-supervised vision transformers learn? In: The Eleventh International Conference on Learning Representations (2023)
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021)
12. Shao, S., Bai, Y., Wang, Y., Liu, B., Liu, B.: Collaborative consortium of foundation models for open-world few-shot learning. Proceedings of the AAAI Conference on Artificial Intelligence **38**(5), 4740–4747 (Mar 2024)
13. Shen, M., Giannarou, S., Shah, P.L., Yang, G.Z.: Branch:bifurcation recognition for airway navigation based on structural characteristics. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) Medical Image Computing and Computer-Assisted Intervention MICCAI 2017. pp. 182–189. Springer International Publishing, Cham (2017)
14. Wang, H., Vasu, P.K.A., Faghri, F., Vemulapalli, R., Farajtabar, M., Mehta, S., Rastegari, M., Tuzel, O., Pouransari, H.: Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3635–3647 (2024). https://doi.org/10.1109/CVPRW63382.2024.00367
15. Wu, Y., Tang, J., Yao, Z., Li, M., Hong, Y., Yu, D., Gao, Z., Chen, B., Zhao, S.: Vertfound: Synergizing semantic and spatial understanding for fine-grained vertebrae classification via foundation models. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 763–772. Springer Nature Switzerland, Cham (2024)
16. Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., Krishnamoorthi, R., Chandra, V.: Efficientsam: Leveraged masked image pretraining for efficient segment anything. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16111–16121 (2024). https://doi.org/10.1109/CVPR52733.2024.01525
17. Yang, Y., Fu, H., Aviles-Rivero, A.I., Xing, Z., Zhu, L.: Diffmic-v2: Medical image classification via improved diffusion network. IEEE Transactions on Medical Imaging pp. 1–1 (2025)
18. Yu, X., Wu, Z., Zhang, L., Zhang, J., Lyu, Y., Zhu, D.: Cp-clip: Core-periphery feature alignment clip for zero-shot medical image analysis. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 88–97. Springer Nature Switzerland, Cham (2024)
19. Zhang, J., Wu, S., Gao, L., Shen, H.T., Song, J.: Dept: Decoupled prompt tuning. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12924–12933 (2024)

20. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (Sep 2022)
21. Zou, Y., Guan, B., Zhao, J., Wang, S., Sun, X., Li, J.: Robotic-assisted automatic orientation and insertion for bronchoscopy based on image guidance. IEEE Transactions on Medical Robotics and Bionics **4**(3), 588–598 (2022)