VertFound: Synergizing Semantic and Spatial Understanding for Fine-grained Vertebrae Classification via Foundation Models

Yinhao Wu[1] *, Jinzhou Tang[1] *, Zequan Yao[1], Mingjie Li[2], Yuan Hong[3], Dongdong Yu[4], Zhifan Gao[5], Bin Chen[4], and Shen Zhao[1,✉]

[1] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China
[2] Radiation Oncology, Stanford University, United States
[3] Zhejiang Normal University, Hangzhou, China
[4] The First Affiliated Hospital, Zhejiang University School of Medicine, China
[5] School of Biomedical Engineering, Sun Yat-sen University, Shenzhen, China
{inhaowu@gmail.com, tangjzh23@mail2.sysu.edu.cn}

Abstract. Achieving automated vertebrae classification in spine images is a crucial yet challenging task due to the repetitive nature of adjacent vertebrae and limited fields of view (FoV). Different from previous methods that leverage the serial information of vertebrae to optimize classification results, we propose VertFound, a framework that harnesses the inherent adaptability and versatility of foundation models for fine-grained vertebrae classification. Specifically, VertFound designs a vertebral positioning with cross-model synergy (VPS) module that efficiently merges semantic information from CLIP and spatial features from SAM, leading to richer feature representations that capture vertebral spatial relationships. Moreover, a novel Wasserstein loss is designed to minimize disparities between image and text feature distributions by continuously optimizing the transport distance between the two distributions, resulting in a more discriminative alignment capability of CLIP for vertebral classification. Extensive evaluations on our vertebral MRI dataset show VertFound exhibits significant improvements in both identification rate (IDR) and identification accuracy (IRA), which underscores its efficacy and further shows the remarkable potential of foundation models for fine-grained recognition tasks in the medical domain. Our code is available at https://github.com/inhaowu/VertFound.

Keywords: Foundation Models · Merging Models · Fine-grained Classification.

1 Introduction

The automated recognition of vertebrae in spinal images is crucial for a wide range of medical applications, including the diagnosis of spinal disorders, surgical

---

* These authors contributed equally to this work. This work is supported by the National Natural Science Foundation of China under Grants 62101607.
✉ Correspondence to Shen Zhao.

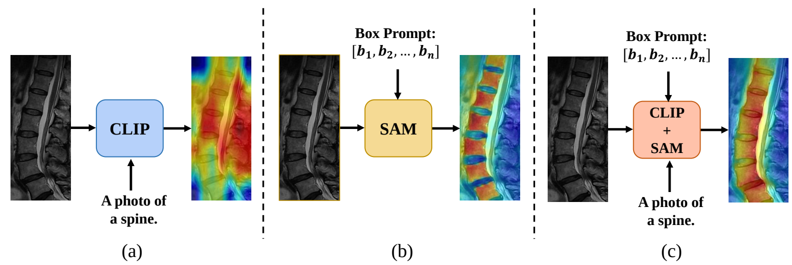


Fig. 1: Illustrations of the individual and combined capabilities of CLIP and SAM for feature representation in spine images. (a) CLIP excels in semantic understanding across the entire image. (b) SAM excels in spatial understanding for precise localization. (c) The merged model obtains features with enhanced local and global information.

planning, and postoperative assessment [1,2]. This process entails not only the precise localization of individual vertebrae but also their subsequent fine-grained classification. Among these tasks, achieving accurate classification poses a great challenge [3], particularly due to the subtle morphological differences among adjacent vertebrae and the constraints arising from arbitrary fields of view (FoV).

Existing methods mainly utilize the spatial relationships of vertebrae to enhance classification results that are aligned with the inherent sequential characteristics of vertebrae. For example, Yang *et al.* [4] enhanced vertebrae classification performance by proposing a message passing scheme to depict the spatial relationship of vertebrae. Wang *et al.* [5] combined key point localization with anatomically-constrained knowledge to facilitate vertebrae classification in CT images. Cui *et al.* [6] introduced a bidirectional relation module to capture the vertebral relationships among vertebrae with a self-attention mechanism. Wu *et al.* [7] designed a sequence loss based on dynamic programming to better preserve the sequential structure of vertebrae. However, the reliance on empirical designs and the scarcity of datasets may limit their robustness to data variability and transferability across different domains.

Recently, large-scale pre-trained foundation models, such as CLIP [8] and SAM [9], have shown remarkable potential with exceptional adaptability and flexibility in the medical domain. For example, Liu *et al.* [10] achieved universal segmentation on partially labeled datasets by leveraging CLIP text embeddings to comprehend the anatomical relationships of different organs and tumors. Lao *et al.* and Qin *et al.* [11,12] introduced different approaches to effectively fuse the different text prompts to enhance the abilities of GLIP [13] for zero-shot lesion detection. Cheng *et al.* and Wang *et al.* [14,15] showed the improved performance of SAM in medical images by introducing additional adapter layers.

However, the use of foundation models in fine-grained vertebrae classification faces considerable challenges due to their deficiencies in effectively exploiting vertebral spatial relationships and their limited discrimination against similar vertebral morphologies. On the one hand, different pre-training objectives endow foundation models with different capabilities in feature representation. As illustrated in Fig.1(a) and (b), contrastive learning-based models such as CLIP excel at capturing high-level semantic information [16], while segmentation models such as SAM excel at capturing low-level spatial details [17]. This disparity indicates that employing a single foundation model might not adequately capture the positional information of vertebrae. On the other hand, foundation models struggle to handle fine-grained classification tasks [18], often lacking the necessary discriminative capacity for the intricate vertebral textures.

To address these challenges, we propose **VertFound**, a framework designed to efficiently synergize the strengths of CLIP and SAM into a unified foundation model for fine-grained vertebrae classification. Firstly, feature extractors from CLIP and SAM are employed to obtain corresponding image-level and region-level features. The proposed vertebral positioning with cross-model synergy (VPS) module then enriches the feature representation by adopting the dot product attention mechanisms. Specifically, VPS utilizes two different attention modules (i.e., CLIP2SAM and SAM2CLIP) to facilitate effective information integration. The CLIP2SAM attention employs image-level features as the queries, while the region-level features serve as both keys and values, which enables local features with global dependencies. Conversely, SAM2CLIP attention employs region-level features as queries and image-level features as keys and values, thereby enriching semantic features with spatial context. This bidirectional attention mechanism amalgamates the strengths of CLIP and SAM, yielding enriched feature representations that encapsulate vertebral spatial relationships. Subsequently, we incorporate textual information as an additional modality input to leverage the image-text alignment capabilities of CLIP for achieving fine-grained classification within vertebral regions. Considering the significant similarities in feature distributions within vertebral images and textual descriptions, we further introduce a Wasserstein loss that transfers contrastive learning into an optimal transport problem by continually optimizing the transport cost between the two distributions. This enables the model to minimize the disparities between the image and text feature distributions and enhance the discriminative alignment capability of CLIP for vertebral classification. Empirical validation underscores the efficacy of VertFound while demonstrating the huge potential of foundation models for fine-grained classification tasks in the medical domain.

2 Methodology

As depicted in Fig.2, VertFound has two main stages. First, CLIP and SAM image encoders are employed to extract image-level and region-level features. The VPS module then combines these features to enhance vertebral position information. In the second stage, fine-grained vertebrae classification is achieved through image-text alignment within CLIP, with an additional Wasserstein loss to improve alignment score discrimination by optimizing the transport distance between image and text feature distributions.

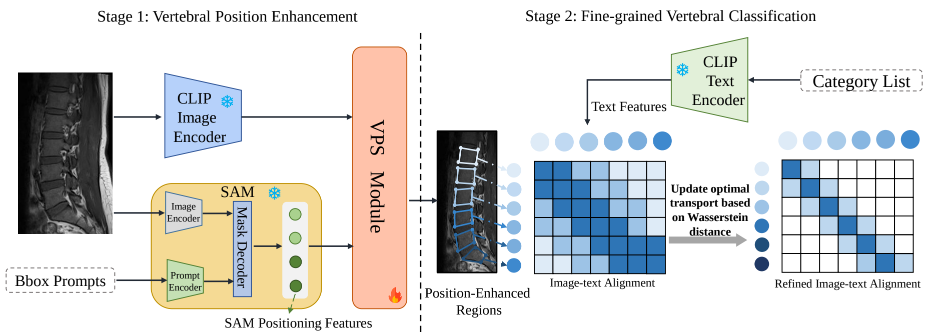


Fig. 2: The two-stage framework of the proposed VertFound. The first stage obtains comprehensive feature representations by employing the VPS module to effectively synergize semantic and spatial features for capturing vertebral position information. The second stage achieves fine-grained vertebrae classification by leveraging image-text alignment within CLIP, while a Wasserstein loss is proposed to enhance the discriminative alignment capability.

2.1 Vertebral Position Enhancement

Image- and Region-Level Feature Extraction Capitalizing on the robust generalization capabilities of foundation models, we directly employ the frozen visual encoders of CLIP and SAM as feature extractors to obtain the corresponding image features at the image-level and region-level. Specifically, we input images I with a shape of $H \times W$ into the vision transformer (ViT) model of CLIP to yield multi-stage features $C_l \in \mathbb{R}^{h,d_c}$ with image-level semantic information, where l represents the l-th stage, h and d_c denote the channels and feature dimensionality, respectively. On the other hand, images combined with their associated annotated box prompts are input into the image encoder and prompt encoders of SAM to produce relevant image and prompt embeddings. The two embeddings are then merged to obtain region-level embeddings $R \in \mathbb{R}^{n,d_s}$ with spatial information, where n represents the number of bounding boxes, d_s denotes feature dimensionality.

Vertebral Positioning with Cross-Model Synergy Module To fully leverage the semantic and spatial knowledge from CLIP and SAM, we propose the **V**ertebral **P**ositioning with Cross-Model **S**ynergy (**VPS**) module that enhances feature representations with more vertebral spatial details. Specifically, as shown in Fig. 3, VPS adopts the dot product attention mechanisms [19] (i.e., SAM2CLIP and CLIP2SAM) to achieve the interactions between global and spatial features. First, CLIP2SAM receives image-level features C_l (e.g., $\{C_{16}, C_{20}, C_{24}\}$) as queries and region-level features R as both keys and values to enrich semantic features with nuanced spatial details:

$$C_l' = \text{softmax}\left(\frac{Q_l K^T}{\sqrt{d}}\right) V \tag{1}$$

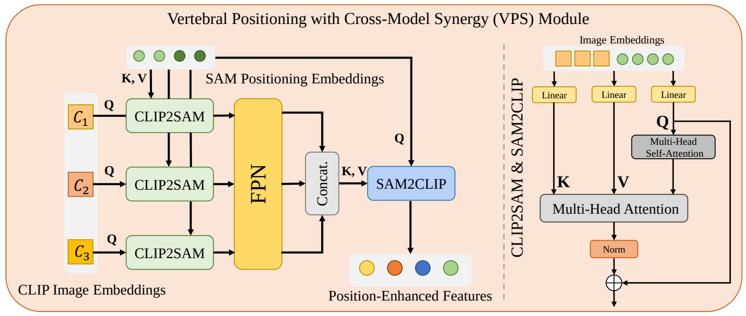


Fig. 3: The paradigm of the VPS module. It employs bidirectional attention mechanisms (i.e., CLIP2SAM and SAM2CLIP) to inherit the advantages of SAM and CLIP and obtain vertebral position-aware features.

$$Q_l = C_l W_q, K = R W_k, V = R W_v \tag{2}$$

where $W_q \in \mathbb{R}^{d_c,d}, W_k \in \mathbb{R}^{d_s,d}$ and $W_v \in \mathbb{R}^{d_s,d}$ are learnable weights, in which d is the dimensionality of the feature space for the queries, keys, and values. Subsequently, a Feature Pyramid Network (FPN) [20] is employed to process C'_l derived from Equation 1 and obtain image-level features $\hat{C} \in \mathbb{R}^{h,d_c}$ with multi-scale information. The refined features $\hat{C}$ are then utilized as keys and values in the SAM2CLIP mechanism, with region-level features R serving as queries, which greatly embeds the spatial understanding with global features. Through the integration of SAM and CLIP, we anticipate that the resultant model will assimilate the representation-level strengths of each, thereby enhancing its understanding of vertebral spatial relationships.

2.2 Fine-grained Vertebral Classification

Region-Text Alignment Beyond learning richer visual representations, we also integrate textual information to harness CLIP's capabilities in image-text alignment for improved vertebral classification. Similar to CLIP, we compute the alignment matrix $S \in \mathbb{R}^{n,m}$ between the refined region features $V \in \mathbb{R}^{n,d_c}$, generated by the VPS, and the text features $T \in \mathbb{R}^{m,d_c}$, extracted from a list of m vertebral categories (e.g., S, L5, L4, ..., C1) by using the text encoder of CLIP. This computation is formulated as follows:

$$S_{ij} = \frac{\phi_v(V)\phi_t(T)^T}{\tau\sqrt{\sum_k \phi_v(V_{ik})^2}\sqrt{\sum_k \phi_t(T_{jk})^2}} \tag{3}$$

Here, ϕ_v and ϕ_t represent different learnable feature projections on image and text features, while τ is a temperature parameter. Based on this, we calculate the cross-entropy loss (CEL) to explicitly push away representations from different image-text pairs while pulling together those that share the same semantics:

$$\mathcal{L}_{CEL} = -\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} \log S_{ij} \tag{4}$$

where $Y \in \{0,1\}^{n,m}$ represents the one-hot vertebral category of the image regions. This approach facilitates precise vertebrae classification by harnessing CLIP's alignment capabilities at the region level.

Wasserstein Loss However, considering the significant similarities between different vertebrae and their corresponding textual descriptions, the image-text alignment score S obtained in Eq. 3 might lack sufficient discrimination for fine-grained vertebral classification. Inspired by the work [21], we propose a Wasserstein loss (WSL) that casts contrastive learning into an *optimal transport problem*. Specifically, we optimize the Wasserstein distance $d_M^\lambda(V,T)$ to reduce the transport cost between two distributions, which results in more discriminative alignment scores for vertebral classification. The calculation of $d_M^\lambda(V,T)$ is formulated as follows:

$$d_M^\lambda(V,T) = \min_{P \in U(V,T)} \sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} M_{ij} + \frac{1}{\lambda}(-\sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} \log P_{ij}) \tag{5}$$

where $U(V,T) = \{P \in \mathbb{R}_+^{n,m} \mid P\mathbf{1}_n = V,\ P^T\mathbf{1}_m = T\}$ represents all possible transport matrices; $\mathbf{1}_n$ and $\mathbf{1}_m$ denote the vectors of ones in dimension n and m; λ is the penalty term associated with the distribution P. M_{ij} quantifies the difference between the i^{th} image and the j^{th} text and is defined as:

$$M_{ij} = -\frac{\exp(S_{ij})}{\sum_j \exp(S_{ij})} \tag{6}$$

The Sinkhorn-Knopp algorithm [22] is used to iteratively solve for the optimal solution. As suggested in Eq.5, a smaller $d_M^\lambda(V,T)$ signifies greater similarity among matched image-text pairs while a larger distance indicates a weaker correlation. Therefore, we directly adopt the Wasserstein distance $d_M^\lambda(V,T)$ as the WSL to achieve the fine-grained alignment between image and text:

$$\mathcal{L}_{WSL} = d_M^\lambda(V,T) \tag{7}$$

2.3 Model Optimization

Finally, VertFound employs a dual-loss optimization strategy to enhance the overall vertebrae classification accuracy:

$$\mathcal{L}_{total} = \mathcal{L}_{CEL} + \mathcal{L}_{WSL} \tag{8}$$

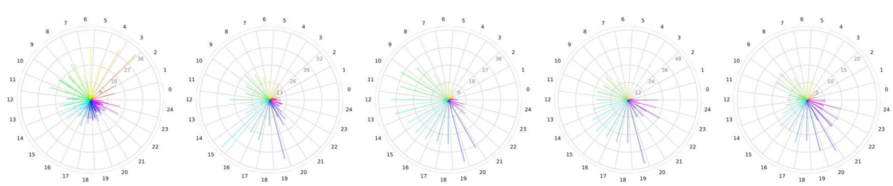
3 Experiment Results

3.1 Datasets and Evaluation Metrics

We evaluate the proposed VertFound on an in-house dataset that contains 1233 2D MRI images from 266 patients with a variety of vertebral appearances and FOVs, and a five-fold cross-validation approach is used for a thorough evaluation. Furthermore, we use identification rate (IDR) and image identification accuracy (IRA) as evaluation metrics. IDR calculates the ratio of vertebrae that are successfully detected, whereas IRA calculates the ratio of images that have all of their vertebrae correctly identified.

3.2 Implementation Details

All input images are resized to 224×224 and 1024×1024 and sent to CLIP and SAM image encoders via frozen ViT-L/14 and ViT-B/16, respectively. During training, the annotated bounding boxes are input to SAM, while in the testing phase the bounding boxes are predicted by a pre-trained detector (YOLOv8). The frozen text encoder of CLIP is employed to extract text features from a total of 25 category names of vertebrae (e.g., S, L5, L4, ... C1). The AdamW optimizer is employed with an initial learning rate of 2.5×10^{-5}, following a warm-up multi-step schedule with weight decay of 0.0001, Adam momentum of 0.9, and batch size of 40. Our methods are implemented in Python using the PyTorch framework and trained on an NVIDIA GTX 1080 Ti GPU. The CEL and WSL are all used for classification optimization, and the loss weights are all empirically set to 1.



(a) wo. CLIP (b) wo. SAM2CLIP (c) wo. CLIP2SAM (d) wo. WSL (e) Full model

Fig. 4: Confusion stars show the classification errors of different ablation studies. Specifically, the circle is divided into 25 regions, each representing a distinct class. Within each region, there are 24 subsections corresponding to the other classes, indicating the number of classification errors for each particular class.

3.3 Results

Comparison with Competing Methods As shown in Table 1, in-depth comparisons with current methods are carried out to evaluate the effectiveness of

Table 1: Comparison with Competing Methods.

Methods	IDR(%)	IRA(%)
DETR [23]	75.58 ± 5.39	46.60 ± 4.26
YOLOv5 [24]	76.46 ± 2.29	54.30 ± 3.27
YOLOv8 [25]	77.21 ± 1.49	55.43 ± 2.46
GLIPv1 [13]	74.31 ± 2.63	47.87 ± 2.19
OWL-ViT [26]	76.68 ± 3.15	54.81 ± 3.56
Ours (VertFound)	**89.05** ± 5.82	**83.73** ± 4.46

Table 2: Ablation study Results.

Methods	IDR(%)	IRA(%)
wo.CLIP	76.31 ± 7.08	64.86 ± 3.62
wo. CLIP2SAM	84.15 ± 2.46	80.18 ± 3.69
wo. SAM2CLIP	61.64 ± 2.12	65.26 ± 4.89
wo. WSL	82.50 ± 4.51	75.97 ± 4.38
Ours (VertFound)	**89.05** ± 5.82	**83.73** ± 4.46

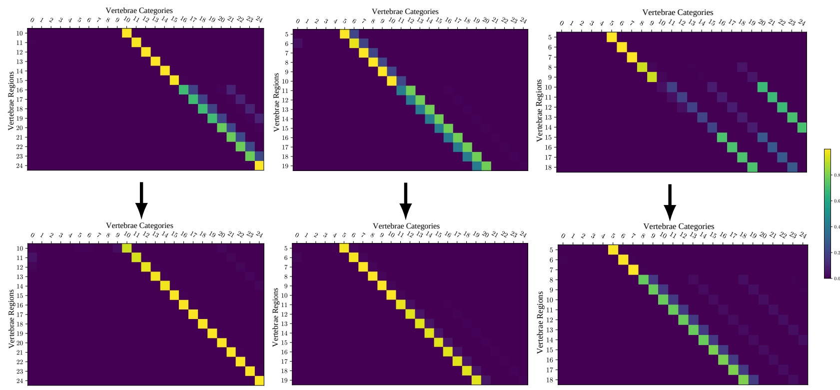
VertFound. All experiments adopt identical experimental settings to ensure fairness and reliability in the comparative evaluation. Our baselines include classical object detection algorithms (i.e., DETR [23], YOLOv5 [24], and YOLOv8 [25]) and recent foundation models (i.e., GLIPv1 [13] and OWL-ViT [26]). VertFound demonstrates superior performance in both IDR and IRA, which underscores its improved capability for fine-grained vertebrae classification.

Ablation Studies We conducted thorough ablation studies to verify the efficacy of the crucial components in VertFound. As presented in Table 2, removing CLIP (relying solely on SAM for classification) results in a noticeable decrease in classification performance. Meanwhile, removing either the CLIP2SAM or SAM2CLIP component in the VPS module also shows a noticeable decrease in both IDR and IRA, which highlights the significance of the aggregation between global and local features for precise vertebral positioning. Besides, the confusion star [27] in Fig.4 vividly illustrates the classification errors arising from different ablation methods. Furthermore, the removal of WSL also leads to a decline in overall performance. Figure 5 further elucidates that the incorporation of WSL optimizes the image-text alignment scores with varying degrees, thereby enhancing the discrimination against fine-grained vertebral features.

4 Conclusion

In this paper, we propose VertFound, which combines semantic and spatial understanding to achieve fine-grained classification of vertebrae. We introduce a novel VPS module that leverages the complementary strengths of CLIP and SAM for enriching vertebral feature representations. A WSL is introduced to enhance the discriminative alignment capability of CLIP for fine-grained vertebrae classification. Empirical validation shows the efficacy of VertFound while demonstrating the remarkable potential of foundation models for fine-grained recognition tasks. In the future, we will extend our experiments to include additional datasets from different sources and modalities to test the generalization capabilities of our method, including 3D MRI and CT scans.

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



(a) Easy case (b) Medium case (c) Hard case

Fig. 5: Different examples show the effectiveness of WSL in improving the discriminative capability of image-text alignment.

References

1. Haofu Liao, Addisu Mesfin, and Jiebo Luo. Joint vertebrae identification and localization in spinal ct images by combining short-and long-range contextual information. *IEEE transactions on medical imaging*, 37(5):1266–1275, 2018.
2. Shen Zhao, Bin Chen, Heyou Chang, Bo Chen, and Shuo Li. Reasoning discriminative dictionary-embedded network for fully automatic vertebrae tumor diagnosis. *Medical Image Analysis*, 79:102456, 2022.
3. Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. A convolutional approach to vertebrae detection and labelling in whole spine mri. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, pages 712–722. Springer, 2020.
4. Dong Yang, Tao Xiong, Daguang Xu, Qiangui Huang, David Liu, S Kevin Zhou, Zhoubing Xu, JinHyeong Park, Mingqing Chen, Trac D Tran, et al. Automatic vertebra labeling in large-scale 3d ct using deep image-to-image network with message passing and sparsity regularization. In *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25*, pages 633–644. Springer, 2017.
5. Fakai Wang, Kang Zheng, Le Lu, Jing Xiao, Min Wu, and Shun Miao. Automatic vertebra localization and identification in ct by spine rectification and anatomically-constrained optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5280–5288, 2021.
6. Zhiming Cui, Changjian Li, Lei Yang, Chunfeng Lian, Feng Shi, Wenping Wang, Dijia Wu, and Dinggang Shen. Vertnet: Accurate vertebra localization and identification network from ct images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 281–290. Springer, 2021.
7. Han Wu, Jiadong Zhang, Yu Fang, Zhentao Liu, Nizhuan Wang, Zhiming Cui, and Dinggang Shen. Multi-view vertebra localization and identification from ct images.

In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 136–145. Springer, 2023.
8. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
9. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.
10. Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023.
11. Miaotian Guo, Huahui Yi, Ziyuan Qin, Haiying Wang, Aidong Men, and Qicheng Lao. Multiple prompt fusion for zero-shot lesion detection using vision-language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 283–292. Springer, 2023.
12. 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.
13. Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.
14. 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.
15. Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, et al. Sam-med3d. *arXiv preprint arXiv:2310.15161*, 2023.
16. Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? *arXiv preprint arXiv:2305.00729*, 2023.
17. Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. *arXiv preprint arXiv:2310.15308*, 2023.
18. Denis Coquenet, Clément Rambour, Emanuele Dalsasso, and Nicolas Thome. Leveraging vision-language foundation models for fine-grained downstream tasks. *arXiv preprint arXiv:2307.06795*, 2023.
19. 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.
20. Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

21. Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019.
22. Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.
23. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
24. Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, imyhxy, Lorna, Zeng Yifu, Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Victor Sonck, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, November 2022.
25. Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, January 2023.
26. Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 728–755, Cham, 2022. Springer Nature Switzerland.
27. Amalia Luque, Mirko Mazzoleni, Alejandro Carrasco, and Antonio Ferramosca. Visualizing classification results: Confusion star and confusion gear. *IEEE Access*, 10:1659–1677, 2021.