CausalCLIPSeg: Unlocking CLIP's Potential in Referring Medical Image Segmentation with Causal Intervention

Yaxiong Chen[1,2], Minghong Wei[1⋆], Zixuan Zheng[3], Jingliang Hu[3], Yilei Shi[3], Shengwu Xiong[1,2], Xiao Xiang Zhu[4], and Lichao Mou[3(✉)]

[1] Wuhan University of Technology, Wuhan, China
[2] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[3] MedAI Technology (Wuxi) Co. Ltd., Wuxi, China
`lichao.mou@medimagingai.com`
[4] Technical University of Munich, Munich, Germany

Abstract. Referring medical image segmentation targets delineating lesions indicated by textual descriptions. Aligning visual and textual cues is challenging due to their distinct data properties. Inspired by large-scale pre-trained vision-language models, we propose CausalCLIPSeg, an end-to-end framework for referring medical image segmentation that leverages CLIP. Despite not being trained on medical data, we enforce CLIP's rich semantic space onto the medical domain by a tailored cross-modal decoding method to achieve text-to-pixel alignment. Furthermore, to mitigate confounding bias that may cause the model to learn spurious correlations instead of meaningful causal relationships, CausalCLIPSeg introduces a causal intervention module which self-annotates confounders and excavates causal features from inputs for segmentation judgments. We also devise an adversarial min-max game to optimize causal features while penalizing confounding ones. Extensive experiments demonstrate the state-of-the-art performance of our proposed method. Code is available at `https://github.com/WUTCM-Lab/CausalCLIPSeg`.

Keywords: referring medical image segmentation · CLIP · causal intervention · cross-modal decoding.

1 Introduction

Medical image segmentation is a crucial prerequisite for numerous clinical applications [1,2] such as computer-aided diagnosis, surgical planning, and image-guided intervention. However, the inherent complexity and variability of medical images often poses significant challenges for automatic segmentation methods. In particular, lesions can induce severe distortions and intensity alterations that confuse segmentation models. Incorporating prior knowledge, e.g., lesion count and location, may help.

⋆ Work done during an internship at MedAI Technology (Wuxi) Co. Ltd.

Medical notes written by physicians provide descriptive details and contextual information about medical images that can aid segmentation models in better delineating lesions. Moreover, accessing these texts incurs marginal additional cost, as they are already stored in electronic medical record systems within medical institutions. Unlike manual annotations which require extra labor, we can utilize the existing medical notes to improve segmentation performance without needing additional data collection and labeling. As a pioneering work, [3] extends the QaTa-COV19 dataset by adding text descriptions and designs a referring segmentation network. Yet so far, there is a paucity of literature focusing on this multi-modal segmentation task utilizing both images and texts.

In the field of multi-modal deep learning, CLIP [4], a large-scale pre-trained vision-language model based on contrastive learning, has recently attracted increasing attention. CLIP, in contrast to uni-modal foundation models like SAM [5], is better suited for multi-modal tasks. It learns image-level visual concepts from 400 million image-text pairs, enabling it to understand semantic relationships between images and texts. CLIP has proven highly effective across a diverse set of uni-modal and multi-modal downstream tasks [4,6,7,8,9,10,11], suggesting its potential utility for the challenging problem of referring medical image segmentation. However, the rich knowledge encapsulated in CLIP is derived from its image-level training objective. Consequently, directly applying CLIP does not yield optimal performance on pixel-level dense prediction tasks. In this paper, we employ CLIP for medical image segmentation from referring expressions, which has not been thoroughly investigated. Our key insight is that even though CLIP has not been trained on medical data, we can inherit much of its latent space's virtue by enforcing its powerful and semantic structure onto the medical domain.

On the other hand, medical images often contain lesions with ambiguous boundaries [12,13,14], while some lesion-free regions have similar appearances to lesions, confusing segmentation models. This arises due to confounders that induce models to learn spurious correlations based on conventional likelihood estimation. Moreover, the use of CLIP exacerbates this problem to some extent, as CLIP's training objective differs from the task of referring medical image segmentation. Mitigating such confounding bias necessitates causal intervention [15]. For instance, collecting images of a specific object in all contexts causes models to focus on the target rather than context. However, exhaustively collecting medical images with all possible lesion backgrounds is infeasible. To address this, we propose a causal intervention module that self-annotates confounders in an unsupervised manner and excavates causal features from inputs, improving segmentation. We assume inputs comprise causal and non-causal factors, with only the former determining segmentation judgments. The module's goal is to suppress confounding features and retain causal ones. Additionally, we devise an adversarial min-max game, optimizing causal features while penalizing confounding ones. Our approach facilitates learning lesion-relevant representations without requiring exhaustive sampling of contexts.

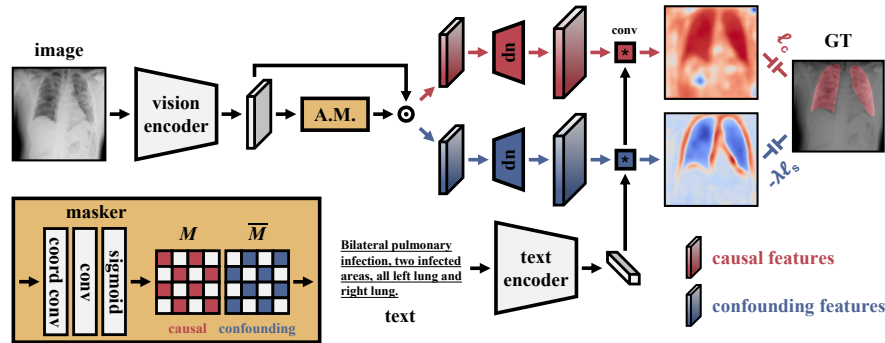Our technical contributions are three-fold:

Fig. 1. Overview of the proposed CausalCLIPSeg model.

- We introduce CLIP to referring medical image segmentation and design a framework called CausalCLIPSeg.
- We propose a causal intervention module to eliminate confounding bias.
- Our approach achieves state-of-the-art performance on the QaTa-COV19 dataset.

Background of Causality In the context of this work, confounding bias refers to the interference of background factors and other extraneous variables in the model's segmentation of lesion areas. Spurious correlation is a related concept to confounding bias. A meaningful causal relationship is the ideal scenario where a model relies solely on visual features of lesion areas to generate lesion masks, without being influenced by confounding factors. Excavating causal features refers to extract visual features that satisfy the meaningful causal relationship. Self-annotating confounders refers to our model's ability to adaptively extract confounders.

2   Method

Fig. 1 illustrates the proposed CausalCLIPSeg model which comprises three key components: (i) CLIP-driven text and vision encoders that extract multimodal feature representations, (ii) a decoder that achieves text-pixel alignment, and (iii) a causal intervention module that eliminates confounding bias between segmentation targets and spurious correlations.

2.1   CLIP-Driven Text and Vision Encoders

Text Encoder We employ the same modified Transformer architecture from [4] as our text encoder. For an input expression, we tokenize it into a text sequence using a byte pair encoding (BPE) tokenizer with a 49,152 vocabulary size, identical to that used in GPT-2 [16]. The text sequence is bracketed with [SOS] and

`[EOS]` tokens. The Transformer performs on this sequence, and activations of its highest layer at the `[EOS]` token are transformed as a global textual representation $\boldsymbol{\tau} \in \mathbb{R}^T$, where T is the feature dimension.

Vision Encoder For the vision encoder, we choose pre-trained ResNet-101. For an input image $\boldsymbol{I} \in \mathbb{R}^{1 \times H \times W}$, we extract multi-scale visual features from stages 2 to 4, denoted as $\boldsymbol{F}_i$. Here, H and W represent the height and width of the image.

2.2   Cross-Modal Decoding for Text-Pixel Alignment

Our decoder $\mathcal{D}$ aligns the text feature $\boldsymbol{\tau}$ with pixels in the input image to generate a segmentation mask. Specifically, $\boldsymbol{\tau}$ is first mapped into a D-dimensional embedding space through a learnable linear projection, where $D = C \times K \times K + 1$. The resulting embedding is further split into $\boldsymbol{W}_\tau \in \mathbb{R}^{C \times K \times K}$ and $b_\tau \in \mathbb{R}$. On the other hand, visual feature maps $\boldsymbol{F}$ are upsampled to the full resolution by interpolation and convolution. Then, the textual and visual features are cross-correlated:

$$\boldsymbol{S} = \mathcal{D}(\boldsymbol{\tau}, \boldsymbol{F}) = \boldsymbol{W}_\tau * \mathrm{up}(\boldsymbol{F}) + b_\tau \mathbb{1}, \tag{1}$$

where $b_\tau \mathbb{1}$ denotes a signal which takes value $b_\tau \in \mathbb{R}$ in every location.

Eq. (1) amounts to performing an exhaustive cross-modal search of the input linguistic expression over the input image. The goal is for activations of the response map (left-hand side of Eq. (1)) to correspond to segmentation targets.

2.3   Causal Intervention Module

To enable the model to learn more robust semantic features, we introduce a causal intervention module that acts directly on the representations learned by the CLIP encoders.

Causal View We introduce the formulation of causality for referring medical image segmentation by using a structural causal model. Specifically, it is constructed by examining causalities among four key components: expression $\boldsymbol{T}$, image $\boldsymbol{I}$, ground truth mask $\boldsymbol{Y}$, and confounder $\boldsymbol{C}$. As illustrated in Fig. 2, each direct link denotes a causal relationship between two nodes. The link $\boldsymbol{I}, \boldsymbol{T} \rightarrow \boldsymbol{Y}$ represents the desired causal effect from lesions in the image $\boldsymbol{I}$ to the segmentation mask $\boldsymbol{Y}$. We refer to a segmentation model as unbiased if it identifies $\boldsymbol{I}, \boldsymbol{T} \rightarrow \boldsymbol{Y}$ based solely on lesion information. The

Fig. 2. Structural causal model for referring medical image segmentation.

path $\boldsymbol{I}, \boldsymbol{T} \leftarrow \boldsymbol{C} \rightarrow \boldsymbol{Y}$ denotes that imaging conditions $\boldsymbol{C}$ determine pixel intensities in $\boldsymbol{I}$; for example, $\boldsymbol{C}$ encapsulates artifacts, acquisition protocols, and

scanner models which influence image characteristics. Since models inevitably utilize imaging cues for segmentation, the link $\boldsymbol{C} \to \boldsymbol{Y}$ also exists. In this structural causal model, we clearly see how $\boldsymbol{C}$ confounds $\boldsymbol{I}, \boldsymbol{T}$ and $\boldsymbol{Y}$ through the backdoor path $\boldsymbol{I}, \boldsymbol{T} \leftarrow \boldsymbol{C} \to \boldsymbol{Y}$. For instance, while no causal link exists between background pixels and the segmentation mask, the backdoor path creates a spurious correlation between them (through $\boldsymbol{C}$), yielding incorrect lesion predictions.

Adversarial Masking To be causally sufficient towards segmentation, we design an adversarial masking mechanism that iteratively detects features that contain relatively more causal information and separate them from confounding ones via a masker. Specifically, the masker produces a pair of complementary attention masks $\boldsymbol{M}$ and $\overline{\boldsymbol{M}}$ controlling every location in feature maps. $\boldsymbol{M}$ is for attending to features of the causal effect, while $\overline{\boldsymbol{M}}$ is for attending to the confounding effect. Note that these masks are generated during both training and inference, and they are learned and input-specific.

We denote the causal and confounding features as $\boldsymbol{F}_i^c$ and $\boldsymbol{F}_i^s$, respectively. They can be obtained by our module as follows:

$$
\begin{aligned}
\boldsymbol{M}_i &= \mathrm{sigmoid}(\mathcal{F}(\boldsymbol{F}_i)) \,, \\
\overline{\boldsymbol{M}}_i &= 1 - \boldsymbol{M}_i \,, \\
\boldsymbol{F}_i^c &= \boldsymbol{M}_i \odot \boldsymbol{F}_i \,, \\
\boldsymbol{F}_i^s &= \overline{\boldsymbol{M}}_i \odot \boldsymbol{F}_i \,,
\end{aligned}
\tag{2}
$$

where $\mathcal{F}$ is implemented by cascading a coordinate convolution and a convolution, and $\odot$ represents element-wise product. By doing so, we disentangle $\boldsymbol{F}_i^c$ and $\boldsymbol{F}_i^s$ from the input features.

Medical images often contain multi-scale lesions that necessitate model attention across scales. To obtain comprehensive representations, we fuse multi-scale visual features as follows. We first adopt the CARAFE operator [17] to transform causal and confounding features learned across stages into a unified scale. The scaled causal and confounding features are then concatenated separately. Finally, fused causal and confounding features, $\boldsymbol{F}^c$ and $\boldsymbol{F}^s$, are obtained via $1{\times}1$ convolutions over the concatenated features.

Adversarial Training We optimize the causal intervention module along with the CLIP encoders and decoder with a min-max loss. We use two separate yet identical decoders $\mathcal{D}_c$ and $\mathcal{D}_s$ to predict two segmentation masks from $\boldsymbol{F}^c$ and $\boldsymbol{F}^s$, respectively. The loss is formulated as follows:

$$
\mathcal{L} = \mathcal{L}_c - \lambda \mathcal{L}_s \,,
\tag{3}
$$

where

$$
\begin{aligned}
\mathcal{L}_c &= \ell_{ce}(\mathcal{D}_c(\boldsymbol{\tau}, \boldsymbol{F}^c), \boldsymbol{Y}) \,, \\
\mathcal{L}_s &= \ell_{ce}(\mathcal{D}_s(\boldsymbol{\tau}, \boldsymbol{F}^s), \boldsymbol{Y}) \,.
\end{aligned}
\tag{4}
$$

Table 1. Quantitative comparison with state-of-the-art methods on the QaTa-COV19 dataset.

Uni-modal			Multi-modal		
Methods	Dice	mIoU	Methods	Dice	mIoU
U-Net [21]	79.02	69.46	ConVIRT [28]	79.72	70.58
UNet++ [22]	79.62	70.25	TGANet [29]	79.87	70.75
AttUNet [23]	79.31	70.04	GLoRIA [30]	79.94	70.68
nnUNet [24]	80.42	70.81	ViLT [31]	79.63	70.12
TransUnet [25]	78.63	69.13	LAVT [32]	79.28	69.89
Swin-Unet [26]	78.07	68.34	LViT [3]	83.66	75.11
UCTransNet [27]	79.15	69.60	CausalCLIPSeg	**85.21**	**76.90**

Here, ℓ_{ce} denotes cross-entropy loss, Y is the ground truth mask of the input image, and λ is a coefficient balancing the two loss terms. We optimize the entire model by minimizing $\mathcal{L}_c$ and maximizing $\mathcal{L}_s$, i.e., minimizing $\mathcal{L}$.

3   Experiments

3.1   Dataset and Evaluation Metrics

We conduct experiments on the QaTa-COV19 dataset, which consists of 9,258 COVID-19 chest X-ray radiographs with manual annotations of COVID-19 lesions. This dataset was compiled by researchers from Qatar University and Tampere University. A very recent work [3] extends the QaTa-COV19 dataset with textual annotations, including whether both lungs are infected, the number of lesion areas, and the approximate location of the infected areas (e.g., "Bilateral pulmonary infection, four infected areas, upper lower left lung and upper middle lower right lung"). The authors of [3] partition the dataset into 5,716 training images, 1,429 validation images, and 2,113 test images.

We adopt two commonly used evaluation metrics [18], Dice coefficient (Dice) and mean intersection-over-union (mIoU), to validate the effectiveness of the proposed and competing methods.

3.2   Implementation Details

We implement our approach using PyTorch and run it on a single NVIDIA RTX 4090 GPU. Input images are resized to 224×224 pixels. The maximum input sentence length is set to 20 tokens. We initialize the encoders with CLIP and use a ResNet-101 backbone for the vision encoder. We train for up to 2,000 epochs with an early stopping criteria after 100 epochs without validation improvement. Optimization uses Adam [20] with a cosine learning rate scheduler, starting from 3e-5, and an adversarial training coefficient λ of 0.05. Additionally, we

Table 2. Ablation study of our proposed CausalCLIPSeg on the QaTa-COV19 dataset. We analyze the impact of different components of our approach.

	CLIP pre-trained	Causal intervention	Dice	mIoU
1	-	-	82.50	73.24
2	-	✓	83.61	74.86
3	✓	-	83.71	74.69
4	✓	✓	**85.21**	**76.90**

conduct multiple experimental runs and observe minimal fluctuations in results. For instance, the standard deviation of Dice for our model does not exceed 0.002.

3.3   Comparison with State-of-the-Art Methods

Table 1 shows results on the QaTa-COV19 dataset, comparing the proposed CausalCLIPSeg to competitors. Our method outperforms the state-of-the-art LViT [3] by 1.55% Dice and 1.79% mIoU. This demonstrates that our model effectively transfers the image-level knowledge of CLIP to pixel-wise segmentation tasks, having better cross-modal matching ability. Besides, our framework achieves remarkable performance gains of 5-7% than vision-only methods [21,22,23,24,25,26,27], highlighting the benefit of incorporating textual cues.

Fig. 3 provides a qualitative comparison between CausalCLIPSeg and two representative methods, U-Net [21] and LViT [3]. It can be seen that our segmentations have less mis-labeled regions and more precise object boundaries. This shows our approach learns more robust semantic features and reduces confounding biases.

3.4   Ablation Study

We present ablation experiments in Table 2 to evaluate the importance of several components. Note that adversarial masking is an indispensable component of our causal intervention module, and the adversarial loss is specifically designed to guide this module's learning. Therefore, we conduct an ablation study on the entire module. Specifically, we remove CLIP's pre-trained weights and the proposed causal intervention module from our framework to establish a baseline. As can be seen in the table, re-introducing both improves Dice by 2.71% and mIoU by 3.66%. Ablating CLIP's pre-trained weights results in a 1.60% decrease in Dice. In addition, removing the proposed causal intervention module also leads to a 1.50% drop. Together, these ablation studies demonstrate the significance of both the CLIP encoders and causal intervention modeling for achieving state-of-the-art referring medical image segmentation.

Further, to gain more insights into the advantage of each component in our proposed approach, we show some visualization results in Fig. 4. Notably, the
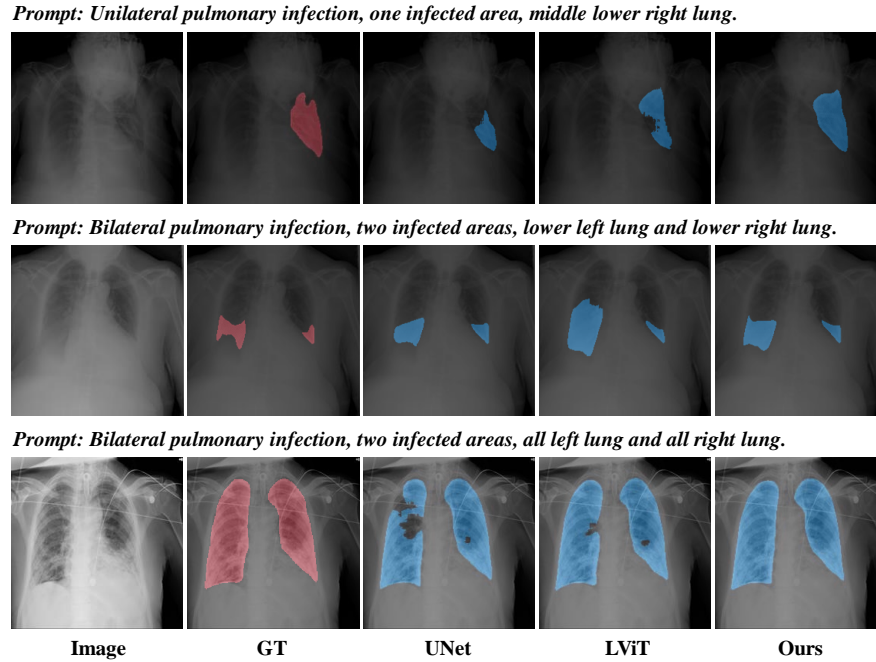
Prompt: Unilateral pulmonary infection, one infected area, middle lower right lung.

Prompt: Bilateral pulmonary infection, two infected areas, lower left lung and lower right lung.

Prompt: Bilateral pulmonary infection, two infected areas, all left lung and all right lung.

| Image | GT | UNet | LViT | Ours |

Fig. 3. Qualitative results on the QaTa-COV19 dataset.

causal intervention module can alleviate poor segmentations when lesion and lesion-free regions are similar in appearance.

4 Conclusion

In this work, we propose CausalCLIPSeg, an end-to-end framework for referring medical image segmentation that leverages CLIP and a tailored cross-modal decoding method to align image pixels and textual cues. To mitigate confounding bias, we introduce a causal intervention module that self-annotates confounders and excavates causal features for more robust segmentation. An adversarial min-max game further optimizes causal features while penalizing confounding ones. Through extensive experiments, CausalCLIPSeg demonstrates state-of-the-art performance for referring medical image segmentation, showing strong potential to translate insights from large-scale pre-trained vision-language models to specialized medical applications. Future work includes extending our approach to other multi-modal medical data analysis tasks and investigating approaches to further improve robustness against spurious correlations.

Prompt: Unilateral pulmonary infection, one infected area, middle left lung.

Prompt: Bilateral pulmonary infection, two infected areas, middle upper left lung and all right lung.

Prompt: Bilateral pulmonary infection, two infected areas, upper left lung and upper middle right lung.

| **Image** | **GT** | **Basline** | **w/o CLIP** | **w/o Causality** | **Ours** |
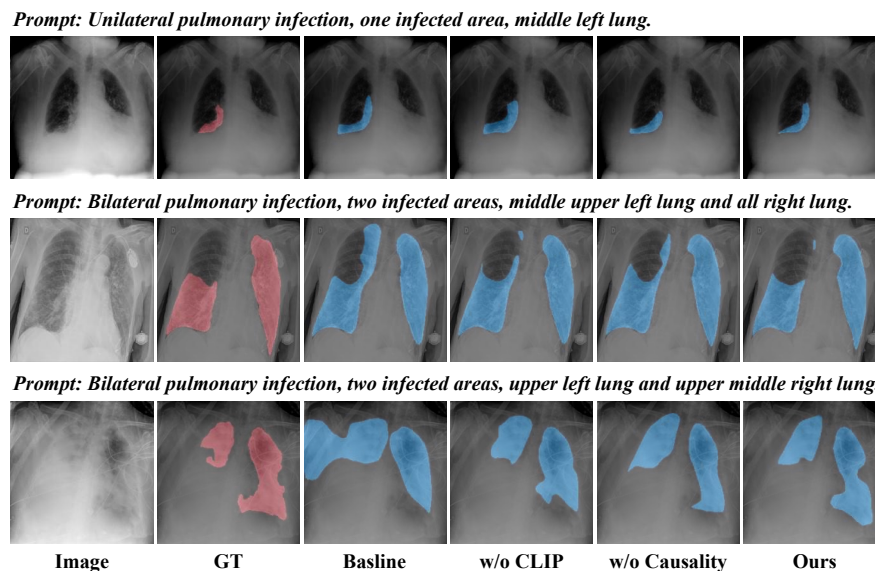
Fig. 4. Qualitative ablation study of CausalCLIPSeg.

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

References

1. Zhou, Y., Xie, L., Shen, W., Wang, Y., Fishman, E.K., Yuille, A.L.: A fixed-point model for pancreas segmentation in abdominal CT scans. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 693–701. (2017)
2. Gering, D.T., Nabavi, A., Kikinis, R., Grimson, W.E.L., Hata, N., Everett, P., Jolesz, F., Wells, W.M.: An integrated visualization system for surgical planning and guidance using image fusion and interventional imaging. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 809–819. (1999)
3. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: LViT: Language meets vision Transformer in medical image segmentation. IEEE Transactions on Medical Imaging **43**(1), 96–107 (2024)
4. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. (2021)

5. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.: Segment anything. In: IEEE/CVF International Conference on Computer Vision, pp. 4015–4026. (2023)

6. Ali, M., Khan, S.: CLIP-Decoder: Zeroshot multilabel classification using multi-modal CLIP aligned representations. In: IEEE/CVF International Conference on Computer Vision, pp. 4675–4679. (2023)

7. Conde, M.V., Turgutlu, K.: CLIP-Art: Contrastive pre-training for fine-grained art classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3956–3960. (2021)

8. Wang, M., Xing, J., Liu, Y.: ActionCLIP: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472 (2021)

9. Zhu, J., Jin, J., Yang, Z., Wu, X., Wang, X.: Learning CLIP guided visual-text fusion Transformer for video-based pedestrian attribute recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2625–2628. (2023)

10. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: CLIP4CLIP: An empirical study of CLIP for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)

11. Fang, H., Xiong, P., Xu, L., Chen, Y.: CLIP2Video: Mastering video-text retrieval via image CLIP. arXiv preprint arXiv:2106.11097 (2021)

12. Xie, Y., Liao, H., Zhang, D., Chen, F.: Uncertainty-aware cascade network for ultrasound image segmentation with ambiguous boundary. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 268–278. (2022)

13. Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., Qin, J.: Boundary-aware Transformers for skin lesion segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 206–216. (2021)

14. Wang, J., Yang, J., Zhou, Q., Wang, L.: Medical boundary diffusion model for skin lesion segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 427–436. (2023)

15. Pearl, J.: Causality. Cambridge University Press (2009)

16. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)

17. Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D.: CARAFE: Content-aware reassembly of features. In: IEEE/CVF International Conference on Computer Vision, pp. 3007–3016. (2019)

18. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**, 98–136 (2015)

19. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L.: PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp. 8024–8035. (2019)

20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

21. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 234–241. (2015)

22. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: A nested U-Net architecture for medical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention Workshops, pp. 3–11. (2018)

23. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B.: Attention U-Net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
24. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
25. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
26. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-UNet: UNet-like pure Transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218. (2022)
27. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with Transformer. In: AAAI Conference on Artificial Intelligence, pp. 2441–2449. (2022)
28. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference, pp. 2–25. (2022)
29. Tomar, N.K., Jha, D., Bagci, U., Ali, S.: TGANet: Text-guided attention for improved polyp segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 151–160. (2022)
30. Huang, S., Shen, L., Lungren, M.P., Yeung, S.: GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: IEEE/CVF International Conference on Computer Vision, pp. 3942–3951. (2021)
31. Kim, W., Son, B., Kim, I.: ViLT: Vision-and-language Transformer without convolution or region supervision. In: International Conference on Machine Learning, pp. 5583–5594. (2021)
32. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: LAVT: Language-aware vision Transformer for referring image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18155–18165. (2022)