Language-Enhanced Local-Global Aggregation Network for Multi-Organ Trauma Detection

Jianxun Yu[1], Qixin Hu[1], Meirui Jiang[1,✉], Yaning Wang[2], Chin Ting Wong[3], Jing Wang[2], Huimao Zhang[2,4,✉], and Qi Dou[1,✉]

[1] Dept. of Computer Science and Engineering, The Chinese University of Hong Kong
[2] Dept. of Radiology, The First Hospital of Jilin University
[3] Guangdong Provincial People's Hospital, Southern Medical University
[4] Jilin Provincial Key Laboratory of Medical Imaging & Big Data

Abstract. Abdominal trauma is one of the leading causes of death in the elderly population and increasingly poses a global challenge. However, interpreting CT scans for abdominal trauma is considerably challenging for deep learning models. Trauma may exist in various organs presenting different shapes and morphologies. In addition, a thorough comprehension of visual cues and various types of trauma is essential, demanding a high level of domain expertise. To address these issues, this paper introduces a language-enhanced local-global aggregation network that aims to fully utilize both global contextual information and local organ-specific information inherent in images for accurate trauma detection. Furthermore, the network is enhanced by text embedding from Large Language Models (LLM). This LLM-based text embedding possesses substantial medical knowledge, enabling the model to capture anatomical relationships of intra-organ and intra-trauma connections. We have conducted experiments on one public dataset of RSNA Abdominal Trauma Detection (ATD) and one in-house dataset. Compared with existing state-of-the-art methods, the F1-score of organ-level trauma detection improves from 51.4% to 62.5% when evaluated on the public dataset and from 61.9% to 65.2% on the private cohort, demonstrating the efficacy of our proposed approach for multi-organ trauma detection. Code is available at: https://github.com/med-air/TraumaDet

Keywords: Multi-organ trauma detection · Language-enhanced model

1 Introduction

Abdominal trauma increasingly affects the elderly population with annual traumatic injury-related deaths exceeding five million worldwide [1]. In the quest for prompt and precise detection of such trauma, deep learning-based methods have ascended to significant importance [2,3]. However, the complexities of interpreting Computed Tomography (CT) scans for abdominal trauma pose considerable challenges for deep learning models. First, the heterogeneity in scale and morphology of abdominal organs, coupled with the often subtle manifestations of hemorrhage, presents a formidable challenge. Second, achieving a precise understanding of visual cues and the diversity of trauma types necessitates a high level of expertise. Such complexities require deep learning models not only to recognize diverse patterns but also to interpret complex clinical signs like expert radiologists [4,5,6]. How to fully exploit the imaging context with varying scales and morphologies, as well as explore the connection between image and clinical signs, remains an open question.

To overcome the heterogeneity in scale and morphology for precise disease diagnosis, various deep learning-based solutions have been increasingly studied. One widely adopted line has been the utilization of three-dimensional spatial information to enhance diagnostics [7,8,9]. For example, Ma et al. [10] combine the convolutional neural network and recurrent neural network to jointly explore sequential information along slices. Hatamizadeh et al. [9] have harnessed the potential of 3D transformer-based architecture to learn 3D representations and capture long-range dependencies between voxels. Despite these advancements, there remains a tendency in these methodologies to prioritize global information at the expense of local subtleties, which are imperative for the detection of organ trauma. Recent studies by Huang et al.[3] and Cheng et al.[2] have attempted to recalibrate the focus towards a more granular examination of scale heterogeneity. Their proposed 3D architectures mix visual features at the organ-specific level, yielding improved results. Nevertheless, these approaches tend to isolate the examination of each organ without sufficiently considering the synergy between local and global information streams or the inter-organ relationships, which are crucial for a holistic understanding of abdominal trauma. Moreover, there exists an imperative to integrate the imaging context with clinical indicators to achieve a comprehensive and nuanced detection of trauma.

In this regard, our insight focuses on two pivotal aspects. The first lies in acquiring imaging context at both local and global scales to tackle the challenges presented by heterogeneity in scale and morphology. The second involves incorporating clinical expertise to guide the detection of trauma. Drawing inspiration from the training paradigm of contrastive language-image pretraining (CLIP) [11], our proposed strategy encompasses interpreting clinical signs through textual input, which subsequently interacts with the image-derived features. Then, the critical question is how to ensure the effective integration of these modalities. The alignment of local-global feature representations or feature-text correlations may not be inherently congruent.

In this paper, we propose a novel language-enhanced local-global aggregation network. Our network considers both global and local visual information and is enhanced by text embeddings obtained from the LLM. In particular, when provided with a CT scan, we initially extract global and local organ-wise image features from a pre-trained vision encoder [12]. Subsequently, we employ a dual attention mechanism, wherein both global and local features serve as keys and values for each other. This architectural design enables the model to capture visual representations encompassing both semantics and details. We utilize LLM text embeddings to further integrate intrinsic anatomical cues into visual representations. This LLM text embedding possesses substantial medical knowledge, enabling the model to capture anatomical relationships of intra-organ and intra-trauma connections. In this study, the text embeddings are used in two ways: first, by enhancing the local vision features through organ-wise prompt, and second, by supplying guidance for the entire network through the trauma-category prompt. Our proposed method is evaluated on one public dataset of contrast-enhanced CT scans and one private dataset of non-enhanced CT scans. The effectiveness of our method has been demonstrated with significant performance improvements and comprehensive analytical studies. Specifically, compared with previous state-of-the-art methods, our approach increases the organ-level F1 score by 11.1% on public data and 3.3% on private data.

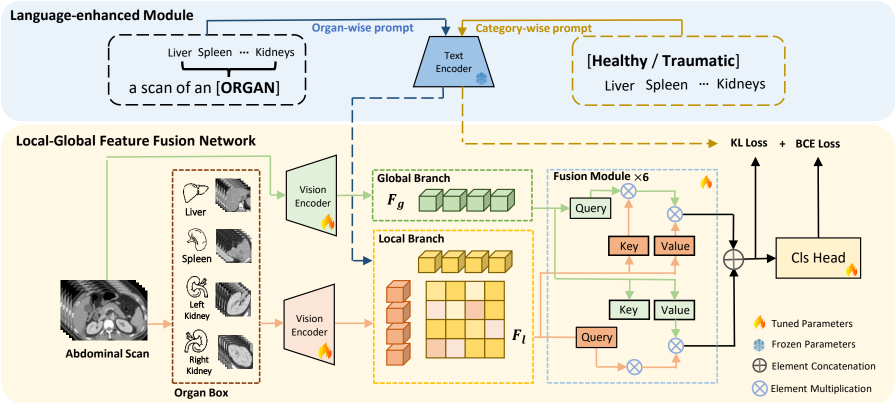


Fig. 1. The overview of our proposed method, which consists of the local-global feature fusion network and the language-enhanced module.

2 Methodology

2.1 Preliminaries

We aim to improve abdominal trauma detection, which has historically been viewed as a problem of directly extracting features from a whole CT scan. Combining with the actual clinical diagnosis knowledge, we refine the information derived from the whole CT scan and complement it with additional anatomical and medical information, which can be formulated as Eq. (1). In this equation, S is the CT scan, S_g is the global semantic information, N is the number of task-related organ, R_i is the detailed information of the i-th organ, ω_l is the intrinsic anatomical relationships among organs, and ω_t is the intra-trauma information:

$$S \leftrightarrow \{S_g, \{R_1, R_2, \ldots, R_i\}\} \cup \{\omega_l, \omega_t\}, i = 1, 2, \ldots, N. \tag{1}$$

The challenge lies in extracting and integrating these multiple pieces of information, which are under different scales and modalities. Based on the above thinking and problem-setting, we consider to use a language-enhanced local-global aggregation network to address the feature fusion problem across different scales and modalities for this specific task.

An overview of our proposed method is shown in Fig. 1. First, we introduce the local-global feature fusion network, which separately extracts global and local features. These features are then integrated to enable a comprehensive vision representation of the input. Subsequently, we present a language-enhanced module that incorporates additional information from both organ-wise and category-wise prompts. Finally, we describe the loss function of the entire framework.

2.2 Local-Global Feature Fusion Network

The network architecture comprises a global branch responsible for extracting global semantic features, a local branch dedicated to capturing local detailed features, and a fusion module that facilitates feature aggregation. As the input of the local branch, organ R_n is obtained from the CT scan S, in which ${R_n}^{C\times h\times w\times d} = \psi_c(n, S^{C\times H\times W\times D})$, $\psi_c(\cdot)$ is a trained segmentation model used to obtain segmentation maps of organs, n is the organ class, C, H, W and D is indicated the channel, height, width, and depth of the CT scan, h, w, d is the spatial size of the bounding box, named the "organ box". In this process, the organ box is cropped from the sparse segmentation maps acquired by the segmentation model. The segmentation model used here can be any currently common segmentation model, such as [13,14].

In the local branch of the network, the local vision encoder component consists of individual encoders with an equal number of organs being considered. These encoders are specifically designed to extract corresponding feature maps for each organ. Therefore, the output of the local branch is acquired by $f_l = \psi_{\mathrm{lv1}}(R_1) \oplus \psi_{\mathrm{lv2}}(R_2)\oplus, \ldots, \oplus\psi_{\mathrm{lv}n}(R_n)$, where n is the number of organs, $\psi_{\mathrm{lv}n}$ is the individual encoders for n-th organ, and $\oplus$ is the element concatenate operation. In the global branch of the network, the global semantic features F_g are encoded directly from the original CT scan S by the global vision encoder. In addition, the feature aggregation module has a transformer-based attention mechanism, which is used to further enhance the connections between the local detailed vision feature and the global semantic vision feature. In practice, the module consists of a multi-head self-attention layer, which is applied six times to the global and local branches. In this procedure, the output of the global branch f_g, and the local branch f_l are reshaped as the query separately, and another as key, value. This process can be formulated as the following:

$$f_{\mathrm{fused}}(f_l, f_g) = \phi(\frac{{f_l}^Q {f_g}^{K^T}}{\sqrt{d}}){f_g}^V \oplus \phi(\frac{{f_g}^Q {f_l}^{K^T}}{\sqrt{d}}){f_l}^V, \qquad (2)$$

where $\phi(\cdot)$ is the softmax function, Q, K, and V refer to the mentioned 'query', 'key', and 'value' components reshaped from f_l and f_g, and d is the dimension of the key, and the output f_{fused} can be seen as an adapted vision representation of the anatomical target in the CT scan.

2.3 Language-Enhanced Module

This module plays a crucial role in leading into semantic information at different scales through the organ-wise prompt and the category-wise prompt. It incorporates a text encoder empowered by the CLIP [11,15] to generate text embedding using a medical prompting template, which can leverage the intrinsic semantic relationships between organs and different types of trauma. The organ-wise prompt is applied during both the training and inference procedures, and the category-wise prompt is exclusively involved in the training process to guide the predictions. During the training stage, the parameters of the text encoder are frozen and used solely for encoding the organ-wise and category-wise prompts. During the inference process, only organ-wise prompts are utilized, as they are generated from organ names in the local branch and are unrelated to labels.

For the organ-wise prompt, let f_k be the text embedding of the k-th organs, produced by the pre-trained text encoder and a medical prompt. In this method, the medical prompt applied the template "a computerized tomography of a [$ORGAN$]", which has been verified as an efficient template [12], where [$ORGAN$] is a concrete task-related organ name, e.g., "liver, spleen, etc.". We combine the text embedding f_k and the output feature f_l in the local branch to implement the intrinsic semantic information and acquire the output feature of the local branch f_l by $f_l \odot f_k$, where $\odot$ is the element-wise multiplication.

For the category-wise prompt, it's generated according to the label and has a similar template to the organ-wise prompt. Different prompt types are tested in experiments, as shown in Fig. 2(b)(c), and the one with the best performance is ultimately chosen. In the chosen type, there are two possible prompts for each organ when the presence or absence of trauma is considered. If the organ label is 0, which means the organ doesn't have trauma, the prompt will be "a computerized tomography of a [$Healthy$][$Organ$]". If the organ label is 1, which represents the organ has trauma, the prompt will be "a computerized tomography of a [$Traumatic$][$Organ$]". In the training procedure, the category-wise prompt p_t, where $p_t = [p_{\text{liver}}, p_{\text{spleen}}, p_{\text{kidneys}}]$, is selected according to the label, and the text embedding f_t of the selected category-wise prompt is produced by the pre-trained text encoder. Then, the text embedding f_t is used to calculate the Kullback-Leibler (KL) loss L_{KL} with the output of the local-global feature fusion network f_{fused} by the Eq. (3), where $\mathcal{X}$ is the set of all possible values:

$$L_{\text{KL}}(f_{\text{fused}}, f_t) = \sum\nolimits_{x \in \mathcal{X}} f_{\text{fused}}(x) \log \left(\frac{f_{\text{fused}}(x)}{f_t(x)} \right). \tag{3}$$

By performing this calculation, the category-wise semantic features contained in the text embedding are distilled in the training procedure, enabling them to exert influence and enhance the prediction in the inference procedure. Finally, the entire framework is trained as follows:

$$L = L_{\text{BCE}}(f, f_{\text{fused}}(f_l, f_g)) + \alpha * L_{\text{KL}}(f_{\text{fused}}(f_l, f_g), f_t), \tag{4}$$

where BCE denotes the Binary Cross-Entropy loss, f is the actual label, and the α is the trade-off parameter weighting the importance of each component.

Table 1. Performance comparison on the private dataset with non-enhanced CT.

Methods	case acc.(%)	case prec.(%)	case F1(%)	organ acc.(%)	organ prec.(%)	organ F1(%)
3D-ViT [16]	79.3±1.5	84.9±3.2	75.6±2.4	80.6±1.2	43.8±4.1	40.8±2.4
TTADC [10]	79.3±3.0	92.9±2.0	73.2±5.1	87.9±0.5	73.8±3.9	57.6±3.1
Ham. et al. [17]	83.3±0.9	90.9±1.2	80.1±1.1	88.1±0.1	69.9±0.6	61.9±0.3
Huang et al.[3]	75.7±1.7	80.6±3.0	71.2±2.1	87.6±0.5	68.6±2.9	60.0±1.0
CBAM [2]	80.3±1.8	92.5±1.5	75.2±2.7	88.3±0.4	70.8±2.8	59.6±2.8
Ours	**84.2±1.3**	**94.3±1.7**	**80.8±1.7**	**89.1±0.3**	**73.8±2.7**	**65.2±0.6**

3 Experiment

Datasets. We use two datasets to evaluate our methods: one public RSNA ATD dataset and one private dataset. The publicly available RSNA ATD dataset has over 4000 *contrast-enhanced* abdominal CT scans with 200 detailed per-voxel segmentation labels for the liver, spleen, and kidney. The private dataset was provided by the First Hospital of Jilin University, which has 600 *non-enhanced* abdominal CT scans, all with per-voxel segmentation labels for abdominal organs. All these datasets provide organ-level trauma injury labels, and the resolution of CT scans is 512 × 512. Annotation of the private dataset for segmentation and trauma injury labels was initially conducted by ten physicians, with subsequent verification by two senior radiologists to ascertain accuracy and consistency. We analyzed their label distribution and presented the results in Fig. 2(a). The blue bars represent the public dataset, while the orange bars represent the private dataset. As demonstrated, there exists a misalignment in the label distribution between the two datasets. This partly explains the performance shift from the public to the private dataset, which will be discussed in the next sections.

Experimental Setting. To assess the effectiveness of our method, both the public RSNA ATD dataset and the private dataset are randomly divided into the training set, validation set, and test set. Specifically, a ratio of 4:1:1 is applied. For the public dataset, 2640 CT scans are utilized for training the model, 680 CT scans for validation, and 680 CT scans for testing. As for the private dataset, the number goes to 400, 100, and 100, respectively.

Evaluation Metrics. We utilize two levels of metrics to evaluate the performance of all models: case-level metrics to assess the model's diagnostic outcome and case-level metrics to evaluate the model's robustness across different organs. We report the results using three metrics: accuracy, precision, and F1-score.

Implementation Details. Our code is implemented in Python with MONAI. Input images are clipped with the window range of [-175,250] and linearly normalized to [0,1]. Isotropic spacing is adopted to re-slice each image to the voxel size of 1.5×1.5×1.5 mm^3. Our method necessitates organ boxes to extract local organ-wise features. Given that abdominal organ segmentation in CT scans is largely resolved, we employ a cutting-edge state-of-the-art approach, e.g. TransUnet [13], to obtain the organ boxes. To accommodate the input size of the pre-trained vision encoder [12], we cropped the local organ images with a fixed-

Table 2. Performance comparison on the public dataset with contrast-enhanced CT.

Methods	case
	acc.(%)
		case
		prec.(%)
			case
			F1(%)
				organ
				acc.(%)
					organ
					prec.(%)
						organ
						F1(%)

3D-ViT [16]	70.9±1.4	44.7±2.6	43.2±1.3	86.9±0.4	32.1±3.2	30.3±1.3
TTADC [10]	80.4±0.3	66.5±1.2	47.7±1.1	90.7±0.1	53.7±1.4	35.9±1.6
Ham. et al. [17]	76.2±3.4	44.4±2.3	52.6±2.5	88.6±0.4	45.3±0.5	45.4±2.8
Huang et al.[3]	84.1±0.7	78.5±5.2	58.8±0.6	92.5±0.3	70.6±4.8	50.3±0.6
CBAM [2]	83.9±0.3	78.6±3.1	57.4±3.1	92.8±0.3	74.7±6.0	51.4±1.5
Ours	**87.7±0.5**	**84.7±2.0**	**66.8±1.1**	**94.1±0.2**	**77.6±1.3**	**62.5±2.1**

sized 96×96×96. The global vision encoder adopts the 3D ResNet-50 model [18]. The ViT-B-32 model with CLIP, which has proven effectiveness and wide applicability [12,19], is used as the text encoder to align text and image features more efficiently. For different types of category-wise prompts, as shown in Fig. 2(c), the GPT-3.5-turbo is used to generate fine-grained descriptions. Our work focuses on 3D medical volumes, whereas current common models primarily address 2D image captioning [20,21], making it challenging to generalize to 3D data. Therefore, we generate clinical fine-grained descriptions based on labels by LLM, without vision inputs. We trained all models using the AdamW optimizer and a warm-up cosine scheduler for 20 epochs. All models are trained for 400 epochs with a batch size of 4. We trained the model with a default initial learning rate of $5e-4$, a momentum of 0.9, and a weight decay of $1e-5$ on a single GPU.

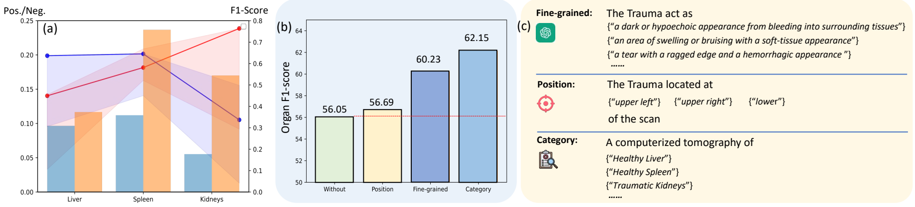
3.1 Comparison with State-of-the-Arts Methods

Public Dataset. We compare our language-enhance local-global aggregation network with several cutting-edge methods. To enhance the performance of these baseline methods, we focus solely on the abdominal region and crop out the background. Table 2 provides a comprehensive comparison with all baseline methods, demonstrating great improvements in our method compared to other state-of-the-art approaches. Specifically, at the case level, the state-of-the-art baseline method only achieves an accuracy of 84.1%, a precision of 78.6%, and an F1-score of 58.8%. In contrast, our method demonstrates superior performance with an accuracy of 87.7%, a precision of 84.7%, and an F1-score of 66.8%, surpassing all metrics significantly. The improvements are even more pronounced at the organ level, with the F1-score increasing from 51.4% to 62.5%. These results highlight the efficacy of our method in multi-organ trauma detection.

Private Dataset. While the efficacy of our method has been proven in the publicly available dataset, the results in the private dataset exhibit more interesting properties, shown in Table 1. Firstly, without question, our method achieves the best performance across all metrics. Secondly, if we compare Table 1 and Table 2, we observe a significant performance shift in the results of other baseline methods. For example, TTADC [10] achieves an organ-level F1-score of 57.6% on the private dataset but only 35.9% on the public dataset. In contrast, our method maintains good performance on both datasets. The underlying reason,

Table 3. Ablation analysis of our method on the public CT dataset.

global branch	local branch	language guidance	case acc.(%)	case prec.(%)	case F1(%)	organ acc.(%)	organ prec.(%)	organ F1(%)
✓			73.24	46.03	54.73	88.48	42.59	48.80
	✓		86.03	80.36	65.46	92.89	69.47	55.66
✓	✓		86.03	82.70	64.42	93.24	74.58	56.05
✓		✓	84.41	75.68	61.31	92.84	70.16	54.38
	✓	✓	86.18	77.17	67.59	93.04	68.49	58.48
✓	✓	✓	**87.35**	**84.70**	**68.61**	**93.97**	**78.30**	**62.15**



The Trauma act as
{"*a dark or hypoechoic appearance from bleeding into surrounding tissues*"}
{"*an area of swelling or bruising with a soft-tissue appearance*"}
{"*a tear with a ragged edge and a hemorrhagic appearance* "}
......

The Trauma located at
{"*upper left*"} {"*upper right*"} {"*lower*"}
of the scan

A computerized tomography of
{"*Healthy Liver*"}
{"*Healthy Spleen*"}
{"*Traumatic Kidneys*"}
......

Fig. 2. (a). Data distribution and organ-wise performance. (b). Performance of using different category-wise prompts. (c). Examples of category-wise prompts.

as indicated by the dataset part, is partially attributed to the more pronounced issue of label imbalance within the public dataset. This finding suggests that our method exhibits greater robustness in handling variations in data distribution.

3.2 Analytical Study

To demonstrate the importance of each module of our network, as well as the effect of language prompts used in this study, we have performed comprehensive analytical studies. The results are shown in Table 3 and Fig. 2(c). First of all, in terms of module design, we find that each module can offer unique information, as can be observed from the Table 3. All the modules can provide useful information, with the local branch and language guidance being particularly significant. Second, regarding the effects of different language prompts, we report the performance in Fig. 2 (b) and (c). From the results, it can be observed that a fine-grained description is not always good. In our scenario, category-wise prompts could provide more precise information than fine-grained descriptions, which could serve as reliable guidance for the network.

4 Conclusion

In this paper, we have proposed a language-enhanced local-global aggregation network for multi-organ trauma detection. We have evaluated the performance of our method on two datasets and demonstrated significant improvements compared to other state-of-the-art methods, both in performance and robustness. This study is among the first to utilize LLM text embedding in multi-organ trauma detection. In the future, one research direction is to eliminate the need for segmentation models and transition our method to weakly-supervised or unsupervised methods to obtain organ boxes. Furthermore, it is promising to extend our work to detect abnormalities across various organ types.

Acknowledgement. This work was supported in part by the Shenzhen Portion of Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under HZQB-KCZYB-20200089, in part by National Natural Science Foundation of China Project No. 12226003 and No. U22A20351, in part by a Hong Kong Research Grants Council Project No. T45-401/22-N, in part by National Key R&D Program of China Project 2022ZD0161100, in part by Jilin Provincial Key Laboratory of Medical Imaging & Big Data Project No. 20200601003JC, and in part by Jilin Provincial International Joint Research Center for Medical Artificial Intelligence Precision Diagnosis and Treatment Project No. 020210504008GH.

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

References

1. Mohamed Tarchouli, Mohamed Elabsi, Noureddine Njoumi, Mohamed Essarghini, Mahjoub Echarrab, et al. Liver trauma: What current management? *Hepatobiliary & Pancreatic Diseases International*, 17(1):39–44, 2018.
2. Chi-Tung Cheng, Hou-Hsien Lin, Chih-Po Hsu, Huan-Wu Chen, Jen-Fu Huang, Chi-Hsun Hsieh, Chih-Yuan Fu, I-Fang Chung, and Chien-Hung Liao. Deep learning for automated detection and localization of traumatic abdominal solid organ injuries on ct scans. *Journal of Imaging Informatics in Medicine*, pages 1–11, 2024.
3. Shungen Huang, Zhiyong Zhou, Xusheng Qian, Dashuang Li, et al. Automated quantitative assessment of pediatric blunt hepatic trauma by deep learning-based ct volumetry. *European Journal of Medical Research*, 27(1):305, 2022.
4. Wenkai Yang, Juanjuan Zhao, Yan Qiang, Xiaotang Yang, Yunyun Dong, Qianqian Du, Guohua Shi, and Muhammad Bilal Zia. Dscgans: Integrate domain knowledge in training dual-path semi-supervised conditional generative adversarial networks and s3vm for ultrasonography thyroid nodules classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 558–566. Springer, 2019.
5. Yutong Xie, Yong Xia, Jianpeng Zhang, Yang Song, Dagan Feng, Michael Fulham, and Weidong Cai. Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct. *IEEE transactions on medical imaging*, 38(4):991–1004, 2018.
6. Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21372–21383, 2023.
7. Nhan T Nguyen, Dat Q Tran, Nghia T Nguyen, et al. A cnn-lstm architecture for detection of intracranial hemorrhage on ct scans. *medRxiv*, pages 2020–04, 2020.
8. Xiyue Wang, Tao Shen, Sen Yang, Jun Lan, Yanming Xu, Minghui Wang, Jing Zhang, and Xiao Han. A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head ct scans. *NeuroImage: Clinical*, 32:102785, 2021.
9. Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.
10. Wenao Ma, Cheng Chen, Shuang Zheng, Jing Qin, Huimao Zhang, and Qi Dou. Test-time adaptation with calibration of medical image classification nets for label distribution shift. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 313–323. Springer, 2022.
11. 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.
12. Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, et al. 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.

13. Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
14. Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
15. Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
16. Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 574–584, January 2022.
17. Mohammad Hamghalam, Robert Moreland, David Gomez, Amber Simpson, Hui Ming Lin, Ali Babaei Jandaghi, Monica Tafur, Paraskevi A Vlachou, Matthew Wu, Michael Brassil, et al. Machine learning detection and characterization of splenic injuries on abdominal computed tomography. *Canadian Association of Radiologists Journal*, page 08465371231221052, 2024.
18. Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018.
19. Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation. In *The Twelfth International Conference on Learning Representations*, volume 1, 2024.
20. Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
21. Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.