Organ Detection Based on Vision-Language Model For Abdominal CT Images

Jun-Hong Ou[1], Bo-Xian Wang[1], Yu-Hong Zheng[1], Sufal K. Chhabra[1], Guo-Shiang Lin[1*], Shen-Lei Yan[2], Chen-Kuo Chiang[3]

[*1] National Chin-Yi University of Technology, Taiwan

[2]Chang Bing Show-Chwan Memorial Hospital, Taiwan

[3]National Chung Cheng University, Taiwan

E-mail: gslin@ncut.edu.tw

***Abstract*— This paper proposes an organ detection method called OGTransVG based on vision-language model. The proposed network developed based on TransVG comprises several components: an image encoder, a text encoder, an image-text feature fusion module, and two predictors. The image encoder and text encoder extract critical features from the image and text, respectively. The image-text feature fusion module integrates multi-model features to achieve useful feature representation. One predictor is used to detect objects and the other indicates whether objects exist or not. The proposed OGTransVG model is trained by some medical images. Here we perform the subjective and objective evaluation for performance analysis. Experimental results demonstrate that the proposed OGTransVG network can not only detect organs well but also successfully deal with the situation: no-target expressions.**

I. INTRODUCTION

Liver Cirrhosis is widely prevalent and is the 11th most common cause of death worldwide [1]. Cirrhosis is a chronic condition resulting from inflammation and fibrosis of the liver. The most common causes of cirrhosis worldwide are alcohol-related liver disease, non-alcoholic fatty liver disease, and chronic viral hepatitis B and C [2,3]. In the clinical setting, patients with decompensated cirrhosis may present with jaundice, ascites, or hepatic encephalopathy [4]. The Child-Turcotte-Pugh score uses serum albumin, bilirubin, prothrombin time, ascites and hepatic encephalopathy to classify cirrhotic patients into classes A, B, and C [5]. Ultrasonography is routinely used during the evaluation of cirrhosis. In decompensated cirrhosis, the liver may appear small, nodular, and increased echogenicity [6]. Although not routinely used in the diagnosis of cirrhosis, computed tomography (CT) is a commonly used imaging tool in patients presenting with focal liver masses. That is why a diagnostic system based on CT images is important to build a computer-aided diagnosis (CAD) system.

Recent studies on abdominal CT have explored organ detection mainly via single-modality approaches, such as 3D U-Net and related volumetric CNNs [15], and unified frameworks such as nnU-Net that self-configure for new segmentation tasks [16]. More recently, large "promptable" segmentation and open-vocabulary detection methods (e.g., Segment Anything (SAM) [17] and GroundingDINO [18]) have been adapted to medical imaging, enabling flexible localization from sparse prompts or textual queries. These methods typically require large-scale mask/box supervision or image–text pairs and do not specifically target text-driven organ detection in abdominal CT with explicit no-target handling. Our OGTransVG fills this gap by adapting a vision–language model to abdominal CT and adding an explicit existence classifier.

Figure 1 shows two CT images. Fig. 1(a) and Fig. 1(b) are normal and cirrhosis cases, respectively. As we can see in Fig. 1, the position, the size, and shape information of the liver is different. Traditional CAD systems in healthcare commonly rely on single-modality image data paired with labels to train and develop predictive models [9-10]. On the other hand, there are some organs in CT images. This means that organ detection is an important issue to analyze CT images.

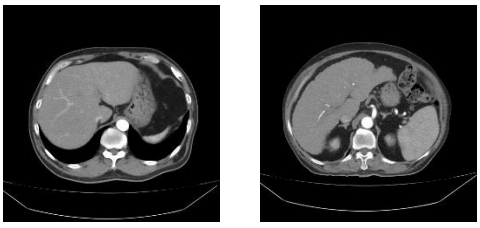

Figure 1 Two abdominal CT images: (a) normal and (b) cirrhosis.

While effective to some extent, these models are inherently limited in their ability to utilize comprehensive clinical information, such as patient history, laboratory results, or descriptive textual data provided by healthcare professionals. Actually, it is expected that integrating image and textual data could offer a promising approach to achieving precise and context-aware detection. Textual descriptions can provide critical contextual cues to resolve ambiguities, particularly for visually similar Liver Cirrhosis levels. On the other hand, the goal of visual grounding (VG) is to detect the most relevant object or region in an image based on a natural language query. Among existing methods [7][14], TransVG [7] is a transformer-based visual grounding (VG) framework that integrates text embeddings, image features, and a regression token as input. It leverages the self-attention mechanism of the

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

transformer to achieve cross-modal feature alignment and reformulates target object detection into an end-to-end regression prediction task. Therefore, this study extends an existing vision-language model [7] to enable the proposed method to not only detect organs but also determine whether the target organ exists in an image or not.

II. Proposed Method

As mentioned in Section I, the proposed method is a visual grounding framework that was built based on the transformer architecture, TransVG [7]. The existing model effectively integrates textual and visual features through a transformer encoder, enabling precise localization of corresponding regions in an image based on language descriptions. To extend its functionality to handle the "No-Target" scenario, we incorporated a classification head into the original architecture to determine whether a target exists in the image. Figure 1 illustrates the proposed network called OGTransVG. In addition, our fusion module adopts Inter-Modality Cross Attention (ICMA) to have interaction between two modalities. Specifically, ICMA first applies self-attention over the visual tokens to capture intra-relationships, followed by cross-attention where the visual tokens query the text embeddings. This design follows the implementation shown in our ICMAFusion block [14], which integrates projected textual embeddings with visual features using self-attention, cross-attention, and feed-forward layers with residual connections. This enables effective alignment of image–text features while also handling padding tokens via attention masks.

As shown in Fig. 2, the visual branch adopts ResNet-50 as the backbone, augmented with 6 stacked transformer encoder layers, initialized with pre-trained weights from the DETR model. Given an input image $I_0 \in \mathbb{R}^{3\times H_0\times W_0}$, the visual branch processes the image to produce a 1D feature map $F_v \in \mathbb{R}^{D_v\times HW}$, where $H = H_0/32$ and $W = W_0/32$. For the language branch, 12 transformer encoder layers initialized with the pre-trained BERT model are utilized. To process the input expression or phrase, the text is first tokenized into a sequence of N tokens using BERT' tokenizer. Specifically, two tokens [CLS] and [SEP] are appended to the beginning and end positions of the tokenized input sequence. This sequence is then fed into the language branch, resulting in an output representation $F_t \in \mathbb{R}^{D_t\times N}$, where N denotes the length of the tokenized input phrase.

As shown in Fig. 2, the sequence x_0 is augmented with learnable positional embeddings before being fed into the visual-language Transformer. These positional embeddings are added to x_0 to encode spatial relationships for the visual modality and sequential relationships for the language modality. This allows the model to better understand the modal relationships within the input sequence. After we apply the visual-language Transformer comprises six Transformer encoder layers, integrating visual and textual features to predict the corresponding target region. These Transformer encoder layers are designed to capture both intra-modal (the self-relations within each modality) and inter-modal (the interactions between visual and textual features) relationships. Furthermore, to enable the model to recognize and handle no-target scenarios, we added a classifier, as indicated by the blue dashed box in Figure 1.

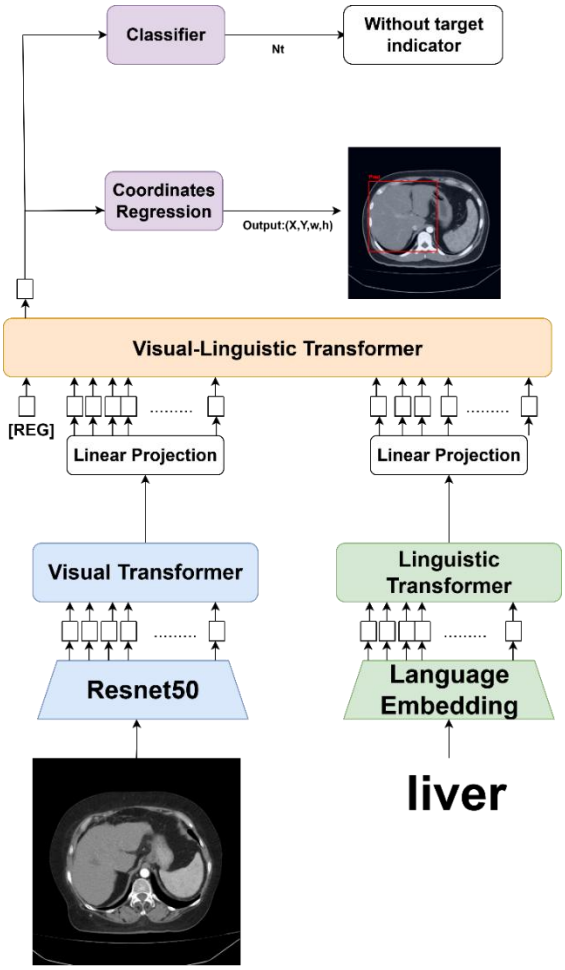

Figure 2 Illustration of organ detection

For organ detection, the proposed model's output consists of two components: a bounding box (BB) predictor and a classifier. The BB predictor and the classifier are both implemented as MLP. In Fig. 2, the [REG] token serves as the input to both components. Since [REG] is responsible for predicting the bounding box, it inherently encodes information about the presence and location of the target. In addition, to determine whether a target exists in the image, there are two cases. The target-present case and the target-absent case represent whether the target is respectively present or absent in an image. Leveraging this property, we introduce a classifier that processes the same [REG] token to determine whether a valid target exists.

The BB predictor directly outputs the center coordinates and size of the referred bounding box, $\tilde{b} = (x, y, h, w)$. Meanwhile, the classifier outputs a two-dimensional probability vector N_t, indicating whether the target is present (w/ target) or absent (w/o target) in the image.

To supervised the training, the loss function has three losses, smooth L1 loss $L_{smooth-l1}$, a generalized IOU loss L_{GIoU}, and cross entropy loss L_{CE} [14]. Then the total loss function is expressed below:

$$L_{total} = L_{smooth-l1} + L_{GIoU} + L_{CE}, \quad (1)$$

To train with a limited dataset, transfer learning [12] is adopted. During the training phase, the image encoder and the text encoder are initialized based on a pre-trained TransVG model trained on a large-scale dataset [11]. Then the proposed whole network was trained for the target task.

III. Experimental results

To build an organ detection model, some CT images are collected. We then created annotations and textual descriptions corresponding to the images and applied data augmentation to enhance the diversity of the dataset. Each image was annotated with corresponding labels and textual descriptions, and data augmentation techniques were applied to improve dataset diversity.

For model training, these organs and the corresponding expressions were distributed across anatomical regions, including liver, stomach, spleen, spine, and aorta. The dataset comprised 827, 103, and 104 samples for training, validation, and testing, respectively.

3.1 Performance indexes

To evaluate the accuracy of the model's predicted bounding boxes, mIoU and Acc@δ were used as evaluation metrics. The measurement, mIoU, measures the average overlap between the predicted bounding boxes and the ground truth across all samples. The definition of mIoU is described as follows:

$$IoU = \frac{Intersection\ \ Area}{Union\ \ Area}, \quad (2)$$

$$mIoU = \frac{1}{N}\sum_{i=1}^{N} IoU_i, \quad (3)$$

The measurement, Acc@δ, evaluates the proportion of predicted bounding boxes that achieve an IoU of at least the threshold δ with the ground truth bounding boxes. The definition of Acc@δ is in Eq. (4). This metric provides a binary measure of whether the model's predictions meet the predefined threshold for correctness, offering a straightforward evaluation of the model's localization capability:

$$Acc@\delta = \frac{Number\ \ of\ \ Correct\ \ Predictions\ \ (IoU\ \geq \delta)}{Total\ \ Number\ \ of\ \ Predctions}, \quad (4)$$

N-acc and T-acc are common metrics used to evaluate the performance of a model in distinguishing between target-absent and target-present samples [13][14]. N-acc measures the model's ability to correctly identify target-absent samples. A True Negative (TN) is defined as the correct prediction of no target without misidentifying any foreground, while a False Positive (FP) occurs when the model incorrectly predicts the presence of a target. Conversely, T-acc evaluates the model's ability to accurately identify target-present samples. A True Positive (TP) indicates a correct prediction of the target's presence, whereas a False Negative (FN) occurs when the model fails to detect a present target. The formulas for calculating these metrics, N-acc and T-acc, are as follows:

$$N - acc = \frac{TN}{TN+FP}, \quad (5)$$

$$T - acc = \frac{TP}{TP+FP}, \quad (6)$$

3.2 Subjective evaluation

To evaluate the prediction head of OGTransVG, we assess its performance in both target-present and target-absent scenarios. visualize the experimental results in Figure 3 illustrates the subjective experimental results under both target-present and target-absent scenarios. The left column of Figure 3 shows the results under the target-absent scenarios. The right column of Fig. 3 shows the results under the target-present scenarios.

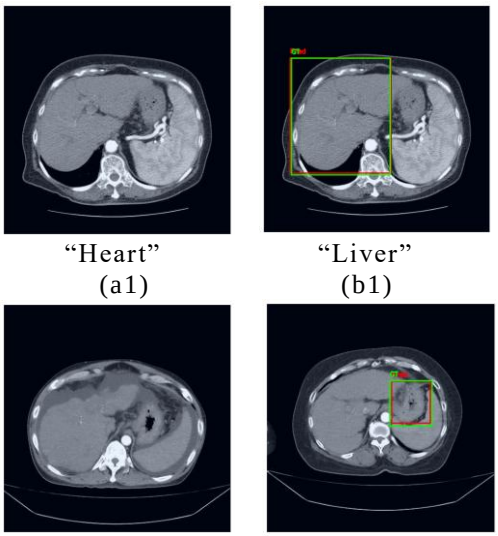

Figure 3 Prediction results of CT images w/o and w target: (a) without target and (b) with target

3.3 Objective evaluation

For testing set, the results of mIoU and Acc@0.7 are 82.49% and 100% for the proposed method, respectively. The results

show that the proposed method can detect organs well in CT images.

Table 1 lists the results of N-acc and T-acc for the proposed method. As shown in Table 1, the OGTransVG model demonstrates high accuracy, achieving over 81% in both target-present and target-absent scenarios. While N-acc is slightly lower than T-acc, both metrics remain at a consistently high level, reflecting the effectiveness of the added prediction head in determining the presence or absence of targets.

Table 1 Results of the proposed method for CT images w/ and w/o target.

	N-acc	T-acc
Total	81.01%	85.05%

IV. CONCLUSIONS

This paper presents an organ detection method called OGTransVG based on vision-language model. The proposed network developed based on the TransVG model and an additional classifier can not only identify organs by combining textual descriptions and image data but also determine the presence or absence of targets. Here we perform the subjective and objective evaluation for performance analysis. The results of mIoU and Acc@0.7 are 82.49% and 100% for the proposed method, respectively. The proposed method demonstrates its high accuracy, achieving over 81% in both target-present and target-absent scenarios. These experimental results indicate that OGTransVG effectively learns to distinguish target-present and target-absent scenarios. Therefore, the proposed OGTransVG network can not only detect organs well but also success-fully deal with the situation: no-target expressions.

REFERENCES

[1] Asrani SK, Devarbhavi H, Eaton J, Kamath PS. Burden of liver diseases in the world. *J Hepatol* 2019; **70:** 151–71.

[2] Crabb DW, Im GY, Szabo G, Mellinger JL, Lucey MR. Diagnosis and treatment of alcohol-associated liver diseases: 2019 practice guidance from the American Association for the Study of Liver Diseases. *Hepatology* 2020; **71:** 306–33.

[3] Eslam M, Sanyal AJ, George J, et al. MAFLD: a consensus-driven proposed nomenclature for metabolic associated fatty liver disease. *Gastroenterology* 2020; **158:** 1999–2014.

[4] Wilson R, Williams DM. Cirrhosis. Med Clin North Am. 2022 May;106(3): 437-446.

[5] Pugh RN, Murray-Lyon IM, Dawson JL, Pietroni MC, Williams R. Transection of the oesophagus for bleeding oesophageal varices. *Br J Surg* 1973; **60:** 646–49.

[6] Di Lelio A, Cestari C, Lomazzi A, Beretta L. Cirrhosis: diagnosis with sonographic study of the liver surface. Radiology 1989; 172: 389.

[7] Deng, J., Yang, Z., Chen, T., Zhou, W., & Li, H. (2021). TransVG: End-to-End Visual Grounding with Transformers. *arXiv preprint arXiv:2104.08541*.

[8] Zihao Zhao, Sheng Wang, Jinchen Gu, Yitao Zhu, Lanzhuju Mei, Zixu Zhuang, "ChatCAD+: Toward a Universal and Reliable Interactive CAD Using LLMs," IEEE Transactions on Medical Imaging, Vol. 43, Issue 11, pp. 3755 – 3766, November 2024.

[9] G.-S. Lin, K.-T. Lai, J.-M. Syu, J.-Y. Lin, and S.-K. Chai, "Instance segmentation based on deep convolutional neural networks and transfer learning for unconstrained psoriasis skin images," *Applied Sciences*, vol. 11, no. 7, p. 3155, 2021.

[10] G.-Z. Jian, G.-S. Lin, C.-M. Wang, and S.-L. Yan, "Classification of Helicobacter Pylori Infection Based on Deep Convolutional Neural Network with Visual Attention and Self-Supervised Learning for Endoscopic Images," *Multimedia Tools and Applications*, 2023.

[11] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, 2014, "Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787-798.

[12] S. J. Pan and Q. Yang, 2009, "A survey on transfer learning. In *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345-1359.

[13] C. Liu, H. Ding, and X. Jiang, 2023, "Gres: Generalized referring expression segmentation." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23592-23601.

[14] Y.-H. Zheng, G.-S. Lin, and K.-Y. Chang, "Transformer-based Visual Grounding with Inter-Modality Cross Attention," In *Proceedings of 19th International Conference on Machine Vision Applications*, 2025.

[15] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2016.

[16] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, vol. 18, pp. 203–211, 2021.

[17] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, "Segment Anything," ICCV, 2023.

[18] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, L. Zhang, "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection," arXiv:2303.05499, 2023.