Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL35249

Overview of Gastric Cancer Recognition Based on Deep Learning

Yufan Liu

Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, China
24106088d@outlook.com

Abstract. Gastric cancer is one of the most common malignant tumors with the top 5 incidence and mortality rates worldwide. Early detection and treatment can significantly improve patient survival. However, current detection mainly relies on manual inspection, with physicians using endoscopic images or pathological images to make a diagnosis. The accuracy of these manual approaches depends on the individual abilities of the physician, which often fails to meets clinical demands. In recent years, deep learning has been introduced into gastric cancer detection to assist physician in diagnosis. This paper reviews three popular deep learning-based methods: Feature fusion methods, Attention mechanism methods, and Self-attention mechanism methods, meanwhile introduces their applications, architecture, performance, and summarizes the commonly used datasets in this field. Based on these observations, this paper concludes about the current limitations of deep learning in gastric cancer detection, with possible optimization in future studies, aiming to provide a reference for the development of this field.

Keywords: Gastric Cancer, CV, CNN, Attention, Self-Attention

1. Introduction

1.1. Research background

Gastric tumors refer to neoplastic diseases that originate from the stomach. Based on their metastatic potential, tumors can be classified as benign and malignant. Malignant gastric tumors are known as gastric cancer, which ranks fifth in both incidence and mortality rates worldwide. According to the 2022 statistics from the Global Cancer Statistics, there were approximately 960,000 new cases of gastric cancer, accounting for 4.9% of all new cancer cases, and approximately 650,000 deaths, accounting for 6.8% of all cancer deaths [1]. Despite the serious threat that gastric cancer has on public health, it is worth noting that early detection and treatment can significantly improve the patient's survival rate. As reported, among 1537 patients with early-stage gastric cancer, 92.6% of them survived after five years by receiving early treatment [2]. This difference emphasizes the importance of improving the early detection rate of gastric cancer for effective treatment.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GL35249

1.2. Existing problems

Currently, early detection and treatment of gastric cancer mainly rely on endoscopic image screening and histopathological examination [3]. However, both methods rely on subjective judgment by physicians. When experience is limited, or experienced physicians are unavailable, misdiagnosis and missed diagnoses will occur easily. The report indicates that traditional manual inspections achieve an accuracy of only 69~79% [4]. Such accuracy is not acceptable for engaging in treatment, creating a clear demand for methods with higher accuracy.

To reduce the subjective interference of manual judgement, some researchers applied machine learning methods (ML) to gastric cancer image analysis. These methods usually use a single neural network or a simple model, by extracting features with different scales, ML can understand the image, thereby assisting in diagnosis. However, the ability of ML is limited by the model's capacity. When facing complex morphological changes of the gastric mucosa or when facing the fine-grained features of early lesions, the model performs poorly. Those cases indicated ML's lacking generalization, robustness, and interpretability, they fail to maintain high accuracy, and thereby lose high clinical reference value in practice [5]. These limitations have encouraged researchers to apply deep learning methods (DL) for stronger feature extraction ability.

1.3. Contributions of this paper

This paper organizes recent deep learning methods for gastric cancer image recognition, and focuses on three mainstream approaches: Feature fusion methods, Attention mechanism method, and Self-attention mechanism method. For each approach, this paper summarizes the most representative applications and outlines their structure with corresponding performance. Meanwhile, this paper also summarizes the current limitations of DL and suggests possible future directions for making improvements. The aim of this paper is to provide a reference for the development of DL in gastric cancer recognition.

2. Feature fusion methods

The performance of traditional machine learning methods depends on whether the features extracted by the model are important. To achieve improved performance, researchers explored strategies that fuse features from multiple sources; these methods are termed "Feature fusion methods". The logic of feature fusion methods is to break through the limitation of extracting all features from a single neural network. By fusing features from other channels, feature fusion methods are expected to enhance the accuracy and robustness of image recognition.

Based on the difference in feature sources, existing feature fusion methods can be divided into 3 types: "Neural network feature fusion", "Dataset feature fusion", and "Multimodal feature fusion". The structure and performance of the 3 types are shown in Table 1.

Table 1. The feature fusion methods

Reference	Method	Structure	Accuracy	Method F1
[6]	Neural network feature fusion	VGG16 + RESNET50 + MobileNetV2	97.81%	98.19%
[7]	Dataset feature fusion	LBP (GasHisSDB) + DenseNet-201	94.41%	95.40%
[8]	Multimodal feature fusion	BERT + ResNeXt (ResNet50)	94.14%	94.26%

DOI: 10.54254/2755-2721/2026.GL35249

2.1. Neural network feature fusion

This fusion method is to connect single models into a fusion model, where individual models perform well in extracting different types of features. The final fusion model aims to complement the feature extraction capabilities of the three models. Ma et al. (2025) selected three models that performed best on the representative gastric cancer dataset GasHisSDB [6]. Based on accuracy, the three models were connected into a fusion model. The final fusion model achieved an accuracy of 97.81%, indicating a significant improvement of at least 6.71% from the three individual models: VGG16 with an accuracy of 91.1%, RESNET50 with 88.3%, and MobileNetV2 with 89.8%. Although this method is not complex in design and practice, it validates the feasibility of fusing features from different neural networks, as well as demonstrates the potential of the feature fusion methods.

2.2. Dataset feature fusion

Features may not only come from neural networks, but also from manual designed features in specific datasets. Loddo et al. (2024) mainly fused handcrafted features (LBP) from GasHisSDB with features extracted from neural networks (DenseNet-201 and EfficientNetB0) [7]. The researchers used three classifiers for the final decision, including SVM, RF and DT. They tested every possible combination, but for clarity, this paper only uses the LBP + DenseNet-201 + SVM combination for comparison. When using only the features from DenseNet-201, SVM has only 86.02% accuracy with 97.89% F1score. When both LBP and DenseNet-201 features were input, SVM reached 94.41% accuracy with 95.40% F1score, achieving an improvement of 7.61% compared with DenseNet-201 + SVM. On the basis of neural network feature fusion, the significant improvement in dataset feature fusion has once again confirmed the potential of feature fusion methods.

2.3. Multimodal feature fusion

Compared with the two previous methods that fuse image features, Multimodal feature fusion fuses both text features and image features. Zhang et al. (2025) extracted features from both endoscopic images and diagnostic texts, mainly using ResNeXt (a ResNet50 variant) and BERT [8]. The final fusion model MSGC was tested on a private dataset collected from a Grade 3A hospital in Shandong Province, achieving 94.14% accuracy and 94.26% F1score, surpassing the single-modality baseline on the same dataset: (image-only ResNeXt: acc = 83.33%, F1 = 82.69%; BERT-only: acc = 87.70%, F1 = 87.67%).

3. Attention mechanism method

In traditional machine learning, convolution neural network (CNN), all spatial and channel features are treated equally. This makes the model hard to extract fine-grained features, and susceptible to complex background and random noise. For early-stage gastric cancer recognition, the target tissue often has subtle changes in texture, color, and morphology, traditional CNNs perform badly in capturing those features. However, the attention mechanism, by simulating human visual processing, uses dynamic weighting to make the model focus on certain important features or key areas. To better capture fine-grained features, researchers introduced attention mechanisms to highlight key

DOI: 10.54254/2755-2721/2026.GL35249

features and suppress irrelevant noise. The aim of this modification is to enhance the model's robustness, thereby boosting performance on early-stage gastric cancer recognition.

By the difference in how attention is introduced, current gastric cancer image recognition methods can be divided into 2 types: "Spatial and channel dual attention", and "Multi-channel attention". The structure and performance of the 2 types are shown in Table 2.

Table 2. The attention mechanism methods

Reference	Method	Structure	Accuracy	Method F1
[9]	Spatial and channel dual attention method	CNN+CBAM (CAM+SAM)	98.30%	98.60%
[10]	Multi-channel attention method	CNN+MCAM (SE+SimAM+ECA)	98.48%	98.72%

3.1. Spatial and channel dual attention

The most representative result of the spatial and channel dual attention mechanism is the Convolutional Block Attention Module (CBAM), proposed by Woo et al. in 2018 [11]. The channel attention (CAM) in CBAM is implemented by applying global average pooling and global max pooling to the feature maps. CAM then generates channel weights, determining which channel deserves more focus. Similarly, spatial attention (SAM) was implemented by calculating the mean value and max value along the channel dimension. SAM then generates spatial weights, determining which area of pixels deserves more focus. Arranging CAM and SAM in series results in CBAM, which is able to extract both channel and spatial attention simultaneously.

For the issue in gastric cancer endoscopic images where cancer cell lesion symptoms are too small for traditional CNNs to capture, Samira et al. (2024) introduced CBAM into their self-designed proposed CNN [9]. The original CNN achieved 93.5% accuracy and 96.5 F1score. After introducing CBAM, the improved CNN reached 98.3% accuracy with 98.6% F1score, achieving increases of 4.8% and 2.1% respectively. The improvement brought by CBAM verified the feasibility of the spatial and channel dual attention method.

3.2. Multi-channel attention (MCAM)

Based on CBAM using two attention channels, more channels can be introduced to emphasize more information. Zubair et al, (2025) the proposed Multi-Channel Attention Mechanism (MSCM), which pairs three distinct attention mechanisms with three CNNs to form three channels; their outputs are combined through weighted voting to produce a final result [10]. The structure and channel functions of MCAM are shown in Table 3.

Table 3. MCAM structure and channel functions

Channel	Attention	Neural Network	Function
Multi-scale Global Information Channel (MGIC)	Squeeze-and-Excitation (SE)	Inception-V3	Weighted feature channel, emphasize more important feature channels
Spatial Information Channel (SIC)	SimAM	VGG16	Weighted spatial position, emphasize more important areas in images

DOI: 10.54254/2755-2721/2026.GL35249

Table 3. (continued)

Multi-scale Spatial Information Channel (MSIC)	ECA	Xception	Weighted spatial position, emphasize different scale of local features in images

In MCAM, the SE attention mechanism introduced channel attention, the SimAM attention mechanism introduced spatial attention, ECA attention mechanism introduced multi-scale spatial attention. They provided the model with the ability to capture global features, key local features, and fine-grained features respectively. The researchers compared CNNs (e.g., VGG-16, Xception, Inception-V3, EfficientNetV2) with MCAM on the dataset GasHisSDB, and the best CNN (EfficientNetV2) achieved 98.06% accuracy and 98.40% F1score. In contrast, the proposed MCAM performed better on the same dataset, reached an average accuracy of 98.48% and an 98.72% F1score. This is an improvement of at least 0.32% over the best CNN, and at least 0.12% over CBAM. Given the baseline of 98%, this improvement is significant.

4. Self-attention mechanism methods

Traditional CNNs perform well in extracting features from a single area or single target, however, the target area in early-stage gastric cancer recognition may not be concentrated in one area. Usually, the lesion areas of gastric cancer appears scattered or interwoven with normal tissue. CNNs performs badly in those cases where the recognition targets have no clear contour.

To address this issue, researchers introduced the Transformer model. The Transformer model was originally designed for natural language processing. It uses a self-attention mechanism to capture relationships between any two positions in a text. Inspired by this mechanism, in 2020, a variant of Transformer, the Vision Transformer (ViT) was applied to image classification [12]. ViT's nice performance proved that the self-attention mechanism can also compensate for, or even surpass traditional CNNs in ambiguous target recognition.

Based on self-attention, researchers have developed methods for early-stage gastric cancer recognition. Depending on the carrier of the self-attention, current methods can be divided into 3 types: "Only ViT method", "Self-attention + CNN method", and "ViT variant method". The structure and performance of the 3 types are shown in Table 4.

Table 4. The self-attention mechanism methods

Reference	Method	Structure	Accuracy	Method F1
[13]	ViT method	Vision Transformer	85.90%	78.00%
[14]	Self-attention + CNN	(ResNet50+MHSA) + Inception-V3	98.55%	97.97%
[15]	ViT variant method	ViT + GAEM + DPSM	96.51%	96.64%

4.1. ViT method

Because ViT is a variant of Transformer, it already carries a self-attention mechanism. Thus, ViT can perform image classification without relying on any additional self-attention modules. Poupi et al. (2024) tested several well-performed CNNs in early-stage gastric cancer recognition on the GasHisSDB dataset, including ResNet50, DenseNet121, and EfficientNetB1 [13]. Their best performance reached 78.71% accuracy, with 76.0% F1score. Meanwhile, the researchers tested the

DOI: 10.54254/2755-2721/2026.GL35249

performance of ViT on the same dataset, which outperformed every baseline CNNs, and reached 85.9% accuracy with 78.0% F1score. Clearly, this comparison has proved the potential of ViT, as well as the self-attention mechanism.

4.2. Self-attention + CNN method

Although ViT has excellent performance in capturing global features, its ability to capture local features is not satisfactory. Traditional CNNs are expected to fill this gap. To address this issue, Chen et al. (2022) proposed a solution combining modular self-attention with CNN [14]. Specifically, they replaced the last 3x3 convolutional layer of ResNet50 with a Multi-Head Self-Attention (MHSA) layer. This turned ResNet50 into a parallel channel that extracts local features. Meanwhile, another parallel channel uses Inception-V3 to extract local features. The two outputs were fused through fully connected layers with SoftMax to determine the final prediction.

The researchers named this fusion architecture "GasHis-Transformer". On the public gastric histopathology image dataset, HE-GHI-DS, GasHis-Transformer achieved 98.55% accuracy and 97.97% F1score. In comparison, the original ResNet50 and Inception-V3 only reached a maximum accuracy of 94.24% with 94.26% F1score. The fusion architecture showed a clear improvement of at least 3.71%. Moreover, the performance far surpasses the pure ViT method on the same dataset, which achieved an accuracy of 84.78% with 84.88% F1score, validating the feasibility and potential of this method.

4.3. ViT variant method

The ViT method has been proven to be accurate in basic gastric cancer recognition. To further optimize the generality of self-attention mechanisms in this task, researchers have developed various variant ViTs for specific tasks. The most representative variant model is the Coarse-to-Fine Inference based Vision Transformer (CFI-ViT) proposed by Wang et al. (2025), specifically designed for gastric cancer subtype classification tasks [15].

The CFI-ViT uses a "coarse to fine" two-stage interference strategy: In the coarse inference stage, the model uses ViT to split the downsampled image into multiple patches, each performing an inference individually. If a patch has low confidence, it enters the fine interference stage. The model uses the Global Attention Extraction Module (GAEM) to produce a global attention score and select the top-M patches that most likely contain useful information. Those patches use original-sized images instead of downsampled ones; the output is then fed into ViT again, combined with the coarse inference results to determine the final result.

CFI-ViT avoids common issues such as scattered lesions, complex backgrounds, and high computational costs for high-resolution images. The CFI-ViT achieved 96.51% accuracy and 96.64% F1score on the HE-GHI-DS dataset, surpassing the ViT method (86.05% accuracy and 86.12% F1score) and EfficientNet-B0 method (95.45% accuracy and 96.15% F1score) on the same dataset.

5. Datasets and samples

The following is an overview of datasets commonly used in early-stage gastric cancer recognition:

HE-GHI-DS: Released in 2022, often used complementarily with GasHisSDB because of the difference in image size. The content of this dataset is 700 full pathological tissue images with a size of 2048x2048. Similar to GasHisSDB, the images in this dataset are binary classified, but the dataset

DOI: 10.54254/2755-2721/2026.GL35249

is unbalanced, containing 560 "Normal" class images and 140 "Abnormal" class images. Figure 1 shows samples of the "Normal" and "Abnormal" classes from HE-GHI-DS.

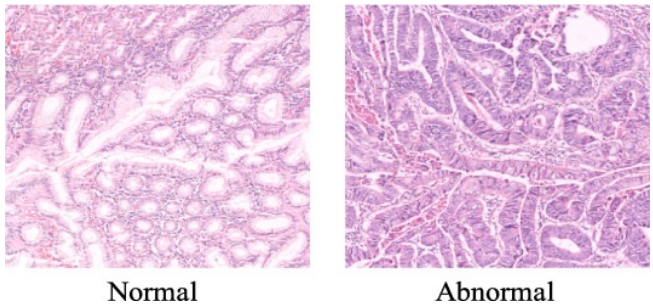


Figure 1. HE-GHI-DS dataset samples (data from: HE-GUI-DS) conclusions

GasHisSDB [16]: Released in 2021, by a team from Northeastern University. It provides large-scale pathology images for computer-aided gastric cancer diagnosis. It is currently the first publicly available large-scale gastric cancer histopathology image dataset, as well as the most commonly used dataset. The content of this dataset was cropped from 600 full-size slide images, including three sub-size images of 80x80, 120x120, and 160x160. It contains 245,196 binary classification images in total, with 148,120 "Normal" class images and 97,076 "Abnormal" class images. Figure 2 shows samples of "Normal" and "Abnormal" class from GasHisSDB, subfigures (a) and (b) are normal tissue, while (c) and (d) are lesion tissue.

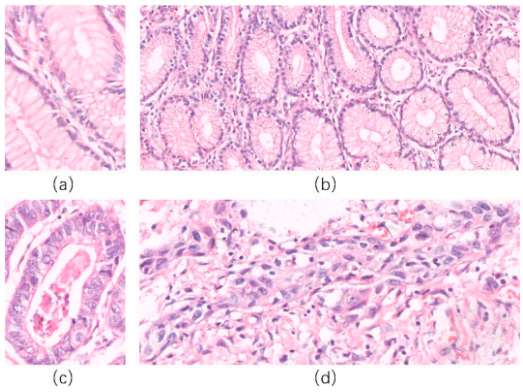


Figure 2. GasHisSDB dataset samples (data from: GasHisSDB)

6. Conclusion

This paper reviews three main deep learning methods for gastric cancer recognition: Feature fusion methods, Attention mechanism method, and Self-attention mechanism method.

Feature fusion methods improve model robustness through multiple models fusion, multiple data sources fusion, and multiple feature sources fusion; Attention mechanism methods emphasized target areas and suppress noise through attention, including spatial and channel attention or multi-channel attention, enhanced model's ability to extract fine-grained features; Self-attention mechanism method uses variant and hybrid architecture models, effectively improved the model's ability to capture scattered lesion areas from complex background images.

The above studies were mainly conducted on the GasHisSDB dataset and partly on the HE-GUI-DS dataset. All achieved performance superior to traditional CNNs, verifying the great potential of

DOI: 10.54254/2755-2721/2026.GL35249

deep learning methods in aiding early-stage gastric cancer diagnosis. However, current research still faces three key issues centered around data sources:

First, insufficient datasets

Most studies, including those reviewed earlier in this paper, rely on a single dataset (GasHisSDB), which lead to insufficient model generalizability. Future research requires more datasets, or more interpretability when trained on a single dataset.

Second, inconsistent standards:

To address the issue of insufficient datasets, researchers have begun using self-made or private datasets for model training. However, this introduces another problem: excellent machine learning models require all data coming from the same distribution and standards. But, differences in scanning protocols, imaging equipment, and image resolution cause data from different regions and hospitals to vary. Therefore, future research needs to establish and promote a standardized gastric scanning protocol to improve model generalizability.

Third, insufficient high-quality annotations:

Creating datasets requires high-quality annotations, and high-quality annotations depend on precise human judgment. However, early gastric cancer recognition is difficult, with limited domain experts; obtaining annotations of high-quality and high quantity remains challenging. In recent years, various self-supervised and semi-supervised learning methods have been developed. Future research should focus on the usage of unlabeled data, or reducing dependence on high-quality annotated data.

References

[1] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians, 74(3), 229–263.

[2] Suzuki, H., Oda, I., Abe, S., Sekiguchi, M., Mori, G., Nonaka, S., Yoshinaga, S., & Saito, Y. (2015). High rate of 5-year survival among patients with early gastric cancer undergoing curative endoscopic submucosal dissection. Gastric Cancer, 19(1), 198–205.

[3] Zhou Di, Lu Mingliang & Jiang Mingming. (2021). Research on early diagnosis of gastric cancer. Journal of Gastroenterology and Hepatology, 30(06), 606-611.

[4] Choi, J., Kim, S. G., Im, J. P., Kim, J. S., Jung, H. C., & Song, I. S. (2010). Comparison of endoscopic ultrasonography and conventional endoscopy for prediction of depth of tumor invasion in early gastric cancer. Endoscopy, 42(09), 705–713.

[5] Huang, Libing. (2023). Interpretability of deep learning models and its research progress in medical image analysis. Journal of Guilin Aerospace University, 28(01), 51-60.

[6] Ma, J., Yang, F., Yang, R., Li, Y., & Chen, Y. (2025). Interpretable deep learning for gastric cancer detection: A fusion of AI architectures and explainability analysis. Frontiers in Immunology, 16, Article 1596085.

[7] Loddo, A., Usai, M., & Di Ruberto, C. (2024). Gastric Cancer Image Classification: A Comparative Analysis and Feature Fusion Strategies. Journal of Imaging, 10(8), 195. https: //doi.org/10.3390/jimaging10080195

[8] Zhang, X., Zheng, X., Dong, M., & Zhang, M. (2025). An endoscopic images and diagnostic records based multimodal method for severity grading of gastric cancer. Biomedical Signal Processing and Control, 109, 107891.

[9] Samira, S., Faruk, Md. F., Islam, M. N., Mahedy Hasan, S. M., Mamun, Md. A., Srizon, A. Y., & Hossain, Md. R. (2024). Enhancing Gastrointestinal Diagnosis: CBAM-Integrated CNN for High-Precision Multi-Class Anatomical Landmark Identification. 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), 1–6.

[10] Zubair, M., Owais, M., Hassan, T., Bendechache, M., Hussain, M., Hussain, I., & Werghi, N. (2025). An interpretable framework for gastric cancer classification using multi-channel attention mechanisms and transfer learning approach on histopathology images. Scientific Reports, 15(1).

[11] Woo, S., Park, J., Lee, JY., & Kweon, I.S. (2018). CBAM: Convolutional Block Attention Module. In Ferrari, V., Hebert, M., Sminchisescu, C., & Weiss, Y. (Eds.), Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in

DOI: 10.54254/2755-2721/2026.GL35249

Computer Science(), 11211. Springer, Cham.
[12] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2010). AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE.
[13] Poupi, T., Armand, T., Bhattacharjee, S., Kim, H.-J., Hussain, A., Ali, S., Choi, H.-K., & Kim, H.-C. (2024). Vision Transformer-based Model for Gastric Cancer Detection and Classification using Weakly Annotated Histopathological Images.
[14] Chen, H., Li, C., Wang, G., Li, X., Mamunur Rahaman, M., Sun, H., Hu, W., Li, Y., Liu, W., Sun, C., Ai, S., & Grzegorzek, M. (2022). GasHis-Transformer: A multi-scale visual transformer approach for gastric histopathological image detection. Pattern Recognition, 130, 108827.
[15] Wang, X., Tao, H., Wang, B., Jin, H., & Li, Z. (2025). CFI-ViT: A coarse-to-fine inference based vision transformer for gastric cancer subtype detection using pathological images. Biomedical Signal Processing and Control, 100, 107160.
[16] Hu, W., Li, C., Li, X., Md Mamunur Rahaman, Ma, J., Zhang, Y., Chen, H., Liu, W., Sun, C., Yao, Y.-D., Sun, H., & Marcin Grzegorzek. (2022). GasHisSDB: A new gastric histopathology image dataset for computer aided diagnosis of gastric cancer. Computers in Biology and Medicine, 142, 105207–105207.