

★ 代表性论著目录:

全部作者名（通讯作者标注*号）	论 著 名 称	出版单位、出版年月/刊物名称、卷、期、发表年月	研究方向
Zifan Chen [#] , Yang Chen [#] , Yu Sun [#] , Lei Tang [#] , Li Zhang[#] , Yajie Hu, Meng He, Zhiwei Li, Siyuan Cheng, Jiajia Yuan, Zhenghang Wang, Yakun Wang, Jie Zhao, Jifang Gong, Liying Zhao, Baoshan Cao, Guoxin Li, Xiaotian Zhang*, Bin Dong*, Lin Shen*	Predicting gastric cancer response to anti-HER2 therapy or anti-HER2 combined immunotherapy based on multi-modal data	Signal Transduction and Targeted Therapy, 9(1), 222, 2024-8	应用多模态医疗数据实现抗胃癌HER2/联合免疫疗法的疗效预测
Mingze Yuan, Yingda Xia*, Hexin Dong, Zifan Chen, Jiawen Yao, Mingyan Qiu, Ke Yan, Xiaoli Yin, Yu Shi, Xin Chen, Zaiyi Liu, Bin Dong, Jingren Zhou, Le Lu, Ling Zhang, Li Zhang	Devil is in the Queries: Advancing Mask Transformers for Real-world Medical Image Segmentation and Out-of-Distribution Localization	Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023-6	面向真是世界医疗数据中近分布外数据的建模分析
Jiajia Yuan, Peng Bao, Zifan Chen, Mingze Yuan, Jie Zhao, Jiahua Pan, Yi Xie, Yanshuo Cao, Yakun Wang, Zhenghang Wang, Zhihao Lu, Xiaotian Zhang, Jian Li, Lei Ma, Yang Chen*, Li Zhang* , Lin Shen* and Bin Dong*	Advanced prompting as a catalyst: Empowering large language models in the management of gastrointestinal cancers	The Innovation Medicine 1(2), 2023-9	大语言模型中提示词技术在医疗中的应用
Hexin Dong, Zifan Chen, Mingze Yuan, Yutong Xie, Jie Zhao, Fei Yu, Bin Dong, Li Zhang*	Region-Aware Metric Learning for Open World Semantic Segmentation via Meta-Channel Aggregation	Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI) Main Track, 2022-7	面向开放世界问题定量分割（可应用于医疗图像数据中少见/罕见数据的定量分析）
Fei Yu, Mo Zhang, Hexin Dong, Sheng Hu, Bin Dong, Li Zhang*	DAST: Unsupervised Domain Adaptation in Semantic Segmentation Based on Discriminator Attention and Self-Training	Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2021-2	基于领域自适应的分割（可应用于解决不同中心数据的分布差异问题）



ARTICLE OPEN

Predicting gastric cancer response to anti-HER2 therapy or anti-HER2 combined immunotherapy based on multi-modal data

Zifan Chen¹, Yang Chen², Yu Sun³, Lei Tang⁴, Li Zhang^{1,5}, Yajie Hu³, Meng He⁴, Zhiwei Li^{6,7}, Siyuan Cheng⁸, Jiajia Yuan², Zhenghang Wang², Yakun Wang², Jie Zhao⁹, Jifang Gong², Liying Zhao^{6,7}, Baoshan Cao⁸, Guoxin Li^{6,7}, Xiaotian Zhang²✉, Bin Dong^{5,10,11}✉ and Lin Shen²✉

The sole use of single modality data often fails to capture the complex heterogeneity among patients, including the variability in resistance to anti-HER2 therapy and outcomes of combined treatment regimens, for the treatment of HER2-positive gastric cancer (GC). This modality deficit has not been fully considered in many studies. Furthermore, the application of artificial intelligence in predicting the treatment response, particularly in complex diseases such as GC, is still in its infancy. Therefore, this study aimed to use a comprehensive analytic approach to accurately predict treatment responses to anti-HER2 therapy or anti-HER2 combined immunotherapy in patients with HER2-positive GC. We collected multi-modal data, comprising radiology, pathology, and clinical information from a cohort of 429 patients: 310 treated with anti-HER2 therapy and 119 treated with a combination of anti-HER2 and anti-PD-1/PD-L1 inhibitors immunotherapy. We introduced a deep learning model, called the Multi-Modal model (MuMo), that integrates these data to make precise treatment response predictions. MuMo achieved an area under the curve score of 0.821 for anti-HER2 therapy and 0.914 for combined immunotherapy. Moreover, patients classified as low-risk by MuMo exhibited significantly prolonged progression-free survival and overall survival (log-rank test, $P < 0.05$). These findings not only highlight the significance of multi-modal data analysis in enhancing treatment evaluation and personalized medicine for HER2-positive gastric cancer, but also the potential and clinical value of our model.

Signal Transduction and Targeted Therapy (2024)9:222

; <https://doi.org/10.1038/s41392-024-01932-y>

INTRODUCTION

Gastric cancer (GC) is the fifth most prevalent cancer globally, and the second most common cancer in China.¹ Approximately 15–30% of advanced gastric or gastroesophageal junction adenocarcinomas exhibit amplification or overexpression of the human epidermal growth factor receptor 2 (ERBB2/HER2).² The heterogeneity of this biomarker poses substantial challenges for effective treatment, with responses varying widely among patients. The trastuzumab for GC trial revealed that less than half patients with HER2-positive responded to a combination of trastuzumab and chemotherapy,³ indicating significant intra-patient and inter-tumor variability. Further complicating the treatment landscape, the KEYNOTE-811 study's interim findings⁴ showed that although adding pembrolizumab to standard therapy substantially increases objective response rates as a first-line therapy, this does not equate to a uniform enhancement in overall survival (OS) for all patients. The discrepancies in survival

rates underscore the complexity of the disease and suggest that conventional monomodal data may be insufficient for understanding the diverse presentations of HER2-positive GC, necessitating a comprehensive evaluation using multi-modal data. By integrating clinical profiles, radiological imaging, and pathological samples, a more nuanced understanding of tumor behavior is possible, which is imperative for refining treatment decisions. Therefore, an integrated multi-modal approach is essential: to fully characterize the heterogeneity of HER2-positive GC and devise personalized and effective treatment strategies.

The potential of artificial intelligence (AI) as an innovative tool for developing multimodal models is high,^{5–9} and its strength lies in its ability to analyze different data types and integrate them at the feature level.^{10–16} However, the application of AI in predicting treatment response is still in its infancy,¹⁷ particularly for predicting treatment response in complex diseases such as GC,^{18,19} a task that is far more difficult than diagnosis.²⁰ Diagnostic

¹Center for Data Science, Peking University, Beijing, China; ²Department of Gastrointestinal Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing, China; ³Department of Pathology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing, China; ⁴Department of Radiology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing, China; ⁵National Biomedical Imaging Center, Peking University, Beijing, China; ⁶Department of General Surgery, Nanfang Hospital, Southern Medical University, Guangzhou, China; ⁷Guangdong Provincial Key Laboratory of Precision Medicine for Gastrointestinal Tumor, Guangzhou, China; ⁸Department of Medical Oncology and Radiation Sickness, Peking University Third Hospital, Beijing, China; ⁹National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing, China; ¹⁰Beijing International Center for Mathematical Research (BICMR), Peking University, Beijing, China and ¹¹Center for Machine Learning Research, Peking University, Beijing, China

Correspondence: Xiaotian Zhang (zhangxiaotianmed@163.com) or Bin Dong (dongbin@math.pku.edu.cn) or Lin Shen (shenlin@bjmu.edu.cn)

These authors contributed equally: Zifan Chen, Yang Chen, Yu Sun, Lei Tang, Li Zhang.

Received: 21 February 2024 Revised: 4 July 2024 Accepted: 17 July 2024

Published online: 26 August 2024

models typically rely on large datasets, including normal and abnormal samples.^{20,21} However, predicting treatment response requires more refined datasets that are specific to a particular disease stage and that reflect the subtle effects of different treatment regimens over time.^{22,23} Additionally, the incompleteness of treatment datasets poses technical challenges in the AI model construction and learning process.^{24,25} In realistic anti-HER2 therapy or anti-HER2 combined immunotherapy scenarios, the patient's imaging history may only be partially available. This modality deficit has not been fully considered in many studies,^{26,27} while it may seriously affect the learning ability of the model and its utility in clinical decision-making.

To address the challenges, this study aimed to use a comprehensive analytical approach to accurately predict the treatment response in patients with GC receiving anti-HER2 therapy or anti-HER2 combined immunotherapy. We assembled a comprehensive multi-center dataset of 429 patients. This valuable dataset integrates diverse modalities of information from the baseline treatment phase, including radiological computerized tomography (CT) scans, pathological whole-slide images, radiological and pathological reports, and general patient information. Our study focused on two distinct cohorts of patients treated at the Peking Cancer Hospital: the anti-HER2 (those receiving anti-HER2 therapy and chemotherapy) and anti-HER2 combined immunotherapy (those receiving anti-HER2 therapy combined with anti-PD-1/PD-L1 immune checkpoint inhibitors [ICI] and chemotherapy) cohorts. We further supplemented these cohorts with additional patients from external medical facilities. With this rich dataset, we developed a unified, transformer-based deep learning model called the **Multi-Modal** model (MuMo), which effectively incorporates multi-modal inputs to predict the treatment response. MuMo represents a major advancement in leveraging diverse data types, even in cases of missing modalities, to improve prediction accuracy. Experimental results demonstrate MuMo's ability to extract complementary insights from multi-modal data and provide more accurate treatment response predictions.

RESULTS

Multi-modal dataset and cohort characteristics in HER2-positive GC study

Our study commenced with collecting an extensive multi-modal dataset encompassing radiology, pathology, and patient information, from a large group of 17,787 patients with GC during the baseline treatment phase at multiple centers. This dataset included data from Peking University Cancer (PKCancer) Hospital, Nanfang Hospital, and Peking University Third Hospital. Rigorous selection criteria were applied to refine the cohort (Fig. 1). These criteria include excluding patients with negative or unknown HER2 status, those not undergoing anti-HER2 therapy, and those lacking multi-modal data. This process resulted in a cohort of 429 patients with HER2-positive GC, between January 2007 and January 2023 (Table 1) (with additional information in Supplementary Tables S1, S2). As shown in Fig. 2, of the 429 patients, 390 were from Peking University Cancer Hospital. Among them, 271 underwent anti-HER2 therapy, forming an anti-HER2 cohort. The remaining 119 patients, forming an anti-HER2 combined immunotherapy cohort, received a combination of anti-HER2 therapy with either anti-PD-1 inhibitors (85 patients) or anti-PD-L1 inhibitors (34 patients). Additionally, we included an external cohort of 39 patients from Nanfang Hospital and Peking University Third Hospital. Most patients in these cohorts were diagnosed with stage IV GC, with a prevalence of 98.52% in the anti-HER2 cohort, 97.48% in the anti-HER2 combined immunotherapy cohort, and 97.44% in the external cohort. The median age of the patients in the three cohorts was 63 (interquartile range [IQR]: 55–69), 65 (IQR: 58–72), and 60 years (IQR: 55–68), respectively. The percentage of men

was 83.03% in the anti-HER2 cohort, 79.83% in the anti-HER2 combined immunotherapy cohort, and 76.92% in the external cohort. Regarding tumor location, the majority were non-gastroesophageal junction (non-GEJ) tumors, accounting for 69.74%, 70.59%, and 79.49% of patients in each cohort, respectively. Additionally, most patients in all cohorts had either moderately differentiated (47.60%, 47.06%, and 35.90%, respectively) or poorly differentiated (48.71%, 52.10%, and 58.97%, respectively) carcinomas.

Our amassed multi-modal dataset offers rich and comprehensive patient data. This included demographic details (that is, age and sex), tumor characteristics (that is, tumor location, degree of differentiation, and Lauren classification), and treatment specifics (that is, lines of treatment received and time elapsed before initiating treatment). These details are visually represented in Fig. 3b, c and Supplementary Fig. S1. Moreover, each patient's dataset included data from at least one modality of both pathology and radiology. However, complete data from both modalities (radiology and pathology) was available for less than half of the patients (Fig. 3a). To address this variability in data availability, learnable embeddings were introduced as placeholders for missing modalities. This technology helps to infer missing information, enhancing the robustness of our multi-modal model (detailed methodology in the methods section). Furthermore, structured clinical reports from both radiological and pathological assessments provided additional crucial clinical insights, aiding in a more comprehensive understanding of each patient's condition (Fig. 2b, c and Supplementary Fig. S1). Specifically, radiological reports included detailed information on post-operative status (whether the patient had undergone gastrectomy), the count and locations of metastatic lymph nodes, occurrences of liver or lung metastases, peritoneal metastasis, and diversity in metastatic lymph node types. In contrast, pathological reports encompassed data on the proportion of tumors, tumor-infiltrating lymphocytes (TILs), and variability in HER2 expression within the tumor.

MuMo's predictive performance in the anti-HER2 cohort

The proposed MuMo demonstrated promising efficacy in predicting treatment responses in the anti-HER2 cohort, achieving an area under the curve (AUC) score of 0.821 (95% Confidence Interval [CI]: 0.692–0.949; Fig. 4a and Supplementary Table S3). Additionally, MuMo exhibited an impressive number needed to treat a value of 1.83 (95% CI: 1.28–4.24; refer to Supplementary Table S4), indicating a high efficiency in predicting treatment response. MuMo's predictive performance surpassed that of the six individual clinicians in a similar test and even matched the combined score of consultation among these clinicians (Supplementary Figs. S2a, b). To evaluate the generalizability of MuMo, we applied it to an external cohort that functioned as an independent test set. MuMo showed a strong discriminative ability in distinguishing non-responders from responders, as indicated by an AUC score of 0.884 (95% CI: 0.745–1.000; Fig. 4b). This performance highlights the effectiveness of MuMo in multi-center data cases. Additionally, we have showcased the flexible extensibility of the MuMo framework by utilizing two public datasets: TCGA-STAD for gastric adenocarcinoma and TCGA-BRCA for invasive breast carcinoma, detailed in Supplementary Figs. S3, S4 and Supplementary Text S1. This extensibility is further illustrated through its application to potential molecular pathology data, as documented in Supplementary Fig. S5 and Supplementary Text S2. Using the Youden index, a statistical measure derived from the receiver operating characteristic (ROC) curves, we stratified each cohort into high- and low-risk groups based on MuMo's predictive scores (see Supplementary Text S3 for details). The low-risk group exhibited significantly longer progression-free survival (PFS) (log-rank test, $P=0.0019$ in the validation set and $P=0.0024$ in the test set; Fig. 4d, e) and

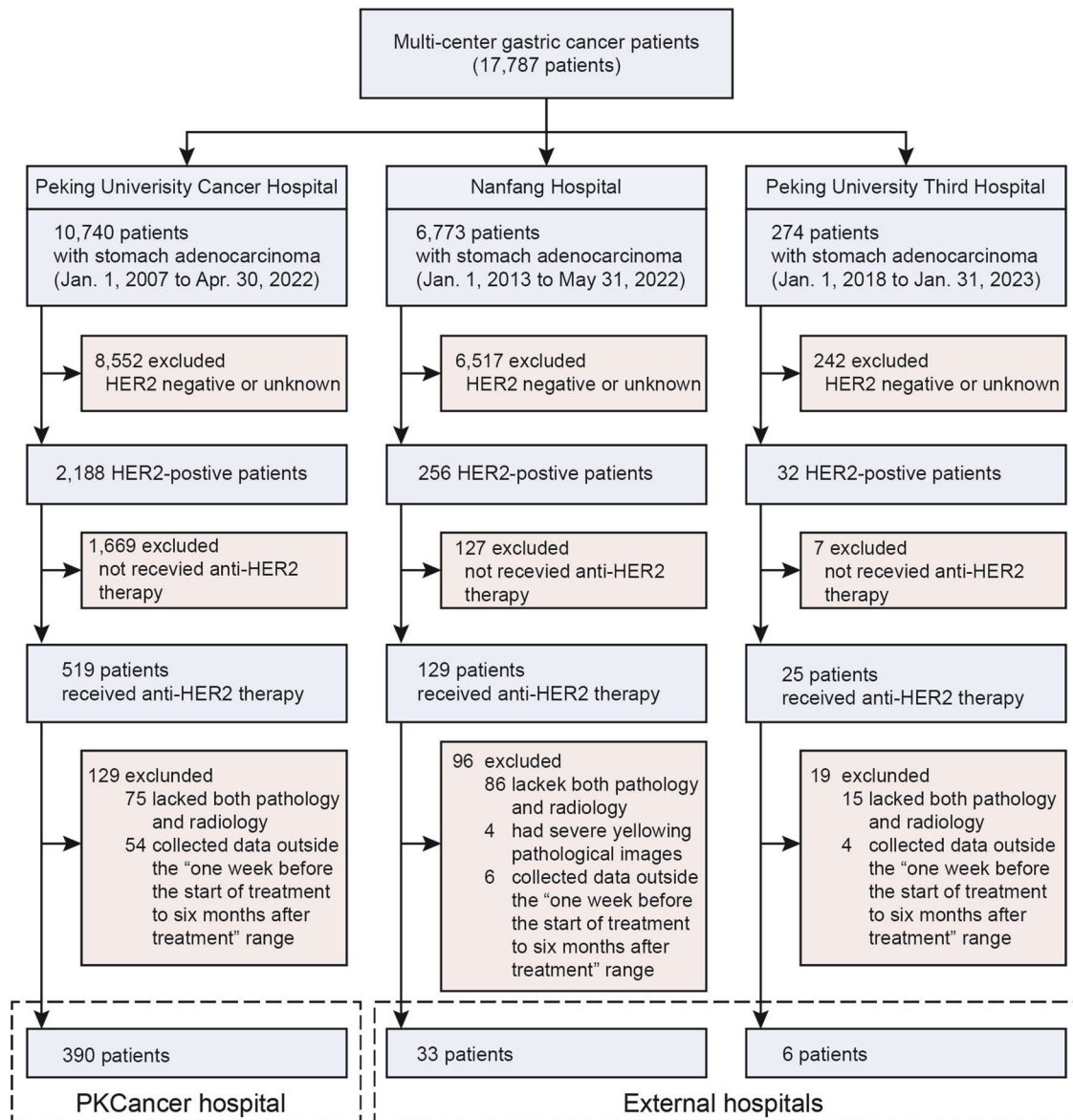


Fig. 1 Multi-modal data collection and filtering flowchart for patients with HER2-positive GC from multi-center study. The cohort, comprising 17,787 patients with GC, was derived from a consecutive series of patients diagnosed with stomach adenocarcinoma at Peking University Cancer (PKCancer) Hospital (10,740 patients), Nanfang Hospital (6773 patients), and Peking University Third Hospital (274 patients) between January 2007 and January 2023. The selection process involved exclusions based on HER2 negativity or unknown status, absence of anti-HER2 therapy, and lack of comprehensive multimodal data, including necessary pathological and radiological information collected within a defined period around treatment initiation. The final multi-modal analysis cohort was segmented into 390 patients from the PKCancer Hospital and 39 patients from external hospitals

increased OS (log-rank test, $P = 0.0067$ in the validation set; Fig. 4g) than the high-risk group. Notably, in the independent test set, a marked difference in median OS was observed between the two groups (6 months for the high-risk group vs. 17 months for the low-risk group; Fig. 4h).

MuMo's adaptability in predicting responses in the anti-HER2 combined immunotherapy cohort
To further evaluate the adaptability of MuMo in diverse treatment cohorts, we analyzed its predictive efficacy in the anti-HER2 combined immunotherapy cohort. In this cohort, MuMo demonstrated remarkable performance with an AUC of 0.914 (95% CI: 0.803–1.000; Fig. 4c, Supplementary Table S3 and Supplementary Fig. S2c), indicating high accuracy in response predictions. Moreover, MuMo proficiently differentiated between high- and

low-risk groups in terms of PFS (log-rank test, $P = 0.0079$; Fig. 4f) and OS (log-rank test, $P = 0.0042$; Fig. 4i), which are essential for patient prognosis and treatment planning. This impressive performance highlights the critical role of MuMo in adapting to a relatively novel treatment regimen, such as anti-HER2 combined immunotherapy, where clinician experience and historical data may be limited. The capability of advanced deep learning models, such as MuMo, to extract meaningful insights from various cohorts demonstrates their potential for broad applications in emerging treatment scenarios.

MuMo's consistent stability in treatment response prediction
To ascertain the reliability and consistency of MuMo in predicting treatment responses, we conducted a confirmatory experiment using 2000 bootstrap replicates, a statistical method to estimate

Table 1. Baseline characteristics of anti-HER2, anti-HER2 combined immunotherapy and external hospital cohorts

Characteristic	Anti-HER2 cohort (n = 271)	Anti-HER2 combined immunotherapy cohort (n = 119)	External cohort (n = 39)
Age			
Median, IQR	63, 55–69	65, 58–72	60, 55–68
Sex			
Male	225 (83.03%)	95 (79.83%)	30 (76.92%)
Female	46 (16.97%)	24 (20.17%)	9 (23.08%)
Tumor site			
GEJ	82 (30.26%)	35 (29.41%)	8 (20.51%)
Non-GEJ	189 (69.74%)	84 (70.59%)	31 (79.49%)
Degree of differentiation			
Poorly	132 (48.71%)	62 (52.10%)	23 (58.97%)
Moderately	129 (47.60%)	56 (47.06%)	14 (35.90%)
Well	10 (3.69%)	1 (0.84%)	2 (5.13%)
Lauren type			
Intestinal	17 (6.384%)	82 (68.91%)	6 (15.79%)
Diffused	34 (12.55%)	12 (10.08%)	1 (2.63%)
Mixed	43 (15.87%)	16 (13.45%)	1 (2.63%)
N/A	21 (7.75%)	9 (7.56%)	31 (79.49%)
PD-L1 expression			
Positive	44 (16.24%)	43 (36.13%)	8 (20.51%)
Negative	60 (22.14%)	25 (21.01%)	8 (20.51%)
N/A	167 (61.62%)	51 (42.86%)	23 (58.97%)
MMR status			
pMMR	137 (50.55%)	104 (87.39%)	18 (46.15%)
dMMR	2 (0.74%)	1 (0.84%)	0 (0.00%)
N/A	132 (48.71%)	14 (11.76%)	21 (53.85%)
EBV status			
Positive	3 (1.11%)	1 (0.84%)	0 (0.00%)
Negative	119 (43.91%)	89 (74.79%)	9 (23.08%)
N/A	149 (54.98%)	29 (24.37%)	30 (76.92%)
TNM stages			
III	4 (1.48%)	3 (2.52%)	1 (2.56%)
IV	267 (98.52%)	116 (97.48%)	38 (97.44%)

the sampling distribution, for both the anti-HER2 and anti-HER2 combined immunotherapy cohorts. We calculated the AUC scores for each replicate, and these scores were visually represented using box plots, which effectively illustrated the distribution and variability of the scores. The results demonstrated that MuMo exhibited a small performance variability, which is a key indicator of its consistent stability and reliability in predicting treatment responses (Fig. 4j). Furthermore, MuMo displayed significantly lower performance variability than both individual clinicians and their collective decision-making processes in group consultations (Levene’s test, $P < 0.05$; Supplementary Fig. S6). Additionally, we demonstrated MuMo’s stable predictive performance (AUC 0.800 to 0.833; Supplementary Fig. S7) across five sets of randomly varied doctor annotations (Supplementary Table S5). These outcomes highlight MuMo’s capability to provide dependable and stable predictions in treatment response scenarios.

Ablation studies of multi-modal information fusion in MuMo
The proposed MuMo offers a comprehensive perspective of patients with GC undergoing anti-HER2 therapy, largely because of its specially designed fusion modules for multi-modal

information fusion. To assess the contribution of the MuMo fusion modules to multi-modal information, we conducted three ablation studies.

In our first experiment, we analyzed patients with both radiological and pathological data (Fig. 4k and Supplementary Table S6). Our results showed that integrating radiological and pathological data improved the predictive AUC score of the model to 0.750, which was superior to the AUC scores achieved by models relying solely on radiological (0.639) or pathological (0.703) data. The pathology-only model performed notably better than the radiology-only model, which can be attributed to the comprehensive visual insights provided by pathological analysis. As the clinical gold standard for diagnosis, pathology offers an effective portrayal of the tumor immune microenvironment, which is a crucial determinant of treatment response.²⁸

Second, we evaluated the effectiveness of MuMo’s specialized inter-modal fusion module, which includes modal-agnostic feature alignment, to integrate disparate information sources (more details in the methods section). The fusion module showed enhanced performance compared with simple combinations of features derived from radiological and pathological data (Fig. 4l and Supplementary Table S7), such as element-wise multiplication (0.577), summation (0.682), and concatenation (0.731). This underscores MuMo’s ability to effectively consolidate inter-modal information. Moreover, when we assessed a MuMo variant that lacked modal-agnostic feature alignment in the latent space, its performance decreased (AUC scores: 0.772) compared to the full MuMo (0.821), demonstrating the advantage of feature alignment across different modalities in a unified embedding space, as supported by the relevant literature.^{10,11}

Finally, we found that incorporating clinical reports led to a promising increase in the AUC score (0.703–0.769; Fig. 4m and Supplementary Table S8). Furthermore, including detailed patient information improved the performance of model (increasing the AUC scores: 0.769–0.821). These findings suggest that comprehensive data from clinical reports and patient-specific information significantly enhanced MuMo’s predictive capabilities.

MuMo’s interpretability with clinical insights

We validated the alignment between MuMo predictions and established clinical knowledge, focusing on two key perspectives: image-focused regions for visual qualitative analysis and clinical information weights for quantitative analysis. In the pathological whole-slide images, we used regional important scores that quantified the model’s focus on specific areas to highlight where the model concentrated its predictions. Notably, these focus areas correlated intuitively with HER2 (3+) expression regions, characterized by a high tumor-to-stroma ratio (over 50%), well-differentiated tumor glands, abundant tumor-associated immune cell infiltration, and significant desmoplastic stroma surrounding tumor cells, suggesting that MuMo effectively deduces vital tumor information for predictions (Fig. 5a and Supplementary Fig. S8). In radiological CT scans, we found that MuMo was primarily concentrated in regions harboring lesions, as identified through gradient-weighted class activation mapping (Grad-CAM),²⁹ aligning with key areas of clinical concern in cancer diagnosis and treatment (Fig. 5b and Supplementary Fig. S9).

Subsequently, we evaluated the risk scores predicted by MuMo across various clinical information subgroups, including patient information, radiologically structured clinical reports, and pathologically structured clinical reports (Fig. 5c–l and Supplementary Fig. S10). We observed that in the anti-HER2 cohort, sex, degree of differentiation, Lauren type, and peritoneal metastasis were identified as key decision variables by the MuMo predicted risk scores. Among these, male patients were predicted to have higher risk scores compared to female patients (Mann–Whitney U test, $P = 0.041$; Fig. 5c). For patients with poorly differentiated tumors (Fig. 5d), MuMo assessed their risk probability as significantly higher

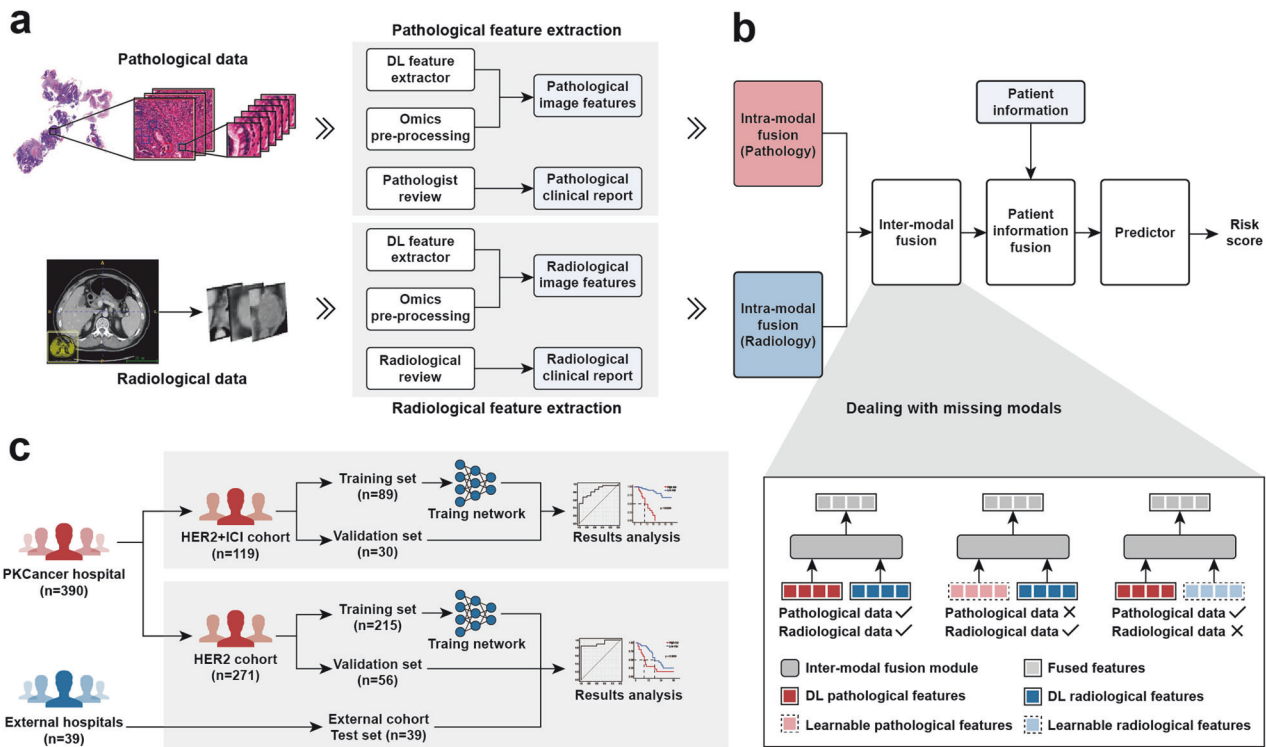


Fig. 2 Workflow of the Multi-Modal Model (MuMo) for predicting treatment response to patients with anti-HER2 GC. **a** Feature extraction process: Pathological WSIs and radiological CT scans were processed to extract deep and omics features, which were correlated with clinical reports provided by pathologists or radiologists. **b** Multi-modal information fusion process: MuMo employs intra-modal fusion modules to integrate image features and clinical reports from pathology and radiology to obtain enhanced features. These features were then amalgamated using an inter-modal fusion module, and the patient information was incorporated using a separate patient information fusion module. Subsequently, a predictor was used to predict the risk scores. MuMo can handle missing modalities by employing learnable modality features as placeholders. **c** Overview of experimental pipeline: Data were sourced from Beijing Cancer Hospital (PKCancer) and external hospitals. The patients were divided into anti-HER2 and anti-HER2 combined immunotherapy, as well as external cohorts. The anti-HER2 cohort was randomly divided into a training set to train the model and a validation set to tune its parameters. The final model with frozen parameters was used to analyze the results. Additionally, an external cohort was used as an independent test set to test the robustness of the model. In the anti-HER2 combined immunotherapy cohort, a similar analytical process was employed

compared to those with moderately differentiated tumors ($P < 0.0001$) and well-differentiated tumors ($P = 0.0003$). For Lauren classification (Fig. 5e), MuMo assessed that the diffuse type had the highest risk probabilities, being significantly higher compared to the intestinal type ($P < 0.0001$) and the mixed type ($P = 0.0038$). Conversely, the intestinal type had relatively lower risk, with the mixed type falling between the two in GC. Additionally, MuMo recognized patients with peritoneal metastasis as having significantly poorer responses than those without peritoneal metastasis ($P < 0.0001$; Fig. 5f). Furthermore, an increase in TILs was also seen to slightly reduce the predicted risk scores (Pearson correlation coefficient $r = -0.092$), indicating MuMo's awareness of the relationship between the abundance and activity of TILs and patient outcomes. In the anti-HER2 combined immunotherapy cohort, the presence of peritoneal metastasis also remained a significant decision variable for predicting high risk scores by MuMo ($P = 0.027$; Fig. 5i); additionally, MuMo identified patients undergoing second-line treatment as having poorer responses compared to those undergoing first-line treatment; similarly, MuMo also recognized a negative correlation ($r = -0.211$; Fig. 5l) between the abundance and activity of TILs and predicted risk probabilities, suggesting that a higher abundance of TILs is associated with better treatment responses. These analyses demonstrate that MuMo can extract appropriate knowledge from clinical reports and patient information to make accurately treatment response predictions, and its recognition of these significant decision variables aligns with current clinical findings,^{30,31} confirming MuMo's reliability and clinical relevance (Fig. 6).

DISCUSSION

This study demonstrated the accuracy and utility of multi-modal data analysis in predicting the response to anti-HER2 therapy in patients with HER2-positive GC. Our dataset, which is the largest available as confirmed by a thorough literature search, covers various types of medical information, including patient demographics, radiological CT scans containing multiple lesions, pathologic whole-slide images with different HER2 expression levels (0–3+), and structured clinical reports. Our proposed MuMo excels in predicting the response to both anti-HER2 therapy and anti-HER2 combined with immunotherapy in patients with HER2-positive GC. Importantly, the external validation of the predictive capabilities of MuMo in an independent cohort from other medical centers underscores its potential applicability in diverse clinical settings.

The advantage of MuMo lies in its efficient integration capabilities with multi-modal data, compared to the widely studied unimodal models.^{18,22,32–34} It utilizes a comprehensive and rich patient profile to improve the accuracy of treatment response predictions. Recent studies have started to investigate the use of multi-modal data in treatment response prediction for various cancers, including clear cell renal cell carcinoma,³⁵ non-small cell lung cancer,^{17,36,37} and hepatocellular carcinoma.³⁸ However, there is a notable scarcity of response prediction research on patients with HER2-positive GC and their response to anti-HER2 therapy. Moreover, current approaches mainly rely on simple integration methods, such as concatenation,³⁹ aligning and then averaging,⁴⁰ or multivariate machine learning analysis,^{35,38,41} which do not take

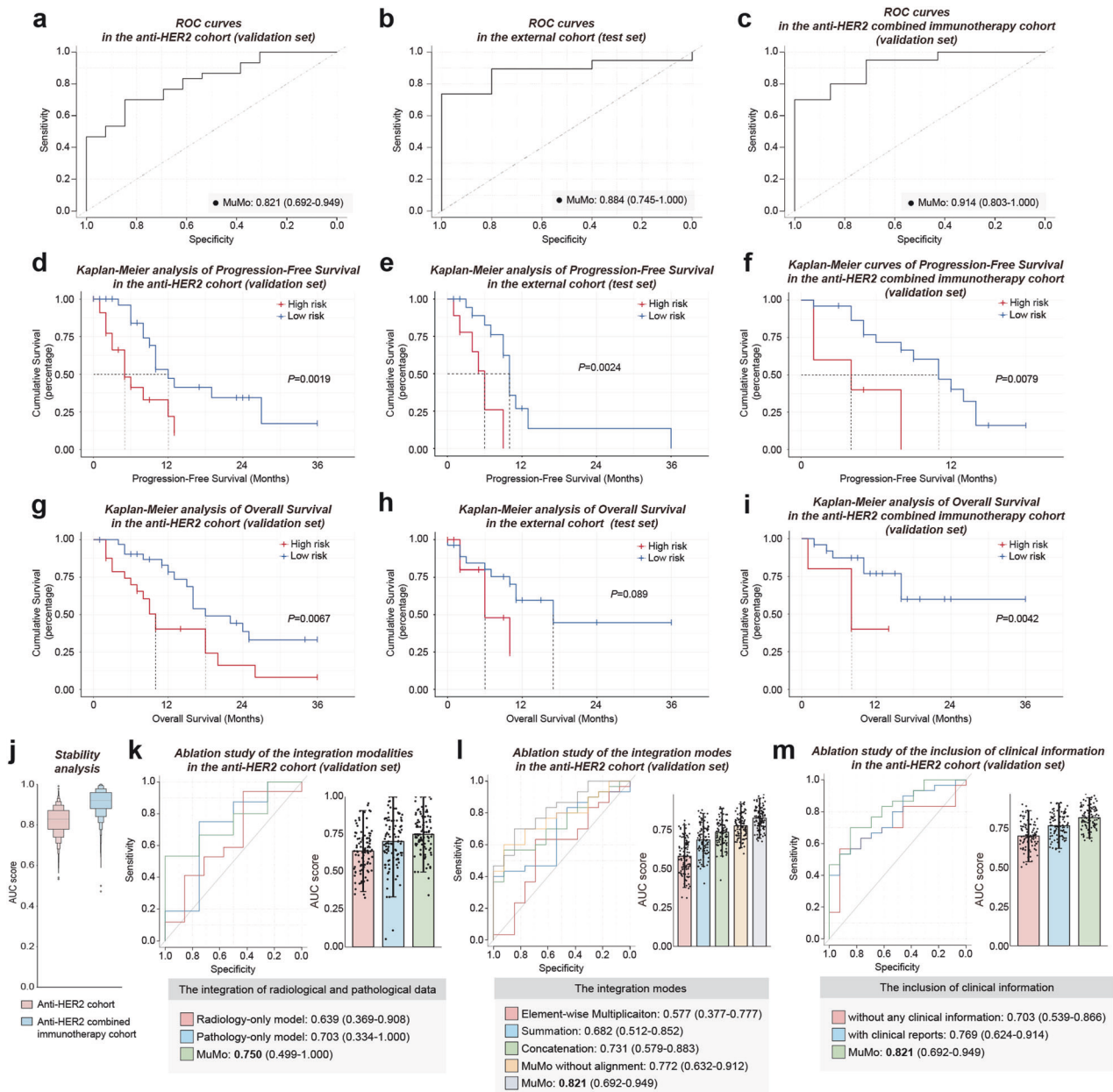


Fig. 4 Performance of the Multi-Modal Model (MuMo). **a–c** Receiver operating characteristic (ROC) curves display MuMo’s performance in predicting treatment responses, distinguishing between non-responders and responders. These curves pertain to the validation and test sets of the anti-HER2 cohort and the validation set of the anti-HER2 combined immunotherapy cohort. **d–f** Kaplan–Meier (KM) curves depict Progression-Free Survival (PFS) based on MuMo predictions. These curves were derived from the validation and test sets of the anti-HER2 cohort and the validation set of the anti-HER2 combined immunotherapy cohort. In the Kaplan–Meier analysis, patients were categorized into high-risk (red line) and low-risk (blue line) groups using the Youden index. The log-rank (Mantel-Cox) test was used to determine statistical significance, with a two-sided P -value of < 0.05 set as significant. **g–i** Kaplan–Meier (KM) curves show Overall Survival (OS) based on MuMo predictions for the validation and test sets of the anti-HER2 cohort and the validation set of the anti-HER2 combined immunotherapy cohort. **j** Stability analysis of MuMo in treatment response prediction. Ablation studies evaluating the integration of radiological and pathological data (**k**), the impact of various integration modes (**l**), and the inclusion of clinical information within MuMo (**m**). Error bars represent the 95% confidence intervals (CI) for the AUC scores

Another major potential of the MuMo is its ability to integrate more modalities in the future, which is key to realizing personalized treatment strategies. Our model consolidated diverse data sources, including radiological CT, pathological images, and clinical reports, and can incorporate even more modalities in the future, such as cancer biomarkers, gene expression, and lifestyle and health history information. Adding such data may significantly improve the model’s accuracy in predicting treatment response while also helping physicians formulate more targeted treatment

plans. For instance, by integrating gene expression data, we can gain deeper insight into patient’s pharmacogenomics, thereby optimizing drug selection and dosage adjustments.

The flexibility of the MuMo makes it an ideal platform for interdisciplinary collaboration, facilitating knowledge fusion between bioinformaticians, clinicians, and data scientists. Through such collaborations, MuMo can continuously assimilate the latest research discoveries and clinical feedback to iteratively update and refine its algorithms. Additionally, the

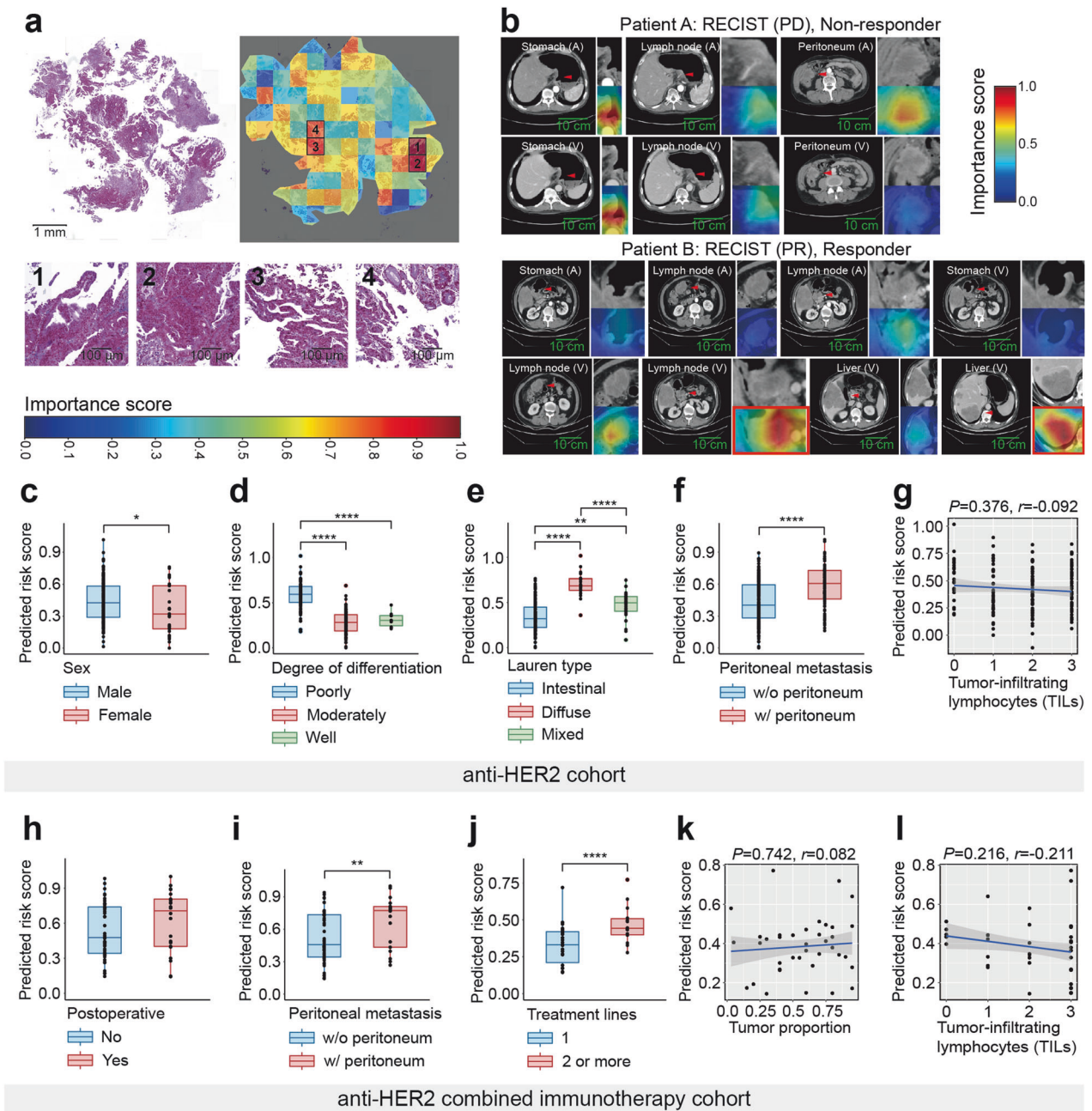


Fig. 5 Interpretability analysis of the Multi-Modal Model (MuMo). **a** Visualization of the importance scores of ‘bags’ on pathological whole-slide images. Darker red regions signify a higher contribution to the response prediction, whereas darker blue regions suggest a diminished influence. The second row shows the four most important bags on the slide image. **b** Visualization of attention maps on radiological lesion images using the Grad-CAM algorithm. Darker red regions signify heightened attention from MuMo, whereas darker blue regions denote reduced attention. The red bounding box emphasizes MuMo’s predominant focus on lymph node and liver tumors in this responder. **c–g** Evaluation of predicted risk scores across various clinical information subgroups in the anti-HER2 cohort. **h–l** Evaluation of predicted risk scores across various clinical information subgroups in the anti-HER2 combined immunotherapy cohort

MuMo framework is highly extensible and can be rapidly expanded to other cancer types and diagnostic markers. By undergoing targeted retraining with specific cohorts, MuMo can swiftly adapt to different cancers such as breast cancer or to other clinically relevant immunohistochemical (IHC) markers like ER, PR, and EGFR. This adaptability not only enhances the model’s utility across various oncological applications but also supports a more comprehensive approach to personalized medicine. Furthermore, we hope to explore the use of data collected from wearable devices and remote monitoring tools, such as patient activity levels and physiological responses,

which can provide the model with comprehensive health information for more accurate personalized treatment.

Although our preliminary findings using the MuMo are promising, we must acknowledge its limitations. First, although our dataset was collected from multiple medical centers and focused primarily on patients with HER2-positive GC, expansion is possible. Secondly, despite performing accurate treatment response predictions, the model still relies on human input for certain sub-tasks, such as requiring experts to annotate bounding boxes around lesions in radiological images and delineate regions with different HER2 expression levels in pathological slides.

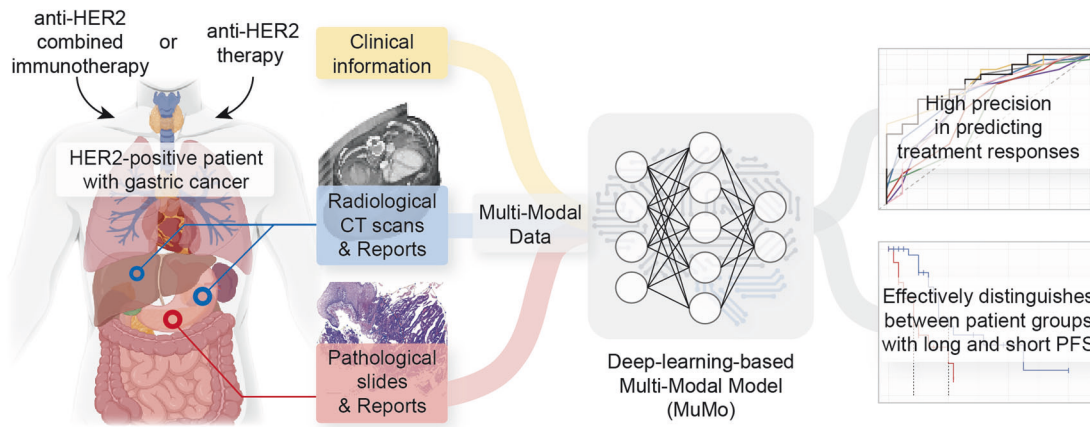


Fig. 6 Comprehensive overview of the multi-modal data analysis on HER2-positive patient with GC. This figure illustrates the step-by-step workflow of our research approach, from data collection through to the analysis techniques used. Key results are highlighted, demonstrating the highly prediction accuracy of treatment response of MuMo in HER2-positive patient with GC

In future studies, we plan to incorporate automation techniques, including the deployment of large language models and AI agents, to further reduce the need for human input and move closer to fully autonomous end-to-end treatment response prediction systems.

Summarily, MuMo represents a promising strategy for leveraging AI capabilities to improve response-predictive accuracy in patients with GC receiving anti-HER2 or anti-HER2 combined immunotherapy. By employing a comprehensive, multi-modal dataset, we are making significant strides toward realizing personalized treatment strategies. The model is an impressive testament to the potential integration of diverse modalities and AI, highlighting an exciting direction for future oncology research.

MATERIALS AND METHODS

Data collection

Our study included patients recruited between January 2007 and January 2023, who were divided into three distinct cohorts: the anti-HER2 cohort, comprising patients who received anti-HER2 therapy; the anti-HER2 combined immunotherapy cohort, comprising patients treated with anti-HER2 and anti-PD-1/PD-L1 ICI; and an external cohort, comprising patients from external hospitals who received anti-HER2 therapy. The study received approval from the Ethics Committees of Peking University Cancer Hospital, Nanfang Hospital, and Peking University Third Hospital (approval numbers 2020KT08, NFEC2017171, and D2021077, respectively). Informed consent was obtained from all participants or their legally authorized representatives. We asked pathologists and radiologists to annotate the medical images and provide structured clinical reports to ensure comprehensive data collection. Epstein-Barr Virus status was ascertained using in situ hybridization, employing probes targeting Epstein-Barr encoded RNA. Mismatch Repair status, a key factor in determining cancer behavior and treatment response, was assessed using IHC analysis to examine the expression levels of DNA mismatch repair proteins, specifically MLH1, MSH2, MSH6, and PMS2, following previously described methods.⁴²

Survival and response metrics

OS is defined as the time from diagnosis until either the death of the patient or the end of the follow-up period, whichever occurred first. PFS refers to the time from the start of treatment to disease progression, recurrence, or death, whichever occurred first. Responders were defined as patients who achieved the response evaluation criteria in solid tumors (RECIST) designation of complete response (CR), partial response (PR), or stable disease

(SD) with PFS exceeding the median PFS reported in KEYNOTE-811⁴ (8 months for the anti-HER2 cohort and 10 months for the anti-HER2 combined immunotherapy cohort). Non-responders included those with a RECIST designation of progressive disease (PD) or SD but did not exceed the median PFS reported in KEYNOTE-811. To maintain the integrity of our data and prevent data distortion, patients lost to follow-up (also referred to as censored cases) before reaching the median PFS threshold were excluded from the response prediction analysis. However, they were considered in the survival analysis to provide a more complete overview of patient outcomes.

Identifying tumor regions in pathological slides

Hematoxylin-Eosin (H&E) and IHC slides are prepared from consecutive tissue sections, typically cut at a thickness of 4 micrometers. This method ensures that the tissue morphology is essentially congruent between the H&E and IHC slides, facilitating precise mapping of HER2 status (as detailed in Supplementary Text S4) onto the H&E images. Based on these carefully aligned slides, experienced pathologists employed an Automated Slide Analysis Platform (ASAP version 1.9, <https://computationalpathologygroup.github.io/ASAP/>) to identify tumor regions within pathological whole-slide H&E images (Supplementary Text S5). These regions exhibited varying expression levels of HER2, including regions with HER2 = 0, HER2 = 1+, HER2 = 2+, and HER2 = 3+, effectively illustrating the heterogeneity of HER2 expression within the tumor (Supplementary Fig. S11 and Supplementary Table S9). This process requires the pathologist to meticulously outline the contours of each region, effectively marking the boundaries in a point-by-point manner while avoiding necrotic areas and normal glands. Concurrently, the pathologist summarized the information in H&E images to write structure pathological clinical report. During the entire process, a senior pathologist reviewed and affirmed all these results to ensure their precision and adherence to standard guidelines. Based on these annotation results, we standardized the bag-level pathological images through the Reinhard algorithm and white balance processing to unify the color distribution differences across different slides and centers (Supplementary Figs. S12, S13 and Supplementary Text S6).

Identifying tumor regions in radiological CT scans

For the radiological data, three radiologists employed the ITK-SNAP software (version 3.6.0, <http://www.itksnap.org>) to identify and annotate primary (GC) and metastatic lesions (liver, lymph nodes, spleen, bone, and soft tissue) within the 3D CT scans. The metastatic lesions were chosen according to the RECIST v1.1 criteria,⁴³ where radiologists selected a maximum of two target lesions per organ and a total of no more than five target lesions.

Then the radiologists used minimal bounding boxes to encompass the entire lesion as fully as possible. Simultaneously, the radiologists critically assessed the radiological data and wrote a structured radiological clinical report. A senior radiologist reviewed and validated all results throughout this process, ensuring their precision and adherence to standard guidelines. Based on these annotations, we initially computed dynamic windows for various lesion types and centers (Supplementary Fig. S14 and Supplementary Table S10), using them to normalize the corresponding radiological lesion images (Supplementary Fig. S15 and Supplementary Text S6).

Overall framework of MuMo

We developed a MuMo, a transformer-based model designed to predict treatment responses to anti-HER2 therapy and anti-HER2 combined immunotherapy (Fig. 2a, b). MuMo begins by extracting diverse features (Fig. 2a), including deep image features, omics features, and clinical reports, from different modalities such as pathological whole-slide images and radiological CT scans, using specialized feature extractors (details in subsequent subsections, Supplementary Fig. S16, and Supplementary Text S7). “Deep features” are image features derived from radiological lesion images and pathological word-level images using a deep learning model MnasNet⁴⁴ (detailed in Supplementary Text S7.1). This method effectively identifies complex patterns that are challenging to distinguish manually. “Omics features” consist of a wide range of radiomics features extracted by PyRadiomics library,⁴⁵ including first-order statistics, shape, texture, and higher-order statistical features, providing a comprehensive quantitative analysis of image data. MuMo then utilizes multi-modal fusion modules (Fig. 2b), including intra-modal fusion, inter-modal fusion, and patient information fusion, to effectively integrate information from different modalities, ensuring a comprehensive and precise analysis for response prediction (details in subsequent subsections, Supplementary Fig. S17, and Supplementary Text S8). Specifically, MuMo employs an intra-modal fusion module, that integrates features within the same modality, such as pathological image features, pathological omics features, and pathological clinical reports, to create a comprehensive set of modality-specific data. MuMo uses an inter-modal fusion module that amalgamates multi-modal features from different modalities into a unified feature. Importantly, this module is tailored to address instances of missing modalities by incorporating learnable modality-specific features, thereby ensuring the robustness of the model even with incomplete datasets. Finally, to enhance the precision of the response predictions, the model considers patient-level clinical information through the patient information fusion module and makes response predictions through a multi-layer perceptron with a softmax activation function.

Feature extraction in MuMo

In our study, we employed three distinct methodologies to extract diverse features, including deep features, omics features, and clinical reports, from available radiological and pathological data (Fig. 2a, Supplementary Fig. S16, and Supplementary Text S7).

In pathology, drawing insights from the prior study⁴⁶ on high-resolution pathological WSIs, we partitioned WSIs into larger segments known as “bags” within the annotated region-of-interest (ROI) areas for more focused analysis. These “bags” were further subdivided into smaller patches, referred to as “words” (Supplementary Fig. S18). Employing MnasNet, a recent lightweight convolutional neural network, enabled efficient conversion of these “words” into word-level deep features. From these, we extracted patient-level deep features using a bottom-up process (Supplementary Fig. S16a). Additionally, we generated pathological omics features from these “bags” using the PyRadiomics library. Pathological image features combine patient-level deep features with pathological omics features. Furthermore, the pathological clinical reports provided by

pathologists were mapped into embeddings using a predefined parameterless encoder (Supplementary Table S11).

In radiology, we began by preprocessing the CT scans to construct focused ROI radiological images following the radiologists’ annotations (Supplementary Fig. S19). Subsequently, we employed MnasNet to extract deep features from the ROI radiological images. Additionally, we segmented the ROI radiological images using a pre-trained lesion segmentor (Supplementary Fig. S20) and utilized the PyRadiomics library to derive the radiological omics features. We combined these two types of features to form radiological image features. Mirroring this approach in pathology, we converted the radiological clinical reports provided by radiologists into embeddings using a predefined parameterless encoder (Supplementary Table S12).

Multi-modal information fusion in MuMo

We developed intra-modal, inter-modal, and patient information fusion modules, each specifically designed to synthesize and integrate different types of data for enhanced treatment response prediction (Fig. 2b, Supplementary Fig. S17, and Supplementary Text S8).

The intra-modal fusion module begins by processing image features and structured clinical reports (Supplementary Fig. S17a). In this module, image features, comprising deep and omics features, are transformed into functional features, including query (Q), key (K), and value (V), through fully connected layers. Subsequently, a cross-attention layer⁴⁷ then merges the information from V and Q , according to the mutual interaction between Q and K . Simultaneously, clinical reports are transformed into embeddings using the parameterless encoder (Supplementary Tables S11–S13), forming another key (K') and value (V'). A similar cross-attention layer then merges the information from V' and Q . Finally, these two aggregated features are integrated with the original deep features via element-wise summation, producing an intra-modal fused feature. The outputs from this module for the pathological and radiological data are denoted as F_{path} and F_{rad} , respectively.

Second, the inter-modal fusion module takes the aggregated features F_{path} and F_{rad} and integrates them, operating under three clinical scenarios:

- I. All modality data are available (left part of Supplementary Fig. S17b): In cases where both radiological and pathological data are available for a patient, the corresponding features F_{path} and F_{rad} first pass through a fully connected layer. These are then divided into modal-specific (F_{path_s} and F_{rad_s}) and modal-agnostic (F_{path_a} and F_{rad_a}) features using dedicated functional layers. Subsequently, an alignment algorithm for modal-agnostic features aligns these features from different modalities via contrastive learning⁴⁸ (see details in Supplementary Text S9.3). An element-wise mean operation computes the averaged modal-agnostic feature. The two modal-specific features and the averaged modal-agnostic feature are then concatenated to form an inter-modal fused feature.
- II. Missing radiological data (middle part of Supplementary Fig. S17b): In this scenario, the radiological aggregated feature F_{rad} is absent. As a substitute, a learnable radiological feature F'_{rad} with the same dimensions as F_{rad} is used as a placeholder. The processing of pathological features F_{path} remains the same as that in scenario I. Only one modal-agnostic feature from the pathological data is generated, serving as the averaged modal-agnostic feature. The pathological modal-specific, averaged modal-agnostic, and learnable pathological features are concatenated to generate the inter-modal fused feature.
- III. Missing pathological data (right part of Supplementary Fig. S17b): This scenario mirrors scenario II but with the use of a learnable pathological feature F'_{path} as a placeholder. The radiological modal-specific and averaged modal-agnostic

features are concatenated with a learnable pathological feature to form the inter-modal fused feature.

The output feature of the inter-modal fusion, denoted as F_{inter} , served as the input for the patient information fusion module (Supplementary Fig. S17c). F_{inter} is first mapped onto a query (Q') via a fully connected layer. Patient information is initially encoded into embeddings using a parameterless encoder (Supplementary Table S13) and subsequently transformed into a key (K'') and value (V'') through fully connected layers. A cross-attention layer integrates patient information from F_{inter} , culminating in patient-level features for response prediction.

Experimental design

For the anti-HER2 cohort with 271 patients, collected from Peking University Cancer Hospital (PKCancer), we randomly split the cohort into a training set of 215 patients for deep learning model training and a validation set of 56 patients for hyperparameter optimization. The specifics of the training configurations are elaborated on in Supplementary Text S6 and S9. Upon completion of the training phase, we averaged the weights of the top seven best-performing models to create our final trained model. This model was subsequently used to further analyze the results (Fig. 2c). To test the generalizability of our model, we compiled an independent test set of 39 individuals from external hospitals, all of whom had received anti-HER2 therapy. The experimental setup for the anti-HER2 combined immunotherapy cohort, also sourced from the Peking University Cancer Hospital, mirrored that of the anti-HER2 cohort. This cohort was partitioned into a training set of 89 patients and a validation set of 30 patients. Analysis of the results was performed for this validation set.

Statistics and reproducibility

Sample sizes were determined based on the availability of suitable patient data that met the inclusion criteria. No statistical method was used to determine the sample size. To our knowledge, our collected dataset is the most comprehensive to date, covering multiple modalities for patients with GC receiving anti-HER2 therapy or anti-HER2 combined immunotherapy. The different distributions of data characteristics between responders and non-responders were evaluated using the two-sided Mann–Whitney U test for two continuous variables (number of metastatic lymph nodes, tumor proportion, and tumor-infiltrating lymphocytes) and the chi-square test for the remaining categorical variables. Survival functions were estimated using the Kaplan–Meier method, and survival distributions across groups were compared using the log-rank (Mantel–Cox) test. We used Levene’s test to assess consistent stability in treatment response prediction. The discriminative performance of the model was evaluated using the ROC–AUC. For statistical analyses, we used R (version 4.1.3) for survival functions, distributions, and stability assessments and Python (version 3.7.10) for model evaluation. We established a P -value threshold of < 0.05 to denote statistical significance in this study.

To ensure reproducibility, we have detailed our methodology in the Supplementary Materials, which covers feature extraction, fusion module operations, loss function definitions, experimental specifics, and evaluation metrics (Supplementary Figs. S16–S20, Supplementary Tables S10–S15, and Supplementary Texts S1–S3 and S6–S9). We also meticulously detailed our data management protocols, including data collection, annotation, and processing (Supplementary Figs. S1–S15 and Supplementary Tables S1–S9 and Supplementary Texts S4–S5). These procedures adhered to good clinical practice and data privacy regulations.

DATA AVAILABILITY

All data referenced in this study are available within the article and on GitHub at <https://github.com/czifan/MuMo>. Due to policy constraints, the raw radiological and

pathological data should be obtained by submitting a reasonable request to the corresponding author.

CODE AVAILABILITY

The source code of the proposed Multi-Modal Model (MuMo) is available on GitHub at <https://github.com/czifan/MuMo>. All R packages employed in this study can be found on CRAN (https://cran.r-project.org/web/packages/available_packages_by_name.html) or Bioconductor (<https://www.bioconductor.org/>).

ACKNOWLEDGEMENTS

We are grateful to Peng Bao and Mingze Yuan for their invaluable advice on writing. This work was supported by the National Natural Science Foundation of China (91959205 to L.S., U22A20327 to L.S., 82203881 to Y.C., 82272627 to X.T.Z., 7232018 to Y.S., 12090022 to B.D., 11831002 to B.D., 81801778 to L.Z.), Beijing Natural Science Foundation (7222021 to Y.C., Z200015 to X.T.Z.), Beijing Hospitals Authority Youth Programme (QML20231115 to Y.C.), Clinical Medicine Plus X-Young Scholars Project of Peking University (PKU2023LCXQ041 to Y.C. and L.Z.).

AUTHOR CONTRIBUTIONS

Zifan Chen, Yang Chen, and Li Zhang were responsible for experimental design and execution and also created the figures. Yu Sun, Lei Tang, Yajie Hu, and Meng He were involved in data collection, annotation, and report generation. Yu Sun and Lei Tang further ensured the accuracy and compliance of all data with standard guidelines through meticulous review. Zhiwei Li, Siyuan Cheng, Jiajia Yuan, Zhenghang Wang, Yakun Wang, Jie Zhao, Jifang Gong, Liying Zhao, Baoshan Cao, and Guoxin Li also contributed to data collection and partial data analysis. The initial manuscript was written by Zifan Chen, Yang Chen, Li Zhang, Yu Sun, and Lei Tang. Xiaotian Zhang, Bin Dong, and Lin Shen managed the project, provided experimental oversight, and made critical revisions to the manuscript drafts. All authors have read and approved the final manuscript.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41392-024-01932-y>.

Competing interests: The authors declare no competing interests.

REFERENCES

- Joshi, S. S. & Badgwell, B. D. Current treatment and recent progress in gastric cancer. *CA Cancer J. Clin.* **71**, 264–279 (2021).
- Cancer Genome Atlas Research, N. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
- Bang, Y. J. et al. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet* **376**, 687–697, (2010).
- Janjigian, Y. Y. et al. Pembrolizumab plus trastuzumab and chemotherapy for HER2-positive gastric or gastro-oesophageal junction adenocarcinoma: interim analyses from the phase 3 KEYNOTE-811 randomised placebo-controlled trial. *Lancet* **402**, 2197–2208 (2023).
- Yu, K. H., Beam, A. L. & Kohane, I. S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2**, 719–731 (2018).
- He, J. et al. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **25**, 30–36 (2019).
- Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022).
- Lipkova, J. et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell* **40**, 1095–1110 (2022).
- Boehm, K. M. et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat. Cancer* **3**, 723–733 (2022).
- Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning*, 8748–8763 (2021).
- Huang, Z. et al. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* **29**, 2307–2316 (2023).
- Ning, Z. et al. Relation-Induced Multi-modal shared representation learning for Alzheimer’s disease diagnosis. *IEEE Trans. Med. Imaging* **40**, 1632–1645 (2021).

13. Chen, R. J. et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging* **41**, 757–770 (2022).
14. Hu, C. et al. Deep learning radio-clinical signatures for predicting neoadjuvant chemotherapy response and prognosis from pretreatment CT images of locally advanced gastric cancer patients. *Int. J. Surg.* **109**, 1980–1992 (2023).
15. Hao, J. et al. PAGE-Net: interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. *Pac. Symp. Biocomput* **25**, 355–366 (2020).
16. Zhou, H. Y. et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat. Biomed. Eng.* **7**, 743–755 (2023).
17. Vanguri, R. S. et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat. Cancer* **3**, 1151–1164 (2022).
18. He, M. et al. Associations of subcutaneous fat area and systemic immune-inflammation index with survival in patients with advanced gastric cancer receiving dual PD-1 and HER2 blockade. *J. Immunother. Cancer* **11**, e007054 (2023).
19. Chen, Y. et al. Predicting response to immunotherapy in gastric cancer via multi-dimensional analyses of the tumour immune microenvironment. *Nat. Commun.* **13**, 4851 (2022).
20. van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nat. Med.* **27**, 775–784 (2021).
21. Aggarwal, R. et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit. Med.* **4**, 65 (2021).
22. Jin, C. et al. Predicting treatment response from longitudinal images using multi-task deep learning. *Nat. Commun.* **12**, 1851 (2021).
23. Adam, G. et al. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis. Oncol.* **4**, 19 (2020).
24. Mittermaier, M., Raza, M. M. & Kvedar, J. C. Bias in AI-based models for medical applications: challenges and mitigation strategies. *NPJ Digit. Med.* **6**, 113 (2023).
25. Kelly, C. J. et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* **17**, 195 (2019).
26. Baltrusaitis, T., Ahuja, C. & Morency, L. P. Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 423–443 (2019).
27. Lee, Y.-L., Tsai, Y.-H., Chiu, W.-C. & Lee, C.-Y. Multimodal prompting with missing modalities for visual recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14943–14952 (2023).
28. Allison, K. H. Prognostic and predictive parameters in breast pathology: a pathologist's primer. *Mod. Pathol.* **34**, 94–106 (2021).
29. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE International Conference on Computer Vision*, 618–626 (2017).
30. Feng, F. et al. Prognostic value of differentiation status in gastric cancer. *BMC Cancer* **18**, 865 (2018).
31. Wang, Z., Chen, J. Q., Liu, J. L. & Tian, L. Issues on peritoneal metastasis of gastric cancer: an update. *World J. Surg. Oncol.* **17**, 215 (2019).
32. Li, J. et al. CT-based delta radiomics in predicting the prognosis of stage IV gastric cancer to immune checkpoint inhibitors. *Front. Oncol.* **12**, 1059874 (2022).
33. Lu, L., Dercle, L., Zhao, B. & Schwartz, L. H. Deep learning for the prediction of early on-treatment response in metastatic colorectal cancer from serial medical imaging. *Nat. Commun.* **12**, 6654 (2021).
34. Krieg, C. et al. High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nat. Med.* **24**, 144–153 (2018).
35. Zhang, Y. & Pei, Z. Multi-omics features-based machine learning method improve immunotherapy response in clear cell renal cell carcinoma. *bioRxiv*, 2023.2011.2024.568360, (2023).
36. Farina, B. et al. Integration of longitudinal deep-radiomics and clinical data improves the prediction of durable benefits to anti-PD-1/PD-L1 immunotherapy in advanced NSCLC patients. *J. Transl. Med.* **21**, 174 (2023).
37. Yang, Y. et al. A multi-omics-based serial deep learning approach to predict clinical outcomes of single-agent anti-PD-1/PD-L1 immunotherapy in advanced stage non-small-cell lung cancer. *Am. J. Transl. Res.* **13**, 743–756 (2021).
38. Liu, Q. P. et al. Prediction of prognostic risk factors in hepatocellular carcinoma with transarterial chemoembolization using multi-modal multi-task deep learning. *EClinicalMedicine* **23**, 100379 (2020).
39. Wang, X. et al. Cancer immunotherapy response prediction from multi-modal clinical and image data using semi-supervised deep learning. *Radiother. Oncol.* **186**, 109793 (2023).
40. Puyol-Anton, E. et al. A multimodal deep learning model for cardiac resynchronisation therapy response prediction. *Med. Image Anal.* **79**, 102465 (2022).
41. He, M. et al. Deep learning model based on multi-lesion and time series CT images for predicting the benefits from anti-HER2 targeted therapy in stage IV gastric cancer. *Insights Imaging* **15**, 59 (2024).
42. Li, J. et al. Subcutaneous envafolelimab monotherapy in patients with advanced defective mismatch repair/microsatellite instability high solid tumors. *J. Hematol. Oncol.* **14**, 95 (2021).
43. Eisenhauer, E. A. et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).
44. Tan, M. et al. Mnasnet: Platform-aware neural architecture search for mobile. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2820–2828 (2019).
45. van Griethuysen, J. J. M. et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res* **77**, e104–e107 (2017).
46. Mehta, S. et al. End-to-End diagnosis of breast biopsy images with transformers. *Med Image Anal.* **79**, 102466 (2022).
47. Vaswani, A. et al. Attention is all you need. In *Proc. International Conferenc on Neural Information Processing Systems*, 5998–6008 (2017).
48. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proc. International Conference on Machine Learning*, 1597–1607 (2020).



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024

Devil is in the Queries: Advancing Mask Transformers for Real-world Medical Image Segmentation and Out-of-Distribution Localization

Mingze Yuan^{1,2,†}, Yingda Xia^{1,*}, Hexin Dong^{1,2}, Zifan Chen², Jiawen Yao¹, Mingyan Qiu¹, Ke Yan¹, Xiaoli Yin⁴, Yu Shi⁴, Xin Chen³, Zaiyi Liu³, Bin Dong^{2,5}, Jingren Zhou¹, Le Lu¹, Ling Zhang¹, Li Zhang²
¹Alibaba Group ²Peking University ³Guangdong Province People’s Hospital
⁴Shengjing Hospital ⁵Peking University Changsha Institute for Computing and Digital Economy

Abstract

Real-world medical image segmentation has tremendous long-tailed complexity of objects, among which tail conditions correlate with relatively rare diseases and are clinically significant. A trustworthy medical AI algorithm should demonstrate its effectiveness on tail conditions to avoid clinically dangerous damage in these out-of-distribution (OOD) cases. In this paper, we adopt the concept of object queries in Mask Transformers to formulate semantic segmentation as a soft cluster assignment. The queries fit the feature-level cluster centers of inliers during training. Therefore, when performing inference on a medical image in real-world scenarios, the similarity between pixels and the queries detects and localizes OOD regions. We term this OOD localization as *MaxQuery*. Furthermore, the foregrounds of real-world medical images, whether OOD objects or inliers, are lesions. The difference between them is less than that between the foreground and background, possibly misleading the object queries to focus redundantly on the background. Thus, we propose a query-distribution (QD) loss to enforce clear boundaries between segmentation targets and other regions at the query level, improving the inlier segmentation and OOD indication. Our proposed framework is tested on two real-world segmentation tasks, i.e., segmentation of pancreatic and liver tumors, outperforming previous state-of-the-art algorithms by an average of 7.39% on AUROC, 14.69% on AUPR, and 13.79% on FPR95 for OOD localization. On the other hand, our framework improves the performance of inlier segmentation by an average of 5.27% DSC when compared with the leading baseline nnUNet.

1. Introduction

Image segmentation is a fundamental task in medical image analysis. With the recent advancements in computer

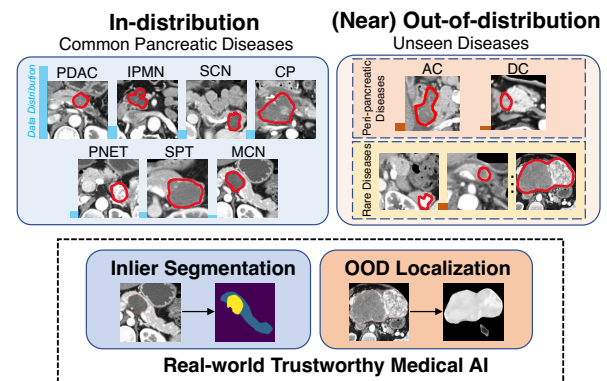


Figure 1. Real-world medical image segmentation. Real-world medical outliers (unseen, usually rare, tumors) are “near” to the inliers (labeled lesions), forming a typical near-OOD problem. A real-world medical OOD detection/localization model should focus more on subtle differences between outliers and inliers than the significant difference between foreground and background..

vision and deep learning, automated medical image segmentation has reached expert-level performance in various applications [3, 28, 54]. Most medical image segmentation methods are based on supervised machine learning that heavily relies on collecting and annotating training data. However, real-world medical images are long-tailed distributed. The tail conditions are outliers and inadequate (or even unable) to train a reliable model [35, 61, 63]. Yet, the model trained with inliers is risky for triggering failures or errors in real-world clinical deployment [43]. For example, in pancreatic tumor image analysis, a miss-detection of metastatic cancer will directly threaten life; an erroneous recognition of a benign cyst as malignant will lead to unnecessary follow-up tests and patient anxiety. Medical image segmentation models should thus demonstrate the ability to detect and localize out-of-distribution (OOD) conditions, especially in some safety-critical clinical applications.

Previous studies have made valuable attempts on medical OOD localization [48, 65], including finding lesions

* Corresponding author. (yingda.xia@alibaba-inc.com)

† Work was done during an internship at Alibaba DAMO Academy

apart from normal cases or simulating OOD conditions for model validation. However, the real-world clinical scenario, such as tumor segmentation, is more complex, where either in-distribution or OOD cases have multiple types of tumors. Establishing a direct relationship between image pixels and excessive semantics (types of tumors) is difficult for real-world medical image segmentation. Using this relationship to distinguish inliers and outliers is even more challenging. Fortunately, several works about Mask Transformers [5, 9] have inspired us to split segmentation as a two-stage process of per-pixel cluster assignment and cluster classification [57, 58]. A well-defined set of inlier clusters may greatly benefit in identifying the OOD conditions from the medical images. Therefore, we propose MaxQuery, a medical image semantic segmentation framework that advances Mask Transformers to localize OOD targets. The framework adopts learnable object queries to iteratively fit inlier cluster centers. Since the affinity between OODs and an inlier cluster center should be less than that within the cluster (between inliers and cluster centers), MaxQuery uses the negative of such affinity as an indicator to detect OODs.

Several recent works further define real-world medical image OOD localization as a near-OOD problem [43, 51], where the distribution gaps between inlier and OOD tumors are overly subtle, as shown in Fig. 1. Thus, the near-OOD problems are more difficult. Our pilot experiments show that the cluster centers redundantly represent the large regions of background and organ rather than tumors, compromising the necessary variability of the cluster assignments for OOD localization. To solve this issue, we propose the query-distribution (QD) loss to regularize specific quantities of object queries on background, organ, and tumors. This enforces the diversity of the cluster assignments, benefiting the segmentation and recognition of OOD tumors.

We curate two real-world medical image datasets of (pancreatic and liver) tumor images from 1,088 patients for image segmentation and OOD localization. Specifically, we collect consecutive patients' contrast-enhanced 3D CT imaging with a full spectrum of tumor types confirmed by pathology. In these scenarios, the OOD targets are rare tumors and diseases. Our method shows robust performance across two datasets, significantly outperforming the previous leading OOD localization methods by an average of 7.39% in AUROC, 14.69% in AUPR, 13.79% in FPR95 for localization, and 3.42% for case-level detection. Meanwhile, our framework also improves the performance of inlier segmentation by an average of 5.27% compared with the strong baseline nnUNet [24].

We summarize our main contributions as follows:

- To the best of our knowledge, we are the first to explore the near-OOD detection and localization problem in medical image segmentation. The proposed method has a strong potential for utility in clinical practice.

- We propose a novel approach, MaxQuery, using the maximum score of query response as a major indicator for OOD localization.
- A query-distribution (QD) loss is proposed to concentrate the queries on important foreground regions, demonstrating superior effectiveness for near-OOD problems.
- We curate two medical image datasets for tumor semantic segmentation/detection of real-world OODs. Our proposed framework substantially outperforms previous leading OOD localization methods and improves upon the inlier segmentation performance.

2. Related Work

Medical Image Segmentation and Diagnosis. U-Net [42] and its variants [31, 34, 37, 56, 64] have been promoting the development of medical image segmentation. A recent self-configuring U-Net (nnUNet) [24, 25] further surpassed existing approaches in various medical image segmentation tasks with minimal manual parameter tuning. Semantic segmentation serves as the core for downstream clinical tasks of disease detection [10], differential diagnosis [12, 61], survival prediction [54], therapy planning [46], and treatment response assessment [28]. Therefore, developing a reliable segmentation method is critical to improving safety in real-world clinical use. After the publication of Vision Transformers (ViTs) [16], integrating subsequent transformer blocks into the backbone of network architecture [7, 17, 18, 47] has been investigated. ViTs achieved improved results over traditional U-Net, particularly for multi-class semantic segmentation tasks. This work greatly focuses on exploring the real-world OOD localization detection problem over medical image segmentation. Current solutions provide limited performance, so we study a novel architecture combining Transformer and nnUNet for improving segmentation performance under clinical tasks, utilizing segmentation to detect and diagnose minority tumors [61].

Mask Transformers. Unlike using Transformers directly as network backbones for natural and medical image segmentation [36, 45, 53, 59, 62], Mask Transformers seek to enhance the CNN-based backbone with stand-alone transformer blocks. MaX-Deeplab [50] interprets object queries in DETR [5] as memory-encoded queries for end-to-end panoptic segmentation. MaskFormer [9] further applies this design to semantic segmentation by unifying the CNN and the transformer branches. Afterward, Mask2Former [8] technically improves over its predecessor. Recently, CMT-Deeplab [57] and KMaX-Deeplab [58] propose to interpret the queries as clustering centers and add regulatory constraints for learning the cluster representations of the queries. The design of Mask Transformers is intuitively suitable for medical image segmentation, especially for the

semantic segmentation and diagnosis of tumors. This task requires the network to be locally sensitive to image textures for tumor segmentation and can globally understand organ-tumor morphological information for tumor sub-type recognition. To our knowledge, we are the first to adapt Mask Transformers for medical image segmentation and further explore its usage of recognizing outliers via queries.

OOD Detection and Localization. OOD Detection aims to detect the out-of-distribution conditions (outliers) that are unseen in the training data. Maximal softmax probability (MSP) [21] serves as a strong baseline. After that, various approaches improved OOD detection from multiple aspects [13, 22, 29, 30]. These approaches focus on image-level OOD detection, and efforts have also been made to localize OOD objects or regions on a large image, e.g., urban driving scenes [4, 6, 14, 21, 26, 32, 39, 52]. Despite the advance of OOD detection and localization on natural images, its application on real-world medical images is challenging. Since the difference between foregrounds in real-world medical images is subtle, their OOD detection/localization becomes a typical near-OOD problem [15, 38, 41, 51]. Therefore, the existing OOD solutions could hardly be recommended for clinical practice [40, 48, 65]. Recent work, HOD [43], paces one step toward real-world OOD detection of rare diseases in dermatology classification.

3. Method

In this section, we first provide an overview of our method and then describe our proposed query-distribution (QD) loss and MaxQuery framework for OOD localization.

3.1. Method Overview

Medical image segmentation aims to segment an image into multiple regions representing anatomical objects of interest. Here, we focus on 3D medical image $\mathbf{X} \in \mathbb{R}^{H \times W \times D}$, and use a segmentation model to partition it into K category-labeled binary masks,

$$\mathbf{G} = \{\mathbf{G}_i\}_{i=1}^K, \quad (1)$$

where $\mathbf{G}_i \in \{0, 1\}^{H \times W \times D}$ is the ground truth mask that belongs to the i -th class, and $\sum_{i=1}^K \mathbf{G}_i = \mathbf{1}^{H \times W \times D}$. In our problem, class 1 refers to background, class 2 stands for specific organ, and the others for tumors. Since the real-world medical image dataset has a long-tail distribution in quantity, its segmentation task should be divided into supervised inlier segmentation and pixel-level OOD localization.

Inlier Segmentation. As shown in Fig. 2, we build our model with a CNN backbone to extract per-pixel features $\mathbf{P} \in \mathbb{R}^{H \times W \times D \times C}$ and a transformer module. The transformer module gradually updates a set of learnable object queries,

$\mathbf{C} \in \mathbb{R}^{N \times C}$, to meaningful mask embedding vectors via cross attention between object queries and per-pixel features,

$$\mathbf{C} \leftarrow \mathbf{C} + \underset{N}{\operatorname{argmax}}(\mathbf{Q}^c (\mathbf{K}^p)^T) \mathbf{V}^p, \quad (2)$$

where the superscripts c and p represent query and pixel features, respectively. We also adopt cluster-wise argmax from KMax-DeepLab [58] to substitute spatial-wise softmax in the original cross attention settings.

Inspired by recent works on cluster analysis of mask transformers [57, 58], we consider semantic segmentation as a two-stage cluster analysis process. First, all pixels are assigned into different clusters. The mask embedding vectors \mathbf{C} from the transformer module are formulated as the cluster centers. The product \mathbf{R} of \mathbf{C} and \mathbf{P}^T represents the query response, which expresses the similarity between each pixel and cluster centers. Then, we use the query-wise softmax activation on query responses \mathbf{R} to generate a mask prediction, which encourage the exclusiveness of cluster assignment. The mask prediction (cluster assignment) \mathbf{M} is defined as,

$$\mathbf{M} = \underset{N}{\operatorname{softmax}}(\mathbf{R}) = \underset{N}{\operatorname{softmax}}(\mathbf{C}\mathbf{P}^T). \quad (3)$$

Notably, different from the sigmoid activation used in [8, 9], the query-wise softmax activation could better guide the object queries (cluster centers) to focus on different regions of the image and encourage diversity in real-world medical image segmentation.

Secondly, the grouped pixels are classified under the guidance of cluster classification. We evaluate the cluster centers \mathbf{C} via a multi-layer perceptron (MLP) to predict the K -channel cluster classifications $\mathbf{C}_K \in \mathbb{R}^{N \times K}$ for all N clusters. We then aggregate the cluster assignments \mathbf{M} of grouped pixels and their classifications \mathbf{C}_K for the final semantic segmentation,

$$\mathbf{Z} = (\mathbf{C}_K)^T \mathbf{M}, \quad (4)$$

where $\mathbf{Z} \in \mathbb{R}^{K \times H \times W \times D}$ represents the final logits. To supervise the final segmentation, we combine the classic segmentation loss and a novel QD loss between final output \mathbf{Z} and ground truth \mathbf{G} , more details in Sec 3.2.

OOD Localization. To further segment abnormal regions unseen in training images, an OOD localization process is required when performing inference on a test image. Formally, given a test image $\mathbf{X} \in \mathbb{R}^{H \times W \times D}$, OOD localization evaluates the query response to find the maximal one that represents the similarity between the pixel and its assigned cluster center. Then, the model can generate an pixel-wise anomalous score map $\mathbf{A} \in [0, 1]^{H \times W \times D}$, where $\mathbf{A}_i = 1$ and $\mathbf{A}_i = 0$ represent that i -th pixel in \mathbf{X} belongs to an OOD class and an in-distribution class, respectively. More details of this novel OOD localization (MaxQuery) is in Sec 3.3

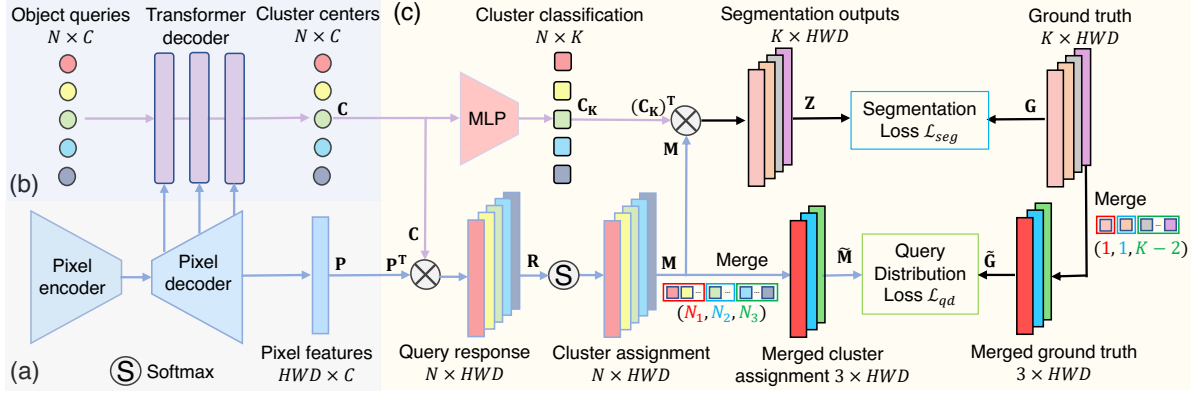


Figure 2. Overview of our proposed framework. (a) A CNN backbone for image segmentation, here we use nnUNet [24]; (b) A transformer decoder iteratively updates the object queries to fit the inlier cluster centers; (c) A two-stage cluster analysis: 1) cluster assignment groups the pixels based on the affinity between pixel features and cluster centers; 2) cluster classification guides the grouped pixels to generate segmentation logits. The overall segmentation is supervised by a classic segmentation loss and a novel query-distribution loss.

3.2. Managing Cluster Distribution with QD Loss

Classic segmentation loss serves as an important learning target of our model. We combine the Cross-Entropy and Dice losses between final output Z and ground truth G in Eq. (1) as the segmentation loss, *i.e.*, $\mathcal{L}_{seg} = \ell_{ce} + \ell_{dc}$. However, when only using classic segmentation loss, object queries focus majorly on the background and organs rather than the tumors. The significant difference between foreground and background greatly distracts the model from focusing on subtle differences between OOD objects and inliers. As later shown in an example in Fig 6, some queries may even have mixed representation on background and foreground which is an unsatisfactory phenomenon for discriminative cluster learning. Therefore, we propose query-distribution (QD) loss to manipulate the object queries and guide them to focus on the foreground, especially the tumors, and encourage concentrated cluster learning. The key idea is to use ground-truth $G \in \mathbb{R}^{K \times HWD}$ to supervise the cluster assignment probability maps M . This motivation also benefits OOD localization as introduced in Sec 3.3.

We thus divide the N channels into three groups, including N_1, N_2, N_3 queries, for background, organ and tumor regions, respectively. Our goal is to associate the first N_1 channels of M (representing the assignment probabilities of the first N_1 cluster centers) with the background class G_1 , the next N_2 channels with the organ class G_2 , and the last N_3 channels with the tumor classes $\sum_{i=3}^K G_i$. We define the merged cluster assignments \tilde{M} and class labels \tilde{G} as the following,

$$\begin{aligned} \tilde{M} &= (\tilde{M}_1, \tilde{M}_2, \tilde{M}_3) \in \mathbb{R}^{3 \times HWD} \\ &= \left(\sum_{i=1}^{N_1} M_i, \sum_{j=1}^{N_2} M_{N_1+j}, \sum_{k=1}^{N_3} M_{N_1+N_2+k} \right), \end{aligned} \quad (5)$$

$$\tilde{G} = (\tilde{G}_1, \tilde{G}_2, \tilde{G}_3) = (G_1, G_2, \sum_{i=3}^K G_i) \in \mathbb{R}^{3 \times HWD}, \quad (6)$$

where the merged \tilde{M} are still probability distributions in each spatial position, *i.e.*, $\sum_{j=1}^3 \tilde{M}_j = \mathbf{1}^{H \times W \times D}$.

Finally, we formulate the QD loss as the negative log likelihood loss between \tilde{M} and \tilde{G} ,

$$\mathcal{L}_{qd} = - \sum_{j=1}^{HWD} \sum_{i=1}^3 \tilde{G}_{ij} \log \tilde{M}_{ij}, \quad (7)$$

which draws strict boundaries between different types of cluster assignments (\tilde{M}_1, \tilde{M}_2 , and \tilde{M}_3) based on the ground truth. The final loss function \mathcal{L} is a combination of segmentation loss \mathcal{L}_{seg} and QD loss \mathcal{L}_{qd} with a balance weight λ , formulated as,

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{qd}. \quad (8)$$

3.3. Localizing OOD Regions with MaxQuery

Given a test image $X \in \mathbb{R}^{H \times W \times D}$, our mask transformer will yield the pixel-level query response $R \in \mathbb{R}^{N \times H \times W \times D}$, representing the affinity of pixel feature and cluster centers. The maximal query response of one pixel then represents the similarity between the pixel and its assigned cluster center. Intuitively, the maximal query response of outliers should be smaller than inliers. We therefore adopt the negative of maximal query response in Eq. (3) as the pixel-wise anomaly score, called MaxQuery, *i.e.*,

$$A = - \max_N R, \quad (9)$$

where $R \in \mathbb{R}^{N \times H \times W \times D}$ represents the query response matrix and $A \in \mathbb{R}^{H \times W \times D}$ indicates the anomaly score

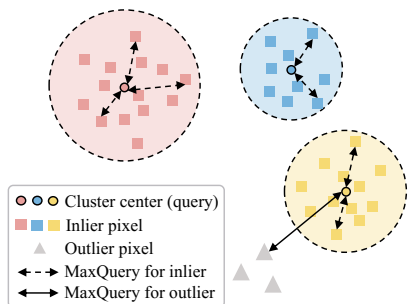


Figure 3. Illustration of how MaxQuery works. MaxQuery, *i.e.*, the negative of maximal query response, reflects the distance of the pixel and its assigned cluster center. MaxQuery of the inlier (dotted arrow) is usually smaller than that of the outlier (solid arrow) and thus is able to identify the anomalous/OOD pixels.

map. The anomaly score can be further normalized into $[0, 1]$ by min-max normalization. Figure 3 illustrates the capability of MaxQuery for OOD pixels identification. The subscript N means that we perform maximum operation on the query dimension. We add a minus sign because when the maximal query response of a pixel is larger, it is less likely to be an OOD pixel.

In addition, we compare the results of anomaly score maps according to the maximum of query responses \mathbf{R} (pre-softmax, $\mathbf{A} = -\max_N \mathbf{R}$) and cluster assignments \mathbf{M} (post-softmax, $\mathbf{A}' = -\max_N \mathbf{M}$). \mathbf{A} greatly outperforms \mathbf{A}' , since if an inlier pixel is evenly close to multiple cluster centers, the maximal score in \mathbf{M} can be very low and easily be mis-classified as an outlier. But with maximum query response \mathbf{R} (pre-softmax), the score is still high enough for an indication of inlier. Thus we choose the maximal query response to imply the anomalous regions.

4. Experiments

4.1. Datasets and Experiment Setting

We collect two datasets, *i.e.*, pancreas and liver tumor segmentation datasets, which includes contrast-enhanced 3D CT scans from consecutive patients before treatment. We register the multi-phase CT scans into arterial late and venous phase using DEEDS [19], respectively. All types of tumors are confirmed by pathology, except for cysts in the liver (confirmed by a radiologist specialized in liver imaging). All tumors are annotated slice-by-slice manually on the CT phase with the best tumor visibility by experienced radiologists specialized in specific diseases. The organ (pancreas or liver) in each dataset is first annotated automatically by a self-learning approach [60] trained on public datasets (*e.g.*, Medical Decathlon [1]) and then edited by engineers.

Pancreatic Multi-type Tumors dataset contains 661 patients. Every patient has five phases of CT scans: noncontrast, arterial-early, arterial-late, venous, and delay. The

median spacing is $3 \times 0.419 \times 0.419$ mm. According to previous clinical studies about pancreatic tumor classification [11, 44], we assign the seven most common conditions (PDAC, PNET, SPT, IPMN, MCN, CP, and SCN) as inliers, and allocate AC, DC, and “other” as outliers. We randomly split 590 inlier data into 378(64%) training, 94(16%) validation, and 118(20%) testing, and leave out all 71 outlier data for OOD testing.

Liver Multi-type Tumors dataset contains 427 patients. Each patient has three phases of CT scans: noncontrast, arterial, and venous. The median spacing is $3 \times 0.760 \times 0.760$ mm. Following [55], we assign the five most common conditions (HCC, ICC, metastasis, hemangiomas, and cyst) as inliers, and allocate hepatoblastoma, FNH, and “other” as outliers. Similarly, We randomly split 327 inlier data into 209(64%) training, 52(16%) validation, and 66(20%) testing, and leave out all 100 outlier data for OOD testing. Notice that the “other” class in both datasets contains multiple rare diseases, reflecting the long-tailed distribution of real-world disease incidence.

4.2. Implementation & Evaluation Metrics

Network Architecture. We use the current benchmark model in medical image segmentation, nnUNet [24], as a CNN backbone, which consists of a pixel encoder and a pixel decoder with skip connections. We adopt four transformer decoder blocks, and each takes pixel features with output stride 32, 16, 8, and 4, respectively. The self-attention layer in the block has 8 heads. Since medical image segmentation is sensitive to local textures, we add a decoder block for output stride 4 compared with previous works [50, 58]. To increase numerical stability, we add an InstanceNorm [49] layer and a LayerNorm [2] at the end of pixel-level and transformer decoder modules, respectively.

Training and Testing. Each CT scan is resampled into the median spacing per tumor dataset (*e.g.*, $3 \times 0.419 \times 0.419$ mm for the pancreatic dataset) and normalized into zero mean and unit variance. Our model is trained using a batch size of 2 on one GPU (with $28 \times 192 \times 320$ patch size for pancreatic, $40 \times 192 \times 224$ for liver). We adopt the drop path [23] strategy with a probability of 0.2 for regularization. During training, extensive data augmentation is utilized on-the-fly [24] to improve the generalization, including random rotation and scaling, elastic deformation, additive brightness, and gamma scaling. The network is trained with RAdam [33] with the initial learning rate as 1×10^{-4} and a polynomial learning rate decay. We first pre-train the nnUNet backbone for 1000 epochs and fine-tune the whole architecture jointly for another 200 epochs. During finetuning, we keep the backbone weights fixed for the first 50 epochs, and then set it with a learning rate multiplier of 0.1 for the next 150 epochs. The number of object queries (*i.e.*, cluster centers) N is 32, and the query distri-

Methods	Pancreatic %				Liver %			
	OOD Localization			OOD _{case}	OOD Localization			OOD _{case}
	AUROC↑	AUPR↑	FPR ₉₅ ↓	AUC↑	AUROC↑	AUPR↑	FPR ₉₅ ↓	AUC↑
MC Dropout [27]	49.08	11.47	84.60	72.91	39.61	16.05	91.13	34.05
MSP [21]	53.81	13.44	86.44	73.38	75.14	25.27	70.04	66.76
MaxLogit [20]	58.46	21.93	83.68	73.42	78.60	35.47	48.73	65.68
SynthCP [52]	69.86	26.50	66.65	68.43	74.93	34.03	57.91	63.34
SML [26]	56.10	30.44	77.81	62.26	86.64	44.59	31.04	63.85
Ours (w/o \mathcal{L}_{qd})	63.54	25.25	67.09	74.87	74.95	42.31	53.52	65.91
Ours	82.52	55.60	46.19	77.97	88.75	48.80	23.93	69.04

Table 1. OOD localization and case-level OOD detection performance on *Pancreatic Tumors* and *Liver Tumors*. Our proposed method achieves state-of-the-art OOD detection performance at both pixel level and case level. All the methods are implemented based on the nnUNet [24] backbone. (OOD_{case}: case-level OOD detection.)

Methods	Pancreatic %								Liver %					
	PDAC	IPMN	PNET	SCN	CP	SPT	MCN	Avg.	HCC	ICC	Meta.	Heman.	Cyst	Avg.
nnUNet [24]	65.65	27.60	32.59	36.46	23.33	31.73	30.96	35.47	57.22	28.16	52.81	77.55	46.49	52.45
Ours (w/o \mathcal{L}_{qd})	65.87	28.3	32.43	40.63	28.93	30.77	30.89	36.84	60.91	30.58	53.21	78.47	46.42	53.92
Ours	67.91	46.92	32.07	42.51	31.36	42.67	28.97	41.77	67.61	30.78	60.40	77.07	47.61	56.69

Table 2. Inlier segmentation Dice scores (%) on *val* set of *Pancreatic Tumors* and *Liver Tumors* (all methods report results with final checkpoint). Compared with the benchmark model (nnUNet [24]) in medical image segmentation, our method noticeably outperforms the strong baseline for the task of inlier tumor segmentation. See the Appendix for other baselines.

bution (N_1, N_2, N_3) is set as (16, 4, 12). We follow KMax-DeepLab [58] to directly add deep supervision on the attention map of (k -means) cross attention to align it with the final segmentation after the segmentation output head. The loss weight λ for QD loss is 0.1.

Evaluation Metrics. For OOD localization, we follow the standard metrics for anomaly segmentation [14, 26, 52]. We compute the area under receptive-operative curve (AUROC) and the area under precision-recall curve (AUPR). We also report FPR at the TPR level of 0.95 (FPR95) as OOD localization metrics since the false positive rate is safety-critical in clinical practice. For case-level OOD detection, we compute the average of anomaly scores in predicted tumor regions as the case-level anomaly score and choose AUC as the case-level OOD detection metric. Meanwhile, we report the average Dice Score of inlier tumors to evaluate the segmentation performance on inlier classes.

Baselines. For OOD localization, we compare our work with a series of representative anomaly segmentation methods in multiple aspects, including uncertainty statistics-based (MSP [21], MaxLogit [20], SML [26]), Bayesian deep learning-based (MC Dropout [27]) and image re-synthesis-based (SynthCP [52]) methods. All of them are implemented using nnUNet [24] backbone. For inlier segmentation, we compare our work with the benchmark model (nnUNet [24]) and previous leading model (SwinUNETR [47]), as well as UNet [42], UNet++ [64] and TransUNet [7], implemented by their officially released code and pre-trained model with same settings.

4.3. Main Results

Comparisons on the real-world datasets, including *Pancreatic Tumors* and *Liver Tumors*, are summarized in Tabs. 1 and 2. We also present visualization examples in Figs. 4 and 5 to better understand the role of object queries in our proposed mask transformer and compare different anomaly segmentation methods.

Pancreatic Tumors. In Tab. 1, we compare MaxQuery with other baselines on *Pancreatic Tumors*. Our framework shows the best performance in all metrics. Specifically, our framework outperforms the previous best method SML [26] by a large margin of 12.66% in AUROC, 25.16% in AUPR, 20.46% in FPR95 for OOD localization, and 4.55% in AUC for case-level OOD detection. For qualitative analysis, we present four visual examples from *Pancreatic Tumors* by visualizing the anomaly score map of MSP [21], MaxLogit [20], SML [26], and ours. As shown in Fig. 4, our method maintains a high anomaly score in the OOD pixels (outlier tumor), while a low anomaly score in the in-distribution pixels (organ). Moreover, the previous methods underestimate the anomalous score map. They tend only to highlight the boundaries of the OOD region, but our method preserves a high anomalous score on the entire OOD region.

In Tab. 2, our segmentation performance for inliers surpasses nnUNet by 6.30% in DSC. These improvements demonstrate that our framework can simultaneously detect common diseases with high accuracy and identify rare dis-

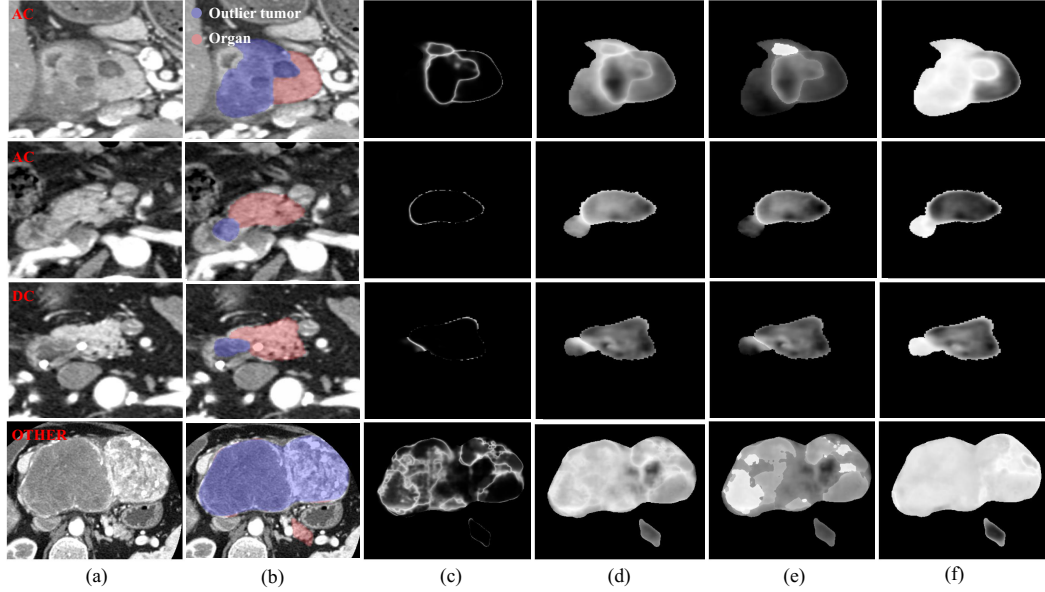


Figure 4. Visualization results of anomaly score map for OOD localization on *Pancreatic Tumors*: (a) CT slice, (b) ground truth (red: pancreas, blue: outlier tumor), (c) MSP [21], (d) MaxLogit [20], (e) SML [26] and (f) Ours. The grayscale level indicates the anomaly score. Our approach maintains a high anomaly score in the OOD pixels (outlier tumor), while a low anomaly score in the in-distribution pixels (organ). The four cases are selected from three different unknown diseases to show our method’s robustness to tumor type.

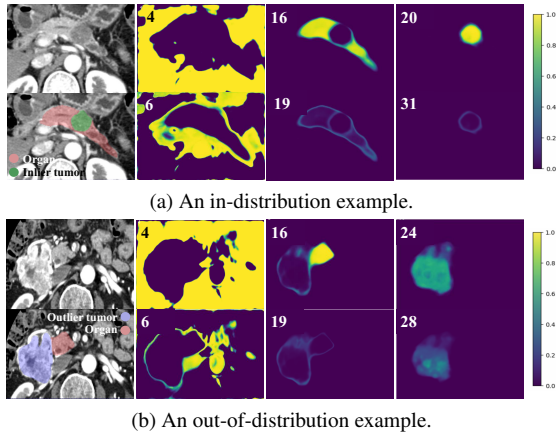


Figure 5. Visual examples of cluster assignments for (a) an in-distribution and (b) an out-of-distribution (OOD) sample. From left to right: (Column 1) image and ground truth with red: organ, green: inlier tumor, blue: outlier tumor; (Columns 2-4) representative object queries for background (C2), organ (C3) and tumor (C4), respectively. Query IDs are at the upper-left corners.

eases in pixel-level localization and case-level diagnosis without requiring very large data samples. (Other baselines can be found in the Appendix.)

We visualize the mask predictions of in-distribution and OOD examples to illustrate the working mechanism of object queries as cluster centers and how MaxQuery identifies the OOD condition. As shown in Fig. 5, for either in-

distribution or OOD example, the background and organ regions are confidently activated by specific queries (Queries 4 and 6 for background, Query 16 for the target organ). Interestingly, regions with distinguishing features, such as the aorta or other abdominal organs, are not activated by the major cluster center (Query 4) but by an independent center (Query 6). This supports that the queries gradually converge to different meaningful centers. Furthermore, the corresponding queries of specific in-distribution tumors usually concentrate at a single center (Query 20 in Fig. 5a). Yet, queries corresponding to the OOD tumors seem to split into multiple centers with lower responses (Query 24 and 28 in Fig. 5b). The visual examples fulfill the motivation of the proposed MaxQuery that no inlier cluster centers can dominantly fit the OOD pixels.

Liver Tumors. Table 1 also shows the quantitative result on *Liver Tumors*. Our method outperforms the baselines in all evaluation metrics. Note that SML [26] improves the performance in OOD localization while dropping its performance in case-level OOD detection compared with MaxLogit [20], whereas our method performs well in both pixel and case level. Particularly, our method reaches a significantly lower FPR95 of 23.93% compared with previous approaches, which is crucial to localizing the OOD regions in medical scenarios. As shown in Tab. 2, our segmentation performance for inliers surpasses nnUNet by 4.24% in DSC. The qualitative analysis on *Liver Tumors* is in the Appendix.

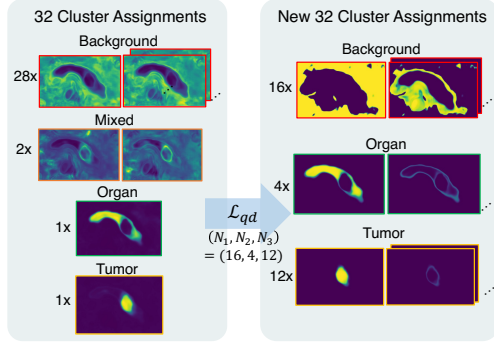


Figure 6. The effect of QD loss by visualizing the cluster assignment maps of the 32 queries on an inlier. **Left:** without QD loss, most queries redundantly focus on the background and some queries mix the background with foreground. **Right:** after using QD loss, we can manage the query distribution on the background, organ, and tumor with better separation. The clear boundaries and high responses shows that QD loss encourages discriminative representation learning of the queries which will benefit both segmentation and OOD localization.

4.4. Ablation Study

The Effect of the Query-Distribution Loss. Without the QD loss, the mean inlier tumor DSC of our framework increases only by a small margin compared to the nnUNet [24] baseline (Tab. 2). Fig. 6 presents query visualizations to show benefits from query-level guidance. Most queries redundantly represent the large and heterogeneous region of the background rather than the tumors without the QD loss (Fig. 6 left). With the QD loss, our framework is manipulated to provide fixed resources (queries for tumors) on distinguishing subtle differences of foregrounds for a near-OOD problem (Fig. 6 right). Final results are thus further improved on all metrics by large margins using QD loss (Tabs. 1 and 2). The results reveal that managing the object queries with QD loss contributes to masking transformers to improve both segmentation and OOD localization/detection performance.

The Distribution of Queries. We also perform an in-depth analysis of query distribution, as shown in Tab. 3. Our method shows robustness to different settings of query distribution. On all settings, our method outperforms the previous leading method, SML [26], by a large margin in OOD localization and inlier segmentation. Eventually, we choose the hyper-parameter (N_1, N_2, N_3) as (16, 4, 12).

Pre-softmax versus Post-softmax for MaxQuery. As shown in Tab. 4, MaxQuery with pre-softmax score **R** exceeds the one with post-softmax **M** by 21.62% in AUPR for OOD localization, which agrees with our explanation in Sec. 3.3.

Query-level versus Category-level Anomaly Score. The debate of pre-softmax versus post-softmax corresponds

Query Dist. (N_1, N_2, N_3)	AUROC \uparrow	AUPR \uparrow	FPR \downarrow	DSC _{inlier} \uparrow
SML [26]	56.10	30.44	77.81	35.47
(8, 4, 20)	84.44	51.32	42.10	36.50
(8, 20, 4)	83.73	49.76	43.32	39.79
(16, 4, 12)	82.52	55.60	46.90	41.77
(20, 4, 8)	85.66	55.17	37.24	38.19
(24, 4, 4)	86.41	52.58	33.70	39.43

Table 3. Ablation study on the distribution of queries. (DSC_{inlier}: mean Dice Score of inlier tumors.)

Level	Softmax	AUROC \uparrow	AUPR \uparrow	FPR95 \downarrow
Category	post	58.14	16.28	79.29
	pre	52.70	24.59	88.40
Query	post	76.88	33.98	55.82
	pre	82.52	55.60	46.19

Table 4. Comparison of category- and query-level anomaly scores. With the same network, the query-level anomaly scores show superiority over the category-level ones for OOD localization. Meanwhile, MaxQuery from the pre-softmax query-level scores outperforms that from post-softmax ones.

to the one of MaxLogit [21] versus MSP [20]. Specifically, MSP calculates the post-softmax score in the final category level, while MaxLogit calculates the pre-softmax one. Unlike MSP and MaxLogit, our MaxQuery produces an anomalous score at the query level. For a fair comparison, we apply MSP and MaxLogit based on the Mask transformer we used in our model. As shown in Tab. 4, MaxQuery (post-softmax) outperforms MSP (category, post-softmax) by 17.70% and MaxQuery (pre-softmax) exceeds MaxLogit (category, pre-softmax) by 32.01% in AUPR. This comparison indicates the superiority of our query-level anomaly score over the category-level ones.

5. Conclusion

Processing a large collection of medical imaging data with long-tailed distributions has always been challenging. The significant performance improvement of our method on two real-world datasets validates its effectiveness. This result proves that interpreting segmentation as (query) cluster assignment is valid and effective. Our novel MaxQuery and QD loss are also evidently helpful for inlier segmentation and (near-)OOD detection/localization, performing in practical scenarios. We believe that the proposed method has the good potential to further boost the adoption of medical image segmentation in designing various clinical applications.

Acknowledgement

This work was supported by Alibaba Group through Alibaba Research Intern Program. Bin Dong was partly supported by NSFC 12090022.

References

- [1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature Communications*, 13(1):1–13, 2022. [5](#)
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *NeurIPS*, 2016. [5](#)
- [3] Yun Bian, Zhilin Zheng, Xu Fang, Hui Jiang, Mengmeng Zhu, Jieyu Yu, Haiyan Zhao, Ling Zhang, Jiawen Yao, Le Lu, et al. Artificial intelligence to predict lymph node metastasis at CT in pancreatic ductal adenocarcinoma. *Radiology*, page 220329, 2022. [1](#)
- [4] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. [3](#)
- [5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. [2](#)
- [6] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [3](#)
- [7] 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. [2](#), [6](#)
- [8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. [2](#), [3](#)
- [9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. [2](#), [3](#)
- [10] Linda C Chu, Seyoun Park, Satomi Kawamoto, Yan Wang, Yuyin Zhou, Wei Shen, Zhuotun Zhu, Yingda Xia, Lingxi Xie, Fengze Liu, et al. Application of deep learning to pancreatic cancer detection: lessons learned from our initial experience. *Journal of the American College of Radiology*, 16(9):1338–1342, 2019. [2](#)
- [11] Linda C Chu, Seyoun Park, Sahar Soleimani, Daniel F Fouladi, Shahab Shayesteh, Jin He, Ammar A Javed, Christopher L Wolfgang, Bert Vogelstein, Kenneth W Kinzler, et al. Classification of pancreatic cystic neoplasms using radiomic feature analysis is equivalent to an experienced academic radiologist: a step toward computer-augmented diagnostics for radiologists. *Abdominal Radiology*, pages 1–12, 2022. [5](#)
- [12] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O’Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018. [2](#)
- [13] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. [3](#)
- [14] Hexin Dong, Zifan Chen, Mingze Yuan, Yutong Xie, Jie Zhao, Fei Yu, Bin Dong, and Li Zhang. Region-aware metric learning for open world semantic segmentation via meta-channel aggregation. In *31th International Joint Conference on Artificial Intelligence (IJCAI-22)*, 2022. [3](#), [6](#)
- [15] Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, and HT Kung. Neural mean discrepancy for efficient out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19217–19227, 2022. [3](#)
- [16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. [2](#)
- [17] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2022. [2](#)
- [18] 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. [2](#)
- [19] Mattias P Heinrich, Mark Jenkinson, Michael Brady, and Julia A Schnabel. Mrf-based deformable registration and ventilation estimation of lung ct. *IEEE Transactions on Medical Imaging*, 32(7):1239–1248, 2013. [5](#)
- [20] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *ICML*, 2022. [6](#), [7](#), [8](#)
- [21] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations, ICLR*, 2017. [3](#), [6](#), [7](#), [8](#)
- [22] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *International Conference on Learning Representations, ICLR*, 2019. [3](#)
- [23] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016. [5](#)
- [24] 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 segmen-

- tation. *Nature methods*, 18(2):203–211, 2021. 2, 4, 5, 6, 8
- [25] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018. 2
- [26] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15425–15434, 2021. 3, 6, 7, 8
- [27] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 6
- [28] Philipp Kickingereder, Fabian Isensee, Irada Tursunova, Jens Petersen, Ulf Neuberger, David Bonekamp, Gianluca Brugnara, Marianne Schell, Tobias Kessler, Martha Foltyn, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *The Lancet Oncology*, 20(5):728–740, 2019. 1, 2
- [29] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *International Conference on Learning Representations, ICLR*, 2018. 3
- [30] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018. 3
- [31] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng Ann Heng. H-DenseUNet: Hybrid densely connected UNet for liver and liver tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging*, 2017. 2
- [32] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2152–2161, 2019. 3
- [33] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2019. 5
- [34] Siqi Liu, Daguang Xu, S Kevin Zhou, Olivier Pauly, Sasa Grbic, Thomas Mertelmeier, Julia Wicklein, Anna Jerebko, Weidong Cai, and Dorin Comaniciu. 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 851–858. Springer, 2018. 2
- [35] Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26(6):900–908, 2020. 1
- [36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2
- [37] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. 2
- [38] Hossein Mirzaei, Mohammadreza Salehi, Sajjad Shahabi, Efstratios Gavves, Cees GM Snoek, Mohammad Sabokrou, and Mohammad Hossein Rohban. Fake it till you make it: Near-distribution novelty detection by score-based generative models. *arXiv preprint arXiv:2205.14297*, 2022. 3
- [39] Philipp Oberdiek, Matthias Rottmann, and Gernot A Fink. Detection and retrieval of out-of-distribution objects in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 328–329, 2020. 3
- [40] Walter HL Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475, 2022. 3
- [41] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021. 3
- [42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 6
- [43] Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, et al. Does your dermatology classifier know what it doesn’t know? detecting the long-tail of unseen conditions. *Medical Image Analysis*, 75:102274, 2022. 1, 2, 3
- [44] Simeon Springer, David L Masica, Marco Dal Molin, Christopher Douville, Christopher J Thoburn, Bahman Afshari, Lu Li, Joshua D Cohen, Elizabeth Thompson, Peter J Allen, et al. A multimodality test to guide the management of patients with a pancreatic cyst. *Science Translational Medicine*, 11(501):eaav4772, 2019. 5
- [45] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 2
- [46] Hao Tang, Xuming Chen, Yang Liu, Zhipeng Lu, Junhua You, Mingzhou Yang, Shengyu Yao, Guoqi Zhao, Yi Xu, Tingfeng Chen, et al. Clinically applicable deep learning framework for organs at risk delineation in ct images. *Nature Machine Intelligence*, 1(10):480–491, 2019. 2
- [47] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022. 2, 6
- [48] Yu Tian, Guansong Pang, Fengbei Liu, Yuanhong Chen, Seon Ho Shin, Johan W Verjans, Rajvinder Singh, and Gustavo Carneiro. Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 128–140. Springer, 2021. 1, 3
- [49] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5
- [50] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5463–5474, 2021. 2, 5
- [51] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020. 2, 3
- [52] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *European Conference on Computer Vision*, pages 145–161. Springer, 2020. 3, 6
- [53] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2
- [54] Jiawen Yao, Kai Cao, Yang Hou, Jian Zhou, Yingda Xia, Isabella Nogues, Qike Song, Hui Jiang, Xianghua Ye, Jianping Lu, et al. Deep learning for fully automated prediction of overall survival in patients undergoing resection for pancreatic cancer: A retrospective multicenter study. *Annals of Surgery*, 2022. 1, 2
- [55] Koichiro Yasaka, Hiroyuki Akai, Osamu Abe, and Shigeru Kiryu. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. *Radiology*, 286(3):887–896, 2018. 5
- [56] Lequan Yu, Xin Yang, Hao Chen, Jing Qin, and Pheng-Ann Heng. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3D MR images. In *AAAI*, 2017. 2
- [57] Qihang Yu, Huiyu Wang, Dahun Kim, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2560–2570, 2022. 2, 3
- [58] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *European Conference on Computer Vision*, pages 288–307. Springer, 2022. 2, 3, 5, 6
- [59] Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan L Yuille, and Wei Shen. Glance-and-gaze vision transformer. *Advances in Neural Information Processing Systems*, 34:12992–13003, 2021. 2
- [60] Ling Zhang, Vissagan Gopalakrishnan, Le Lu, Ronald M Summers, Joel Moss, and Jianhua Yao. Self-learning to detect and segment cysts in lung ct images without manual annotation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1100–1103. IEEE, 2018. 5
- [61] Tianyi Zhao, Kai Cao, Jiawen Yao, Isabella Nogues, Le Lu, Lingyun Huang, Jing Xiao, Zhaozheng Yin, and Ling Zhang. 3D graph anatomy geometry-integrated network for pancreatic mass segmentation, diagnosis, and quantitative patient management. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13743–13752, 2021. 1, 2
- [62] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 2
- [63] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021. 1
- [64] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2019. 2, 6
- [65] David Zimmerer, Peter M Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas, et al. Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. *IEEE Transactions on Medical Imaging*, 41(10):2728–2738, 2022. 1, 3



Advanced prompting as a catalyst: Empowering large language models in the management of gastrointestinal cancers

Jiajia Yuan,^{1,8} Peng Bao,^{2,8} Zifan Chen,^{2,8} Mingze Yuan,^{2,8} Jie Zhao,^{3,7} Jiahua Pan,⁷ Yi Xie,¹ Yanshuo Cao,¹ Yakun Wang,¹ Zhenghang Wang,¹ Zhihao Lu,¹ Xiaotian Zhang,¹ Jian Li,¹ Lei Ma,⁶ Yang Chen,^{1,*} Li Zhang,^{2,6,*} Lin Shen,^{1,*} and Bin Dong^{4,5,6,7,*}

*Correspondence: yang_chen@bjcancer.org (Y.C.); zhangli_pku@pku.edu.cn (L.Z.); shenlin@bjmu.edu.cn (L.S.); dongbin@math.pku.edu.cn (B.D.)

Received: July 24, 2023; Accepted: August 8, 2023; Published Online: August 14, 2023; <https://doi.org/10.59717/j.xinn-med.2023.100019>

© 2023 The Author(s). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Prompt engineering affects large language models' performance in GI oncology.
- Prompts with templates and in-context learning enhance large language models' output.
- Multi-round interaction helps large language models to reach the best performance.
- Such performance meets the need of senior GI oncologists for effective AI agents.



Advanced prompting as a catalyst: Empowering large language models in the management of gastrointestinal cancers

Jiajia Yuan,^{1,8} Peng Bao,^{2,8} Zifan Chen,^{2,8} Mingze Yuan,^{2,8} Jie Zhao,^{3,7} Jiahua Pan,⁷ Yi Xie,¹ Yanshuo Cao,¹ Yakun Wang,¹ Zhenghang Wang,¹ Zhihao Lu,¹ Xiaotian Zhang,¹ Jian Li,¹ Lei Ma,⁶ Yang Chen,^{1,*} Li Zhang,^{2,6,*} Lin Shen,^{1,*} and Bin Dong^{4,5,6,7,*}

¹Department of Gastrointestinal Oncology, Key Laboratory of Carcinogenesis and Translational Research, Peking University Cancer Hospital and Institute, Beijing 100142, China

²Center for Data Science, Peking University, Beijing 100871, China

³National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing 100871, China

⁴Beijing International Center for Mathematical Research, Peking University, Beijing 100871, China

⁵Center for Machine Learning Research, Peking University, Beijing 100871, China

⁶National Biomedical Imaging Center, Peking University, Beijing 100871, China

⁷Peking University Changsha Institute for Computing and Digital Economy, Changsha 410205, China

⁸These authors contributed equally

*Correspondence: yang_chen@bjcancer.org (Y.C.); zhangli_pku@pku.edu.cn (L.Z.); shenlin@bjmu.edu.cn (L.S.); dongbin@math.pku.edu.cn (B.D.)

Received: July 24, 2023; Accepted: August 8, 2023; Published Online: August 14, 2023; <https://doi.org/10.59717/j.xinn-med.2023.100019>

© 2023 The Author(s). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Citation: Yuan J, Bao P, Chen Z, et al., (2023). Advanced prompting as a catalyst: Empowering large language models in the management of gastrointestinal cancers. *The Innovation Medicine* 1(2), 100019.

Large Language Models' (LLMs) performance in healthcare can be significantly impacted by prompt engineering. However, the area of study remains relatively uncharted in gastrointestinal oncology until now. Our research delves into this unexplored territory, investigating the efficacy of varied prompting strategies, including simple prompts, templated prompts, in-context learning (ICL), and multi-round iterative questioning, for optimizing the performance of LLMs within a medical setting. We develop a comprehensive evaluation system to assess the performance of LLMs across multiple dimensions. This robust evaluation system ensures a thorough assessment of the LLMs' capabilities in the field of medicine. Our findings suggest a positive relationship between the comprehensiveness of the prompts and the LLMs' performance. Notably, the multi-round strategy, which is characterized by iterative question-and-answer rounds, consistently yields the best results. ICL, a strategy that capitalizes on interrelated contextual learning, also displays significant promise, surpassing the outcomes achieved with simpler prompts. The research underscores the potential of advanced prompt engineering and iterative learning approaches for boosting the applicability of LLMs in healthcare. We recommend that additional research be conducted to refine these strategies and investigate their potential integration, to truly harness the full potential of LLMs in medical applications.

INTRODUCTION

Large Language Models (LLMs), exemplified by cutting-edge architectures like GPT-4,¹ have demonstrated considerable potential in transforming healthcare delivery²⁻⁴ and competency in medical examinations.⁵ This influence is manifested across various healthcare sectors, including online patient interaction,⁶ preventive oncology,⁷⁻⁹ neuropsychiatry,¹⁰ dermatology,¹¹ and aesthetic surgery consultation,^{12,13} underscoring their remarkable versatility. However, the application of LLMs such as GPT-4 in digestive system cancer treatment remains an underexplored area. The complexities inherent to this field, from patient consultation, and diagnosis to treatment planning and follow-up care, pose formidable challenges for LLMs. Additionally, the existing body of research^{2,5,7} primarily evaluates LLMs' responses to common medical inquiries via rudimentary prompting, which may not fully leverage their potential in medical settings. This highlights the need for a more comprehensive assessment of GPT-4's capability to provide personalized cancer treatment recommendations via sophisticated prompts.

To harness the full potential of LLMs, it is crucial to employ effective prompt engineering.¹⁴⁻¹⁹ Prompt engineering, a process of creating, testing, and optimizing input prompts, serves as a crucial tool in controlling and enhancing interactions with LLMs. Various techniques such as in-context learning,¹⁵ retrieval-augmented generation,¹⁶ chain-of-thought,¹⁷ and least-to-most prompting¹⁸ have been shown to significantly improve the performance of LLMs in tasks demanding logical thinking and reasoning. In-context learn-

ing offers models a few demonstrations before attempting a task, while retrieval-augmented generation enhances this process by retrieving relevant examples from a given database. Chain-of-thought prompting improves LLMs' reasoning ability by directing them to generate a series of intermediate steps toward a solution, and least-to-most prompting dissects complex problems into simpler sub-problems to be solved sequentially. Intuitively, these techniques could effectively boost LLMs' performance in complex medical tasks, including cancer treatment recommendations.

In this study, we aimed to unleash GPT-4's potential to provide personalized digestive system cancer treatment plans through prompt engineering. Inspired by the thinking, reasoning, and action processes of digestive oncologists, we initially conceived the iterative procedure of prompt engineering as a method of amassing information regarding gastrointestinal tumors within a distinct storage of knowledge and in turn, educating the LLM. However, these knowledge repositories, when embedded in rudimentary prompts, are often devoid of substantial content, thus limiting their potential to effectively guide LLMs. Consequently, we established an empirically effective multi-step prompt template consisting of: (i) declaring the role, this process involves assigning a particular role to GPT-4 that emulates a real-world professional or function; (ii) stating the main task, this step essentially provides GPT-4 with a clear directive of what it is required to accomplish; (iii) declaring the workflow, which we view as a generalized chain of thought that allows GPT-4 to approach problem-solving or deliver answers in an organized, step-by-step manner; (iv) specifying constraints, it involves defining the boundaries within which GPT-4 should operate. Then, we iteratively refined this template to align GPT-4's responses with physicians' requirements and added elements to generate comforting responses for patients. An experienced oncology specialist subsequently interacted with GPT-4 over multiple rounds to further guide and optimize the recommended treatment plans. Furthermore, motivated by the exemplar-based teaching approach in medicine, we also assessed the impact of in-context learning by providing GPT-4 with examples of ideal treatment suggestions through document retrieval. We evaluated the performance of diverse prompt engineering strategies on 43 case reports, encompassing a wide range of digestive system cancer types, utilizing a clinically standardized evaluation metric.

In summary, we are the first to conduct a comprehensive assessment of prompt engineering on GPT-4's ability to provide personalized digestive system cancer treatment recommendations, as per our comprehensive search in the existing literature. We developed a sophisticated prompt template to generate personalized cancer treatment plans that emphasize patient comfort, which significantly outperforms rudimentary prompts and offers valuable insights for prompt design in the medical domain. We evaluated various prompt engineering strategies, including rudimentary prompts, templated prompts, in-context learning, and multi-round interaction, using a clinically standardized metric. Our results highlight the promise of prompt engineering for medical applications of LLMs.

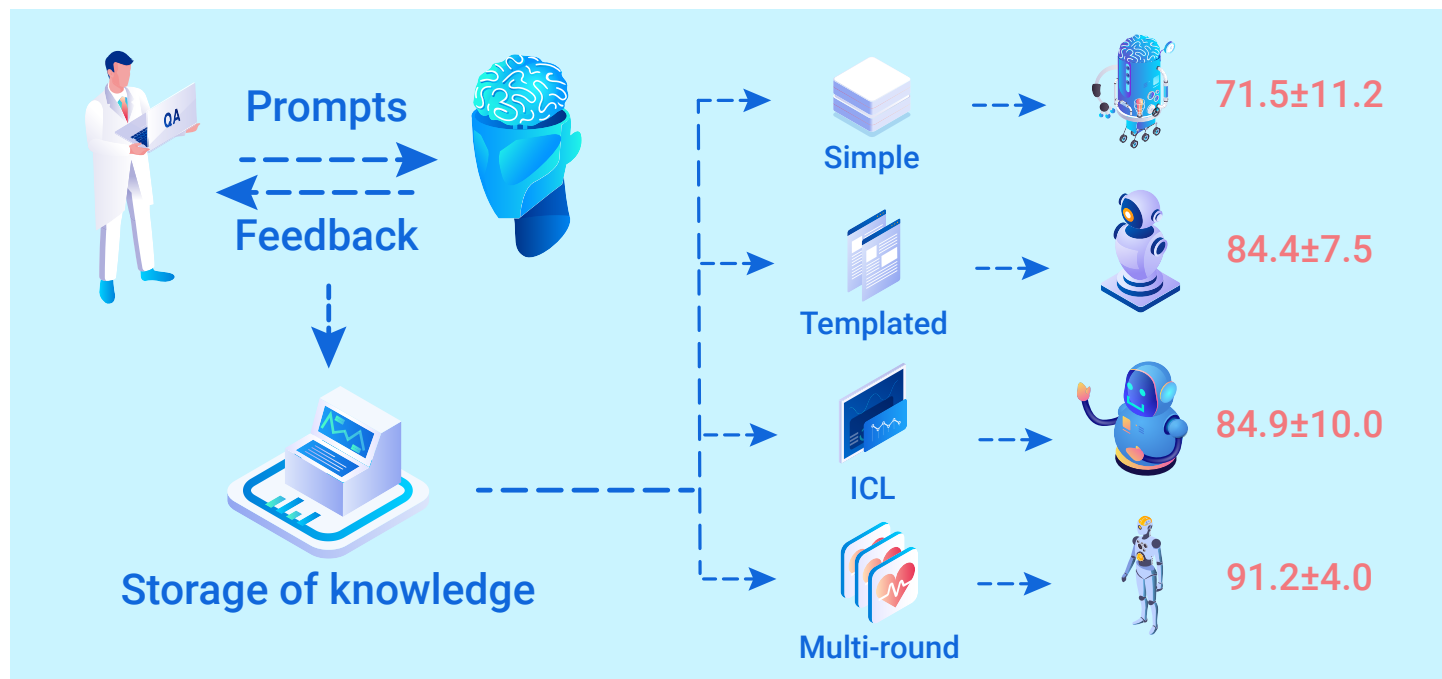


Figure 1. An illustration showcasing the effects of various prompting strategies on Language Learning Models' (LLMs) performance, mediated by a 'storage of knowledge'. Simple prompts leave this storage empty, offering no enhancement for GI tumor decision-making. Conversely, templated prompts and ICL populate the storage with role assumptions and case examples, respectively, helping to standardize LLMs' output, thus improving performance. The multi-round interaction strategy fills the storage with the complete physician-LLM dialogue, potentially allowing more accurate comprehension and utilization of decision-assisting information.

MATERIALS AND METHODS

Materials

In this study, we propose an innovative methodology to augment the learning capability of LLMs by incorporating multifaceted prompt design and dynamic training approaches. As shown in Figure 1, diverse prompt designs can be perceived as varying modifications to the storage of knowledge, encompassing manual alterations meticulously orchestrated based on GI tumor expertise, automatic modifications that explore the hospital's pre-existing data for analogous cases as pedagogical instances for the LLM, and dynamic modifications consistently interrogated and addressed during the deployment of the consultation process. Consequently, the design of the prompts was executed as follows: Initially, the models are subjected to a more sophisticated introduction prompt, intricately crafted with complex semantic and structural nuances, thereby priming the LLMs to comprehend and respond to intricate queries. Furthermore, an advanced method of in-context learning is introduced, encouraging the models to extract knowledge and patterns from various contexts rather than individual sentences, fostering a more comprehensive understanding of the text. To accommodate evolving data patterns, we also incorporate online learning techniques, enabling the LLMs to continually learn and adapt from real-time, dynamic data. Lastly, we implement an iterative feedback loop through multi-round question-and-answer sessions, reinforcing the model's ability to comprehend, retain, and apply information over successive interactions. This combination of sophisticated prompt architecture, in-context learning, online learning, and iterative interactions aims to substantially enhance the LLM's predictive and interpretative capabilities, pushing the frontiers of AI language understanding. We used publicly available medical licensing examination cases, oncology residency and attending physician exam cases as text source.

Templated prompts

Past studies have shown that a good use of different prompt engineering,^{17,20,21} as well as properly designed prompt templates²² can significantly improve the problem-solving ability of large language models, and this phenomenon was similarly observed in our study. As shown in Figure S1, we developed our prompt template by adopting a four-pronged approach as follows:

Declaring the role. Assigning a 'role' or 'identity' to large language models is one of the commonly used techniques for interacting with these models. Previous research²² supports that this method can effectively guide what type of output the models generate and what details they prioritize. In our study, we assigned the role of a digestive oncology specialist to GPT-4, emphasizing its range of skills that included clinical diagnosis, treatment, and communication techniques. We found this strategy successfully influenced GPT-4's behaviors, responses, and interaction styles to align with the expectations of the role.

Stating the main task. This approach essentially provides GPT-4 with a clear directive on what it is expected to accomplish. In our study, the primary task of our model is to deliver detailed and accurate advice to patients with digestive system cancers. This involves defining the central task that GPT-4 needs to perform. Given the context of our research, our model, acting as a digestive oncology specialist, is tasked with generating personalized treatment plans for digestive system cancer patients. By articulating the main task, we direct GPT-4's focus, streamline its reasoning process, and enhance its ability to produce task-specific, relevant, and actionable outputs. In addition, to enable GPT-4 to produce complex and contextually accurate responses, we've included a wide range of scenarios and contexts, from simple situations to the complexity of academic discourse in hospitals. We also encourage GPT-4 to link different pieces of information together. This approach aids GPT-4 in moving beyond simple pattern recognition, facilitating a deeper understanding when executing tasks.

Declaring the workflow. We have defined a comprehensive workflow in the prompt templates, which includes case analysis, clinical examination, scheduling examination, diagnosis and treatment, execution and adjustment of treatment, and follow-ups. This is also the general workflow of a professional digestive oncology specialist. We believe this represents a generalized chain of thought and many studies^{17,20,21,23} have already demonstrated that this approach can stimulate LLM's reasoning ability. We find this strategy ensures that GPT-4's output is more consistent and logical, using a planned, step-by-step approach to accomplish tasks, which is very similar to the process a human expert uses to solve problems. By structuring GPT-4's thinking in this way, we can effectively manage its output, improve overall consistency, and reduce the likelihood of generating irrelevant or erroneous information.

Specifying constraints. In this process, we've incorporated certain constraints into the prompt templates. We require GPT-4 not to make responses when uncertain or additional information is needed, but rather, it must first gather sufficient information. In addition, we require GPT-4 to provide detailed and correct guidance for a specific case, as GPT-4 tends to give general and non-specific answers that may not be wrong but lack specificity. This approach ensures that GPT-4 avoids generating responses that are undesirable or beyond its scope, thereby enhancing its effectiveness and minimizing potential deviations. We also advised GPT-4 to build a trusting doctor-patient relationship in a warm, humorous manner rather than in a cold and impersonal way when answering.

In-context learning

In this study, we introduce an automated in-context learning (ICL) approach to refine GPT-4's capabilities, focusing on the integration of doctors' habits and cognition. This method assimilates insights drawn from analogous past cases and is comprised of three main components: firstly, transposing past patient conditions into a designated embedding space; secondly, gauging the similarity between the current condition and these archived cases to identify its k-nearest counterparts; and finally, building in-context learning prompts based on these identified cases. We provide a detailed exposition of these three components in the following:

Encoding patient conditions using pre-trained chinese BERT model. A pre-trained Chinese BERT model in Hugging Face (<https://huggingface.co/hfl/chinese-bert-wwm-ext>), specifically the "hfl/chinese-bert-wwm-ext", is utilized to translate patient conditions into a high-dimensional embedding space (768 dimensions in this study), capturing the context of the condition effectively. The BERT tokenizer is used to convert condition text into input vectors, which are then fed into the BERT model. Operating in a no-gradient update setting, the "pooler_output" from the model serves as the sentence embedding for each patient condition.

Calculation of cosine similarity and identification of k-nearest neighbors. Once the embeddings for all patient conditions have been computed, we calculate the cosine similarity between them to derive a similarity score. This metric provides a measure of the contextual similarity between different patient conditions. Based on these similarity scores, we identify up to k-nearest neighbors for each patient condition (with k being up to four depending on the token limitation of GPT-4).

Generation of in-context learning prompts. For each patient condition, we generate an enriched prompt that includes the top-k similar past cases and the corresponding doctor's suggestions. To ensure consistency and readability in these prompts, a pre-defined template is used: "As an experienced clinician, your responsibilities include understanding and analyzing patient information and chief complaints, [...]. Now, let's look at these examples: [...]. After analyzing these examples, here is a new patient: [...]. Please give specific treatment plan suggestions based on the above examples and relevant literature. (see [Figures 4 & S6](#) for details).

Metrics

We have developed a unique set of metrics, drawing from those typically used for evaluating clinicians' examinations, to quantitatively assess the results generated by various methods. These metrics encompass six key aspects:

1. Understanding medical history (0-20): This metric assesses how accurately and comprehensively an LLM captures and interprets a patient's medical history. This includes consideration of the patient's previous diagnoses, surgeries, hospitalizations, allergies, family history, lifestyle, and other relevant information.

2. Diagnosis and differential diagnosis (0-20): This metric assesses the ability of the LLM to accurately diagnose the patient's condition based on the medical history. It includes both the primary diagnosis and any differential diagnoses.

3. Further examination and reason (0-10): This metric evaluates the appropriateness of any additional examinations suggested by the LLM. It measures not only whether the recommended examinations are suitable, but also if they are justified based on the patient's condition and symptoms. The LLM

should also provide a clear rationale for why these additional examinations are needed.

4. Principles and plans of treatment (0-20): This metric evaluates the LLM's ability to propose a suitable treatment plan. The plan should be personalized for the patient, taking into account factors like age, overall health, potential side effects, and patient preferences.

5. Breadth and depth of results (0-20): This metric measures how comprehensively the LLM covers the scope of medical knowledge in its results (breadth), as well as how much detail it provides (depth). Breadth refers to the range of different topics or areas covered in the results, while depth refers to the level of detail or complexity within those topics.

6. Thinking and expressing ability (0-10): This is a measure of how effectively the LLM reasons and communicates its findings. Thinking refers to the LLM's ability to logically process and interpret data, make connections, draw conclusions, and anticipate potential outcomes. The expressing ability should not only be clear and accurate but also demonstrate empathy in line with a real clinician's interaction. This includes sensitivity to the patient's emotional state, using comforting and supportive language, and showing understanding and respect for the patient's experiences and concerns. By effectively incorporating empathy, the LLM can build trust, encourage open communication, and provide emotional support in addition to addressing physical health issues.

To gain a clearer understanding of performance based on the total scores, we have defined the following expertise levels:

1. Level A (90-100 points): Top-level expertise, capable of independently managing complex and rare cases, demonstrating exceptional skills and professional knowledge.

2. Level B (80-89 points): Experienced level, capable of handling most cases, but requires guidance for complex or rare cases.

3. Level C (70-79 points): Mid-level competence, capable of independently addressing common cases, requires guidance for complex ones.

4. Level D (60-69 points): Junior level, capable of handling some common cases, but requires close guidance for complex cases.

5. Level E (below 60 points): Initial training level, needs guidance from experienced clinicians in all aspects.

RESULTS

Templated evaluation

[Figures 2 & S2](#) provide a comparison between our designed prompting template ([Figure 2B](#)) and the standard, direct prompting ([Figure 2A](#)) utilized by GPT-4. The findings underscore that the designed template for role assumption ([Figure S1](#)) can improve GPT-4 to make more complex decisions based on the patient's individual circumstances. In the provided example, our designed prompting can prioritize the control of disease progression, symptom relief, enhancement of life quality, and survival extension, instead of merely pursuing a cure unconditionally. Moreover, the template manifests an exceptional ability to interweave quality-of-life considerations within the treatment strategies and provides comprehensive guidance ([Figure S3](#)). It also underscores the significance of continuous patient assessment and the pursuit of innovative, custom treatments ([Figure S4](#)). As opposed to direct prompting, our designed prompting template possesses the ability to mimic the intricate treatment ideation process, enhancing GPT-4's efficacy as a therapeutic advisory tool when acting as a senior oncologist.

Multi-round evaluation

[Figures 3 & S5](#) illustrate an interaction with GPT-4 for cancer treatment advice. Initially, GPT-4 prematurely diagnosed the patient with late-stage cancer and proposed a treatment plan. However, this was inappropriate, given the necessity for a more accurate staging diagnosis for this patient. As highlighted in [Figure 3](#), the clinician directed GPT-4 to offer a detailed staging diagnosis, subsequently pointing out its error. Following multiple question-and-answer interactions with the clinician, GPT-4 acknowledged its mistake and adjusted its response. It began by determining the cancer's stage, before suggesting a specific treatment plan. This revised response is not only more suitable for the patient but also provides her with hope. This multi-round interaction demonstrates the learning capability of large language models like GPT-4, highlighting their ability to quickly integrate human logical reasoning within the context of intricate medical scenarios.

A Direct prompting



Doctor

Male, 68 years old. The patient was found to have a gastric mass in an external hospital. Gastroscopy showed that at 41-47cm from the incisors, there is an ulcerative tumor at the esophagogastric junction. Pathology: Esophagogastric junction. Biopsy: moderately differentiated adenocarcinoma. Lauren classification: intestinal type. Immunohistochemical results: HER2(0) [HER2 positive control (3+), HER2 negative control (0)]. CT: Thickening of the gastric wall, considered as Borrmann type 3 gastric cancer(ct4aN3a) ; Multiple filling defects in the splenic vein and portal vein, suggesting tumor thrombus. The tumor markers are normal. Considering the patient's multiple portal vein tumor thrombi, it is not suitable for clinical trial enrollment.

This is the information of a patient. Please search for relevant literature and provide specific treatment plan recommendations, and explain the reasons.

- This prompt is summarized into "{Sex}, {Age}, {Disease description}, {Question}." and will be referred to in other content.

Based on the information provided, immunohistochemistry shows HER2 positive. Due to the presence of multiple tumor thrombi in the portal vein, the patient is not suitable for clinical trial enrollment.

The treatment plan is as follows:

1. Chemotherapy: Considering the patient's condition, neoadjuvant chemotherapy can be administered to reduce tumor size, alleviate symptoms, and increase the possibility of surgery. Common chemotherapy regimens include the FLOT regimen and the EOX regimen.
2. Surgery: If the tumor shrinks and the condition stabilizes after chemotherapy, the possibility of surgical removal can be evaluated. Depending on the patient's specific situation, a gastric cancer radical operation (D2 or D1+ lymph node dissection) or palliative surgery can be chosen.
3. Targeted therapy: For HER2-positive patients, HER2-targeted therapy drugs such as trastuzumab can be added during chemotherapy.
4. Supportive therapy: Nutritional support and symptomatic treatment.



GPT-4

B Make role assumptions before prompting



Doctor

{Role declaration}
{Definition of a Gastrointestinal Oncologist}
{Skills of a Gastrointestinal Oncologist}
{Work Process of a Gastrointestinal Oncologist}
{Specific Requirements and Rules for the Response}

- {Sex}, {Age}, {Disease description}.

Based on the information you provided, the patient has been diagnosed with moderately differentiated adenocarcinoma of the esophagogastric junction, with a pathological type of intestinal, HER2-negative. CT scans show the existence of tumor thrombus, which may suggest the disease is in the late stage. However, please note that a more detailed clinical data and expert evaluation are needed for a definite diagnosis.

For patients with advanced gastric cancer of this type, systemic chemotherapy is often the first choice, such as the FOLFOX regimen (fluorouracil + cisplatin + oxaliplatin) or the SOX regimen (S-1 + oxaliplatin). Moreover, considering the HER2-negative pathological result, targeted drugs like trastuzumab may not be applicable here. Additionally, if the disease is well-controlled, palliative surgery can be considered to alleviate symptoms.

Regarding the tumor thrombus, we need to evaluate in detail whether there is a possibility for surgery or interventional treatment, or if anticoagulant therapy is needed. If the tumor thrombus causes severe clinical symptoms, surgery or interventional treatment may be considered, but this requires a comprehensive assessment of the patient's overall condition and other complications.

At the same time, it is also recommended to conduct relevant genetic testing, such as MLH1, MSH2, MSH6, PMS2, etc., because a family history of cancer may be related to hereditary cancer syndromes.

This is a possible diagnosis and treatment plan based on the information I currently have. The actual treatment plan needs to be made based on the patient's specific situation and a comprehensive assessment by the medical team. Please discuss these suggestions with your medical team to develop the most suitable treatment plan for the patient.



GPT-4

Figure 2. Evaluation of prompting template design (A) Direct prompting. GPT-4 is directly provided with a patient's condition and asked to suggest a treatment plan. (B) The proposed templated prompting (make role assumptions before prompting). GPT-4 is initially primed with a specifically designed role-embedded prompt, ensuring the AI understands the task requirements and responsibilities. Subsequently, the patient's condition is provided, and GPT-4 is asked to suggest a treatment plan. Note that this figure is a reduced version of [Figure S2](#).

ICL evaluation

As demonstrated in [Tables 1-4](#), the performance of In-Context Learning (ICL) exceeded that of rudimentary prompting by a substantial margin across

various types of digestive system cancer treatments, with a notable difference of 13.4 points in overall performance. [Figures 4 & S6](#) provide an illustrative comparison between in-context learning and rudimentary prompting.



Doctor

{Role assumption first}

Female, 76 years old. Poor appetite. Gastroscopy revealed a raised lesion with a concave surface at the cardia, extending to the gastric fundus (Siewert II type, adenocarcinoma), and coarse gastric mucosa consistent with Borrmann IV type gastric cancer presentation (adenocarcinoma). Pathology report: a small amount of poorly differentiated cancer (gastric body), poorly differentiated adenocarcinoma; Cmet(1+), EGFR(0), HER2(0), pMMR, pan-TRK(-), PD-L1 (CPS<1), EBER(-); (Cardia) moderately differentiated adenocarcinoma, intestinal type; Cmet(2+), EGFR(1+), HER2(0), pMMR, pan-TRK(-), PD-L1 (CPS=10), EBER(-). Normal lab results, chest-abdominal-pelvic enhanced CT showed diffuse thickening of the entire stomach, lymph node metastasis between the liver and stomach. Currently, the patient's food intake has decreased, bowel movements are not abnormal, and she has lost about 5 kg in the past 2 months.

Based on the above information, the treatment plan should focus on alleviating symptoms and improving quality of life. Specific suggestions are as follows:

1. Chemotherapy: Common chemotherapy regimens include FOLFOX, XELOX, SOX, etc.
2. Targeted Drugs: C-MET and EGFR are slightly positive, so targeted drugs against these two markers, such as crizotinib (a C-MET inhibitor) and erlotinib (an EGFR inhibitor) can be considered.
3. Immunotherapy: Although the expression level of PD-L1 is low, immunotherapy checkpoint inhibitors like Keytruda (pembrolizumab) can still be considered.
4. Palliative Care: This includes pain management, nutritional support, and psychosocial support.
5. Regular Follow-up: Regular gastroscopy review and chest, abdominal and pelvic CT scans are necessary.



GPT-4



Doctor

How do you analyze and determine the staging diagnosis and future prognosis for this patient?

The staging of gastric cancer primarily refers to the American Joint Committee on Cancer (AJCC) TNM staging system, which assesses the depth of tumor invasion (T), the status of lymph node metastasis (N), and the condition of distant metastasis (M). According to the information you provided, the patient's gastric cancer has developed lymph node metastasis (LNM), indicating that the N part of the TNM stage is at least N1 (1-2 regional lymph nodes metastasis) or higher. Simultaneously, the gastric cancer is Borrmann IV type, suggesting that the tumor has extensively invaded the stomach wall, and in the TNM stage, the grading of the T part is usually T3 (tumor invades the serosa) or T4 (tumor penetrates the serosa or directly invades adjacent structures). Based on this information, it can be preliminarily judged that the patient is in at least stage III of gastric cancer (including T3N1M0, T4aN1M0, T4bN1M0, etc.) or a higher stage.



GPT-4



Doctor

The patient's lymph node metastasis (LNM) between the liver and stomach could be classified as either N1 or N3, and a detailed interpretation of the lymph node metastasis is needed. Will there be a chance for surgery in the future? This will have a significant impact on future treatment outcomes.

In terms of the resectability of the disease, the patient's lymph node metastasis between the liver and stomach (LNM) may affect surgical decisions. The N stage in the TNM staging is at least N1 (1-2 regional lymph node metastases). In some cases, for locally advanced gastric cancer, even with lymph node metastasis, radical surgery (such as D2 gastrectomy) may be performed. This surgery can remove as many lymph nodes as possible, but the risk of surgery is high, and the rate of postoperative complications and mortality is also relatively high, especially in the elderly population. However, for advanced gastric cancer (such as the patient's Borrmann IV type gastric cancer), especially in cases accompanied by extensive lymph node metastasis or distant metastasis, surgery is usually not the first choice of treatment, because surgery at this time cannot significantly improve prognosis and may instead bring greater risks and complications.



GPT-4



Doctor

Based on the above discussion, please provide a new recommendation for this patient.

I suggest the following:

1. Detailed Staging: First, I suggest more detailed tumor staging. This may include PET-CT scanning to find any other unknown metastatic lesions. Additionally, for lymph node metastasis, we need to further clarify its extent in order to stage more accurately.
2. Peritoneal Metastasis Examination: Considering the possibility of common peritoneal metastasis in gastric cancer, I recommend laparoscopic examination for direct observation of the abdominal cavity and to obtain living tissues for pathological examination, to find any potential hidden peritoneal metastasis.
3. Individualized Treatment Plan: Based on the detailed staging and peritoneal metastasis examination results, we can further discuss the treatment plan. This may include chemotherapy, targeted therapy, and immunotherapy. Especially, immunotherapy might be tried considering her PD-L1(22C3)(CPS=10).
4. Nutrition and Symptom Management: Given her age and current condition (weight loss of about 5kg in the past 2 months, decreased food intake), I recommend asking a dietitian for evaluation and management, and there may also need to be appropriate palliative treatment for symptoms such as decreased appetite.



GPT-4

Figure 3. A Multi-round Interaction with GPT-4 for Gastric Cancer Treatment Advice Note that this figure is a reduced version of [Figure S5](#).

A Direct prompting



Doctor

Male, 59 years old. Visited the clinic due to abdominal pain after eating. Colonoscopy indicated a tumor 30cm from the hepatic flexure. Pathology: Sigmoid colon: Moderately differentiated adenocarcinoma with mucinous adenocarcinoma. PET-CT: Considering colon cancer, multiple metastatic lymph nodes in the abdominal pelvic cavity, peritoneum, and retroperitoneum, multiple liver metastases. Tumor markers: CEA 37.93, CA199 368.3, CA724 103.4, CA242 463.6. Recent weight loss of 8KG.
This is the information of a patient. Please search for relevant literature and provide specific treatment plan recommendations, and explain the reasons.

● This prompt is summarized into "{Target case}, {Question}." and will be referred to in other content.

- 1. Chemotherapy:** Common chemotherapy drugs include fluorouracil (5-FU), oxaliplatin, and irinotecan. Chemotherapy regimens can use FOLFOX or FOLFIRI.
- 2. Targeted therapy:** Based on the genetic testing results of the patient's tumor, you can consider using drugs that target specific targets for treatment, such as EGFR inhibitors (such as Herceptin, cetuximab or panitumumab), VEGF inhibitors (such as bevacizumab) or other applicable targeted therapy drugs.
- 3. Immunotherapy:** PD-1/PD-L1 inhibitors (like pembrolizumab or nivolumab) may be effective for some patients. The applicability of immunotherapy needs to be determined based on the patient's tumor microsatellite instability (MSI) or PD-L1 expression.
- 4. Supportive care:** During treatment, the patient may need hydration, nutritional support, pain relief, and antibiotic treatment, among other supportive care.
- 5. Follow-up and monitoring:** During treatment, tumor markers should be monitored regularly to assess the effectiveness of the treatment.



GPT-4

B In-context learning prompting



Doctor

As an experienced clinical physician, your responsibilities include understanding and analyzing patient information and chief complaints, then based on previous cases and the treatment inspection suggestions of real doctors in those cases, give professional treatment inspection suggestions in a similar style. You need to be professional, careful, and thoughtful, and your answer must be based on existing medical knowledge and consider the patient's personal situation and needs. Now, let's look at these examples:

Case 1: {Case description 1}

Doctor's suggestion: {Doctor's suggestion 1}

Case 2: {Case description 2}

Doctor's suggestion: {Doctor's suggestion 2}

After analyzing these examples, here is a new patient: [Target Case in Panel A]. Please give specific treatment plan suggestions based on the above examples and relevant literature.

Based on the information you provided, I suggest the following treatment plan:

1. Complete gene testing, including KRAS/NRAS/BRAF gene testing, as well as immunohistochemistry of MLH1/PMS2/MSH2/MSH6 to clarify the molecular characteristics of the tumor. If it is dMMR/MSI, it is recommended to treat with PD1 monotherapy ± CTLA4 monotherapy.
2. Choose the appropriate chemotherapy plan based on the results of gene testing. If it is a KRAS/NRAS/BRAF wild type, consider the FOLFIRI scheme combined with targeted therapy (such as cetuximab or bevacizumab). If the gene testing results show mutations, select the FOLFIRI scheme or other suitable chemotherapy plans.
3. Strengthen nutritional support in conjunction with the patient's recent weight loss, to improve the patient's nutritional status.
4. Regularly review tumor markers and imaging examinations (such as CT or MRI of the peritoneum, retroperitoneal lymph nodes, and liver) to assess the treatment effect.
5. Based on the treatment effect and the patient's physical condition, assess the possibility of local treatment, such as surgery, ablation, or radiation therapy.
6. If conditions permit, consider participating in clinical research to obtain more treatment opportunities.
7. Pay attention to observe the patient's symptoms and side effects, and conduct symptomatic treatment accordingly.



GPT-4

Figure 4. Utilizing GPT-4 to Suggest Oncological Treatment Regimens with Distinct Prompting Techniques (A) Direct prompting and (B) In-Context Learning (ICL) prompting. Note that GPT-4 was not provided with the designed prompt, as our objective was to exclusively evaluate the performance of ICL. Note that this figure is a reduced version of Figure S6.

Panel A portrays a clinical scenario in which a patient is diagnosed with rectal cancer and is accompanied by a naive prompt for generating treatment strategy recommendations. In contrast, Panel B highlights the implementation of in-context learning, where GPT-4 is provided with two comparable patient cases, each accompanied by treatment suggestions endorsed by an experi-

enced oncologist. Through in-context learning, GPT-4 successfully absorbed the treatment approach and reasoning process utilized by physicians, leading to the generation of personalized and targeted treatment plans, as opposed to general guideline recommendations offered through rudimentary prompts. Furthermore, ICL emphasized clinically critical aspects such as clin-

ical research and potential side effects, thus enhancing the overall quality and relevance of the generated treatment recommendations.

Overall evaluation

Tables 1-4 provided the performance evaluation of different prompt engineering strategies-Simple, Templated, ICL, and Multi (Multiple Rounds)-on several aspects of understanding and knowledge organization across different disease conditions (Overall, Gastric Cancer, Colorectal Cancer, Other GI

cancers). Broadly, the tables show a general trend of increased performance as we move from the Simple strategy to the Multi-strategy, but templated prompts and ICL show similar performance. The mean scores for all aspects improve noticeably as the complexity of the prompts increases as shown in Table S1.

A few specific observations can be highlighted. First, 'Understanding Medical History' consistently receives full marks in the Multi-strategy, underscoring the effectiveness of iterative questioning in gathering comprehensive

Table 1. Overall performance in all GI cancers

Metrics	Simple	Templated	ICL	Multi-round
Understanding Medical History (0-20)	19.1±2.5	19.9±0.8	20.0±0.0	20.0±0.0
Diagnosis and Differential Diagnosis (0-20)	18.3±3.4	19.4±2.2	19.4±1.9	19.9±0.8
Further Examination and Reason (0-10)	6.6±3.8	8.8±2.6	9.0±2.5	9.8±1.1
Principles and Plans of Treatment (0-20)	10.8±4.0	14.0±3.3	14.2±4.7	16.5±3.2
Breadth and Depth of Results (0-20)	11.4±2.2	14.2±2.1	14.3±1.7	15.2±1.5
Thinking and Expressing Ability (0-10)	5.3±2.0	8.1±2.4	8.0±2.7	9.8±1.1
Total Score (0-100)	71.5±11.2	84.4±7.5	84.9±10.0	91.2±4.0

Table 2. Overall performance in gastric cancer

Metrics	Simple	Templated	ICL	Multi-round
Understanding Medical History (0-20)	18.1±3.3	19.8±1.1	20.0±0.0	20.0±0.0
Diagnosis and Differential Diagnosis (0-20)	16.4±4.1	18.8±3.0	18.8±2.6	19.8±1.1
Further Examination and Reason (0-10)	5.5±3.7	8.8±2.6	8.8±2.6	9.5±1.5
Principles and Plans of Treatment (0-20)	9.8±4.2	13.6±3.8	14.3±4.7	16.4±3.8
Breadth and Depth of Results (0-20)	11.0±2.0	15.0±1.5	14.5±1.5	15.7±1.7
Thinking and Expressing Ability (0-10)	5.0±1.5	9.0±2.0	8.6±2.3	9.5±1.5
Total Score (0-100)	65.7±10.6	85.0±7.7	85.0±10.9	91.0±5.0

Table 3. Overall performance in colorectal cancer

Metrics	Simple	Templated	ICL	Multi-round
Understanding Medical History (0-20)	20.0±0.0	20.0±0.0	20.0±0.0	20.0±0.0
Diagnosis and Differential Diagnosis (0-20)	20.0±0.0	20.0±0.0	20.0±0.0	20.0±0.0
Further Examination and Reason (0-10)	7.3±3.3	9.1±1.9	9.1±1.9	10.0±0.0
Principles and Plans of Treatment (0-20)	10.5±4.0	13.6±3.1	12.3±5.8	17.3±2.5
Breadth and Depth of Results (0-20)	11.4±2.2	14.1±1.9	14.1±1.9	15.0±0.0
Thinking and Expressing Ability (0-10)	4.5±1.4	7.3±2.5	6.8±3.2	10.0±0.0
Total Score (0-100)	73.6±6.4	84.1±6.3	82.3±10.7	92.3±2.5

Table 4. Overall performance in other GI cancer

Metrics	Simple	Templated	ICL	Multi-round
Understanding Medical History (0-20)	19.4±1.6	20.0±0.0	20.0±0.0	20.0±0.0
Diagnosis and Differential Diagnosis (0-20)	19.4±1.6	20.0±0.0	20.0±0.0	20.0±0.0
Further Examination and Reason (0-10)	7.6±3.9	9.1±2.6	9.4±2.4	10.0±0.0
Principles and Plans of Treatment (0-20)	11.2±4.0	14.7±3.2	15.6±3.4	15.6±3.4
Breadth and Depth of Results (0-20)	11.5±2.3	13.5±2.3	14.1±1.9	14.7±1.2
Thinking and Expressing Ability (0-10)	5.9±2.6	7.9±2.5	8.5±2.3	10.0±0.0
Total Score (0-100)	75.0±11.6	85.3±7.8	87.6±7.1	90.3±3.6

information. Second, in the context of 'Principles and Plans of Treatment', the marked improvement along the four types of prompts indicates the importance of diverse and complex prompts in formulating an effective treatment plan. The total scores also follow the same trend, with the Multi-strategy achieving the highest scores across all disease conditions consistently. These results provide strong evidence supporting the effectiveness of employing various prompt engineering strategies, as well as iterative questioning, in enhancing the performance of GPT-4 in medical contexts. However, the exact impact and effectiveness may vary depending on the specific disease condition, necessitating further nuanced analysis.

DISCUSSION

To the best of our knowledge, it is evident that our research is pioneering in exploring the techniques to optimize LLMs specifically for the recommendation of treatments for gastrointestinal cancers. In contrast to studies that solely use simple prompts,^{2,5,6} our research assessed a series of prompt engineering strategies including simple prompts, templated prompts, ICL, and multi-round interaction. Our results demonstrate that complex prompting approaches, especially multi-round interaction, are capable of accruing sufficient diagnostic and therapeutic information pertinent to a specific case. This approach facilitates the rational and efficient expansion of the storage of knowledge, thereby substantially enhancing model performance in collecting medical histories, forming accurate diagnoses, and recommending effective treatments for digestive cancers. The iterative nature of multi-round interaction consistently yielded the highest scores across evaluation metrics, highlighting its reliability and broad applicability. Our study also necessitates further exploration. Firstly, our adopted metric, based on clinicians' examinations, retains some degree of subjectivity. This accentuates the necessity for a more objective clinical evaluation method. We are currently collaborating with statisticians to devise novel evaluation tools to measure model performance more accurately and objectively. Moreover, our investigation was solely focused on tumors in the digestive system, indicating that future research could extend the application of LLMs to other types of cancers. Additionally, our primary clinical scenario was set in China, with the study conducted in Chinese before being translated into English using GPT-4. Although GPT-4 exhibits robust cross-language performance¹, the influence of the selected language on performance deserves further study. Moreover, we have conducted preliminary assessments of various LLMs including GPT-4, Claude, ChatGLM, Wenxin Yiyao, and PaLM. Our findings indicate that GPT-4 exhibits superior underlying capabilities compared to its counterparts, leading us to select it as a representative of LLMs. Nevertheless, a comprehensive examination of the diverse and continually evolving LLMs is still an imperative area of future research. Last but not least, it is crucial to note that the data used to train GPT-4 predominantly originates from sources outside of China. However, due to variations in clinical guidelines, available medical technologies, perceptions of risk and benefit by patients and physicians, as well as disease prevalence trends in different regions, treatment approaches for gastrointestinal cancers can significantly differ across various regions. Subsequently, the outputs generated by GPT-4 may not entirely apply to Chinese patients. This particular aspect could potentially impact the evaluation scores during our comparative experiments of different prompt engineering strategies. To mitigate this potential issue, fine-tuning the LLMs with data specifically sourced from China could provide a more appropriate approach.

Recognizing the constraints and potential biases of LLMs is essential for their responsible and ethical application. One major concern is that LLMs gather knowledge from vast amounts of internet data that may contain inherent biases or inaccuracies. To mitigate potential bias and increase the reliability of our results, we employed a method of inter-rater reliability where each output from the model was independently evaluated by two separate individuals. Their evaluations were then compared and reconciled, ensuring a more objective and balanced assessment of the model's performance. Data privacy and security must be underscored when providing medical records to online LLMs. Thus, we have implemented stringent data protection measures, ensuring all patient data is anonymized and encrypted to protect privacy. Furthermore, the inadequate judgment and critical thinking skills of LLMs when interpreting medical records limit their performance in highly

specialized tasks. To address this, we've fostered close collaboration with expert clinicians and used prompt engineering to assist the model in understanding and handling complex medical information. We envision LLMs not as replacements for healthcare professionals, but rather as effective aid for clinical decision-making when properly guided. Future technological advancements, such as parameter-efficient fine-tuning for specialized tasks and the use of vectorized databases, may further contribute to solving these issues, offering better solutions for data security and private model deployment.

As we move forward, our findings open up avenues to further refine prompt engineering techniques to optimize LLMs for analyzing patient data and medical literature to recommend evidence-based treatments for digestive system cancers. We aim to explore how different prompts impact the model's ability to accurately recommend optimal interventions based on tumor characteristics and patient factors. For instance, certain prompts may enhance the model's capacity to suggest appropriate surgical procedures depending on tumor size, location, and staging. Other prompts could optimize the recommendation of systemic therapies like chemotherapy regimens and radiation therapy protocols tailored to the individual's medical history and cancer biomarkers. Advances in prompt engineering to account for all relevant clinical variables could enable the generation of more personalized and effective treatment plans for each unique patient case. However, more research is still urgently needed to ensure patient safety, avoid biases, and enable reliable interpretation of model outputs before these systems are ready for real-world clinical implementation. We must rigorously test prompts to identify any that skew recommendations in inappropriate or unsafe ways. Transparent reporting of model limitations and close collaboration with medical experts will be critical to responsible prompt engineering. While our results demonstrate immense promise for LLMs to enhance evidence-based decision support, translating these tools into practice will require thoughtful and ethical design paired with extensive validation to evolve prompt engineering strategies that provide trustworthy guidance without ever replacing human clinical judgment. Overall, steering LLMs through carefully crafted prompts shows great potential to augment clinicians' abilities to optimize and personalize treatment plans, propelling more effective cancer care. But as this technology continues maturing, maintaining patient well-being through rigorous prompt optimization and evaluation remains imperative.

CONCLUSION

This study has underscored the potential and challenges associated with the application of prompt engineering techniques to large language models (LLMs) in the field of clinical oncology. Through careful crafting of simple, templated prompts and more complex strategies, like in-context learning (ICL) and multi-round interaction, we have seen promising capabilities of these models in processing and interpreting intricate medical data related to gastrointestinal cancers. This can substantially support healthcare professionals in making decisions about recommended treatments. However, it is crucial to continuously address the inherent limitations of these models, including potential biases, data privacy concerns, and their specific interpretative limitations in this clinical context. Although complex prompts, especially those allowing for iterative questioning, have shown great promise in optimizing the performance of LLMs, it's evident that further investigations are needed to refine these strategies and explore their potential integrations. As our study was conducted in a clearly defined and constrained environment to ensure consistency, further exploration in diverse settings is warranted to fully exploit the potential of LLMs in healthcare scenarios.

REFERENCES

1. OpenAI, R. (2023). GPT-4 technical report. arXiv 2303.08774. DOI: 10.48550/arXiv.2303.08774.
2. Lee, P., Bubeck, S., and Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N. Engl. J. Med.* **388**: 1233-1239. DOI: 10.1056/NEJMs2214184.
3. Lee, P., Goldberg, C., and Kohane, I. (2023). *The AI revolution in medicine: GPT-4 and beyond* (Pearson Education, Limited).
4. Xu, Y., Liu, X., Cao, X., et al. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation* **2**: 100179. DOI: 10.1016/j.xinn.2021.100179.
5. Nori, H., King, N., McKinney, S.M., et al. (2023). Capabilities of GPT-4 on medical

- challenge problems. arXiv preprint arXiv:2303.13375. DOI: 10.48550/arXiv.2303.13375.
6. Ayers, J.W., Poliak, A., Dredze, M., et al. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**: 589-596. DOI: 10.1001/jamainternmed.2023.1838.
 7. Haver, H.L., Ambinder, E.B., Bahl, M., et al. (2023). Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology* **307**: e230424. DOI: 10.1148/radiol.230424.
 8. Zhu, L., Mou, W., and Chen, R. (2023). Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *J. Transl. Med.* **21**: 1-4. DOI: 10.1186/s12967-022-03835-4.
 9. Uprety, D., Zhu, D., and West, H.J. (2023). ChatGPT—a promising generative AI tool and its implications for cancer care. *Cancer* **129**: 2284-2289. DOI: 10.1002/cncr.34827.
 10. Zhong, Y., Chen, Y.J., Zhou, Y., et al. (2023). The artificial intelligence large language models and neuropsychiatry practice and research ethic. *Asian J. Psychiatr.* **84**: 103577. DOI: 10.1016/j.ajp.2023.103577.
 11. Young, J.N., Ross, O.H., Poplousky, D., et al. (2023). The utility of ChatGPT in generating patient-facing and clinical responses for melanoma. *J. Am. Acad. Dermatol.* **89**: 602-604. DOI: 10.1016/j.jaad.2023.05.024.
 12. Xie, Y., Seth, I., Hunter-Smith, D.J., et al. (2023). Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. *Aesth. Plast. Surg.* **47**: 1985-1993. DOI: 10.1007/s00266-023-03338-7.
 13. Buzzaccarini, G., Degliuomini, R.S., and Borin, M. (2023). The artificial intelligence application in aesthetic medicine: How ChatGPT can revolutionize the aesthetic world. *Aesth. Plast. Surg.* **47**: 2211-2212. DOI: 10.1007/s00266-023-03416-w.
 14. Radford, A., Wu, J., Child, R., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog* **1**: 9. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
 15. Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *NeurIPS* **33**: 1877-1901. DOI: 10.48550/arXiv.2005.14165.
 16. Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS* **33**: 9459-9474. DOI: 10.48550/arXiv.2005.11401.
 17. Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS* **35**: 24824-24837. DOI: 10.48550/arXiv.2201.11903.
 18. Zhou, D., Schärli, N., Hou, L., et al. (2022). Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625. DOI: 10.48550/arXiv.2205.10625.
 19. Yao, S., Yu, D., Zhao, J., et al. (2023). Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601. DOI: 10.48550/arXiv.2305.10601.
 20. Fu, Y., Peng, H., Sabharwal, A., et al. (2022). Complexity-based prompting for multi-step reasoning. arXiv preprint arXiv:2210.00720. DOI: 10.48550/arXiv.2210.00720.
 21. Khot, T., Trivedi, H., Finlayson, M., et al. (2022). Decomposed prompting: A modular approach for solving complex tasks. arXiv preprint arXiv:2210.02406. DOI: 10.48550/arXiv.2210.02406.
 22. White, J., Fu, Q., Hays, S., et al. (2023). A prompt pattern catalog to enhance prompt

engineering with ChatGPT. arXiv preprint arXiv:2302.11382. DOI: 10.48550/arXiv.2302.11382.

23. Suzgun, M., Scales, N., Schärli, N., et al. (2022). Challenging big-bench tasks and whether chain-of-thought can solve them. arXiv preprint arXiv:2210.09261. DOI: 10.48550/arXiv.2210.09261.

FUNDING AND ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (91959205 to L.S., U22A20327 to L.S., 82203881 to Y.C., 82272627 to X.T.Z., 7232018 to Y.S., 12090022 to B.D., 11831002 to B.D., 81801778 to L.Z.), Beijing Natural Science Foundation (7222021 to Y.C., Z200015 to X.T.Z.), Beijing Hospitals Authority Youth Programme (QML20231115 to Y.C.), Clinical Medicine Plus X-Young Scholars Project of Peking University (PKU2023LCXQ041 to Y.C. and L.Z.). Guangdong Provincial Key Laboratory of Precision Medicine for Gastrointestinal Cancer (2020B121201004). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

J. Yuan, P. Bao, Z. Chen, M. Yuan, and J. Pan contributed to data analysis and interpretation, and drafted the manuscript. J. Zhao, Y. Xie, Y. Cao, Y. Wang, Z. Wang, Z. Lu, X. Zhang, J. Li and L. Ma performed the sample preparation. Y. Chen, L. Zhang, L. Shen, and B. Dong planned the study and participated in manuscript revision. All authors have given final approval for the manuscript to be published and have agreed to be responsible for all aspects of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

Not applicable.

DATA AND CODE AVAILABILITY

A pre-trained Chinese BERT model in Hugging Face (<https://huggingface.co/hfl/chinese-bert-wwm-ext>), specifically the "hfl/chinese-bert-wwm-ext", is utilized to translate patient conditions into a high-dimensional embedding space (768 dimensions in this study), capturing the context of the condition effectively.

SUPPLEMENTAL INFORMATION

It can be found online at <https://doi.org/10.59717/j.xinn-med.2023.100019>

LEAD CONTACT WEBSITE

<http://faculty.bicmr.pku.edu.cn/~dongbin/>

Region-Aware Metric Learning for Open World Semantic Segmentation via Meta-Channel Aggregation

Hexin Dong^{1*}, Zifan Chen^{1*}, Mingze Yuan¹, Yutong Xie¹, Jie Zhao^{1,2}, Fei Yu¹,
Bin Dong^{4,3,2} and Li Zhang^{1,2†}

¹Center for Data Science, Peking University, Beijing, China

²National Biomedical Imaging Center, Peking University, Beijing, China

³Center for Machine Learning Research, Peking University, Beijing, China

⁴Beijing International Center for Mathematical Research (BICMR), Peking University, Beijing, China

Abstract

As one of the most challenging and practical segmentation tasks, open-world semantic segmentation requires the model to segment the anomaly regions in the images and incrementally learn to segment out-of-distribution (OOD) objects, especially under a few-shot condition. The current state-of-the-art (SOTA) method, Deep Metric Learning Network (DMLNet), relies on pixel-level metric learning, with which the identification of similar regions having different semantics is difficult. Therefore, we propose a method called region-aware metric learning (RAML), which first separates the regions of the images and generates region-aware features for further metric learning. RAML improves the integrity of the segmented anomaly regions. Moreover, we propose a novel meta-channel aggregation (MCA) module to further separate anomaly regions, forming high-quality sub-region candidates and thereby improving the model performance for OOD objects. To evaluate the proposed RAML, we have conducted extensive experiments and ablation studies on *Lost And Found* and *Road Anomaly* datasets for anomaly segmentation and the *CityScapes* dataset for incremental few-shot learning. The results show that the proposed RAML achieves SOTA performance in both stages of open world segmentation. Our code and appendix are available at <https://github.com/czifan/RAML>.

1 Introduction

The breakthrough of deep learning in many fields of computer vision is based on the closed set assumption, which means that all classes in the test should be covered in the training set. However, this assumption rarely holds in the open world. Since most computer vision applications have to deal with unknown classes, models, especially the deep models, must

* Equal contribution.

†Correspondence to Li Zhang: zhangli_pku@pku.edu.cn

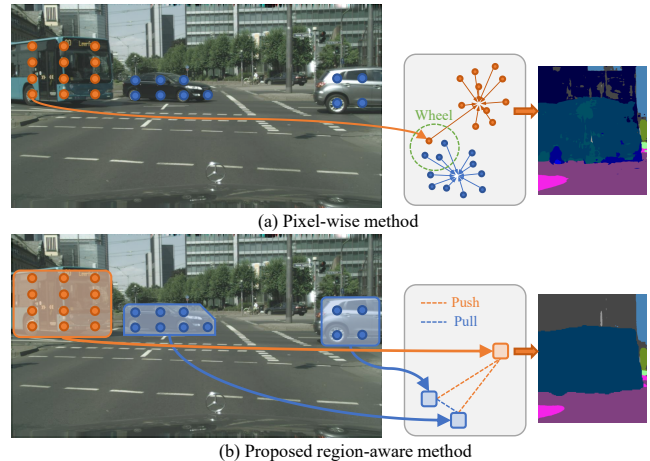


Figure 1: Main idea of our proposed method. (a) Existing methods focus on pixel-level which may result in fine-grained segmentation errors. (b) Our proposed Region-Aware Metric Learning (RAML) method maintains the semantic integrity of the OOD objects.

handle the out-of-distribution (OOD) data. Quite a number of work for image recognition and classification in the open world has been proposed since the first introduction of the concept “open world” in [Bendale and Boulton, 2015]. However, the work about open world segmentation is scarce. It is not until recently that [Cen *et al.*, 2021] proposes a two-step framework to achieve open world semantic segmentation. The framework consists of (1) an **anomaly segmentation** module that extends the close-set model of in-distribution objects to delineate the unknown regions of the OOD objects correctly, and (2) an **incremental few-shot learning** module that separates the unknown regions into OOD objects with novel classes. They also introduce metric learning into both stages of open world segmentation, and the results prove that their proposed criteria of metric learning can improve the model’s segmentation of OOD objects.

Although this pilot work provides a good framework for open world segmentation tasks, the model can be improved in two aspects for better performance. First, the metric learning in [Cen *et al.*, 2021] relies on the pixel-wise feature embeddings, which may falsely split the object into pieces and re-

sult in numerous fine-grained segmentation errors. For example, as shown in Figure 1, the *bus wheels* and the *car wheels* have similar feature embeddings and are highly likely to be classified into one group according to the pixel-wise feature embeddings, but they apparently belong to different classes in semantic segmentation. To solve this kind of problems, we propose region-aware metric learning (RAML) for open world segmentation, which significantly outperforms pixel-wise metric learning (PML) in multiple experiments.

Moreover, we improve the model performance, especially for the incremental few-shot learning stage, by introducing a novel region separation module named meta-channel aggregation (MCA). MCA first aims at over-segmenting the unknown regions into several meta channels. Regions belonging to different meta channels are aggregated to form a segmentation of the objects and then evaluated by the Region-aware Metric Learning module.

In addition, [Cen *et al.*, 2021] sets a fixed center embedding for each in-distribution class, i.e., a one-hot vector in the feature space. Although the fixed center embedding can effectively create a distance between the distribution of different classes, it ignores the relative similarity between them. For example, in the *Cityscapes* dataset, the method fails to reveal that the difference between *person* and *rider* is smaller than the difference between either of them and *sky*. This paper aims to overcome the drawback by exploiting a more natural metric learning to constrain the distance between the inter-class region-aware features. Specifically, we replace the one-hot setting in [Cen *et al.*, 2021] with Circleloss [Sun *et al.*, 2020] as the objective of the metric learning, which not only maintains a fine inter-class distance but also shapes the intra-class distribution more concentrated. Experiments show that such division of the feature space is more conducive to segmenting the OOD data.

In summary, we propose a region-aware metric learning method for open world semantic segmentation. Our contributions are as follows:

- We propose using the region-aware over pixel-wise features for open world semantic segmentation to ensure better semantic integrity of the segmented OOD objects.
- We introduce the MCA module as a novel region separation method that suits incremental few-shot learning.
- We adopt Circleloss [Sun *et al.*, 2020] to enlarge the inter-class distance and reduce the intra-class distance of the data samples, improving the performance of the RAML module.

2 Related Work

2.1 Region-aware Semantic Segmentation

The ideas of how to apply regional information to improve semantic segmentation have been discussed by many research groups recently, including two main threads. First, several works have shown that region-aware information has better contextual representation than pixel-level information to achieve pixel labeling [Yuan *et al.*, 2020]. Secondly, for image segmentation tasks, region-aware information can be better combined with metric or contrastive learning to manip-

ulate the feature space more effectively [Wang *et al.*, 2021; Hu *et al.*, 2021]. These ideas inspire our paper, but the above works require a sufficient number of training samples to obtain the reasonable region-aware feature representation, while our work is in an open world setting that can only access a few images with unseen class labels. Therefore, we have to design novel region-separation modules (such as MCA) that fit the open world segmentation tasks.

2.2 Anomaly Segmentation

There are two types of approaches for anomaly segmentation, including uncertainty-based methods and generative model-based methods. Uncertainty refers to the level of not belonging to known classes, widely used to determine abnormal states. The baseline of uncertainty-based methods is maximum softmax probability (MSP) reported by [Hendrycks and Gimpel, 2017]. [Hendrycks *et al.*, 2019] then improves MSP using maximum logit (MaxLogit) for better performance on large-scale datasets. Other uncertainty-based methods include using Bayesian neural networks [Gal and Ghahramani, 2016] and maximizing the entropy of OOD objects in the images [Chan *et al.*, 2021]. On the other hand, generative model-based methods also perform well, including autoencoder (AE) [Baur *et al.*, 2018] and GAN-based methods [Xia *et al.*, 2020]. However, generative models suffer from unstable training and usually have complex network backbones.

In this work, we follow the idea of MaxLogit and develop our anomaly segmentation based on non-normalized logit.

2.3 Open World Problem

[Bendale and Boulton, 2015] is the first research that gives the formal definition of “open world”, i.e., an open world model must incrementally learn and extend its generality, thereby making the objects with novel classes “known” to itself. Since then, the research on open world problems has increased, including classification [Zhong *et al.*, 2021], object detection [Joseph *et al.*, 2021], instance segmentation [Saito *et al.*, 2021], among others. However, it is not until recently that [Cen *et al.*, 2021] proposes the first framework of open world semantic segmentation. Our work follows the settings in [Cen *et al.*, 2021] and divides the problem into anomaly segmentation and incremental few-shot learning. However, to ensure semantic integrity and improve the segmentation performance, we use region-aware feature embedding instead of pixel-wise feature extraction in their original method.

2.4 Metric Learning

Deep metric learning constrains the distance between feature embedding of learning samples to manipulate the feature distribution. Its applications are seen in various computer vision tasks, such as open set recognition [Chen *et al.*, 2020], few-shot learning [Oreshkin *et al.*, 2018] and open world semantic segmentation [Cen *et al.*, 2021]. Classic metric learning includes two paradigms. The first is to learn with pair-wise labels, under the guidance of triplet loss [Schroff *et al.*, 2015] and center loss [Wen *et al.*, 2016]. The second consists of softmax cross-entropy and variants that train the model with class-level labels. A recently proposed method called Circle loss [Sun *et al.*, 2020] unifies the above two paradigms and

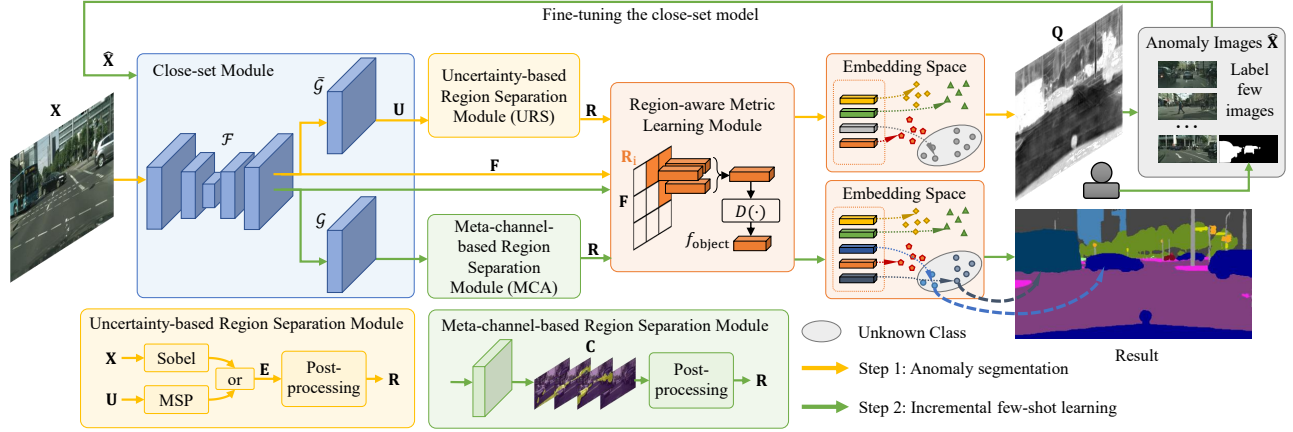


Figure 2: The pipeline of Region-aware Metric Learning for Open World Semantic Segmentation: 1) train a close-set segmentation model with known classes (*bluish square*); 2) **Anomaly Segmentation** (*yellowish arrows*): separate regions based on edge prediction (*yellowish squares*) and segment the anomaly regions using metric learning (*orangish squares*); 3) annotate for unknown objects ($\hat{\mathbf{X}}$) to fine-tune the close-set model; 4) **Incremental few-shot learning** (*greenish arrows*): separate regions based on MCA (*greenish squares*) and segment the OOD objects using metric learning (*orangish squares*). (Best view in color)

forms the feature space with large inter-class distances and small intra-class distances. We thus adopt Circle loss as the key objective of our proposed RAML module.

3 Methods

As shown in Figure 2, our proposed method contains: 1) a backbone model for close-set segmentation, 2) an anomaly segmentation process to delineate the unknown regions of OOD data, and 3) an incremental few-shot learning step for splitting the unknown regions into objects with novel classes.

3.1 Close-set Segmentation Module

Suppose $\mathcal{C}_{in} = \{C_{in,1}, C_{in,2}, \dots, C_{in,N}\}$ are N in-distribution classes, which are all annotated in training datasets, and $\mathcal{C}_{out} = \{C_{out,1}, C_{out,2}, \dots, C_{out,M}\}$ are M novel classes not involved in the training datasets. Here, the semantic segmentation network \mathcal{S} is divided into a feature extractor \mathcal{F} and a label predictor \mathcal{G} , where $\mathcal{S} = \mathcal{G} \circ \mathcal{F}$.

For the close-set segmentation, we minimize the following loss $\mathcal{L}_{seg}(\mathcal{F}, \mathcal{G})$ which guides \mathcal{S} to produce a pixel-level segmentation for in-distribution classes.

$$\mathcal{L}_{seg}(\mathcal{F}, \mathcal{G}) = \mathbb{E}_{\mathbf{X}, \mathbf{Y}}(\ell_{ce}(\mathcal{G} \circ \mathcal{F}(\mathbf{X}), \mathbf{Y})) \quad (1)$$

where $\ell_{ce}(\cdot, \cdot)$ indicates the multi-class cross entropy loss, $\mathbf{X} \in \mathbb{R}^{3 \times H \times W}$ is an input image, \mathbf{Y} is the corresponding label.

After training this module, we obtain a trained feature extractor \mathcal{F} and a trained label predictor \mathcal{G} . The feature map $\mathbf{F} = \mathcal{F}(\mathbf{X}) \in \mathbb{R}^{N_1 \times H \times W}$ and the non-normalized logit $\mathbf{U} = \mathcal{G}(\mathbf{F}) \in \mathbb{R}^{N \times H \times W}$ can then be generated for in-distribution classes, where $\bar{\mathcal{G}}$ is obtained by removing the softmax layer of \mathcal{G} . The feature map \mathbf{F} and the non-normalized logit \mathbf{U} will be used in later modules.

3.2 Anomaly Segmentation

To identify the candidate regions of region-aware anomaly segmentation, we adopt an uncertainty-based OOD objects



Figure 3: Visual examples of maximum softmax probability. Borders between objects have higher uncertainty because the semantics of the borders are usually ambiguous.

detection method, MSP [Hendrycks and Gimpel, 2017], as our region separation module, named Uncertainty-based Region Separation (URS). Its high uncertainty response around the object edges could be used as a promising initialization of the region separation, as shown in Figure 3.

To further enhance the edges, we introduce Sobel filtering over the original input image. The final edge prediction map \mathbf{E} can be generated as follow,

$$\mathbf{E} = \mathbb{I}(\text{Sobel}(\mathbf{X}) \geq \alpha \text{ or } \text{MSP}(\mathbf{U}) \geq \beta), \quad (2)$$

where \mathbf{X} is the input image, \mathbf{U} is the non-normalized logit, and $\mathbb{I}(\cdot)$ is an indicator function, α and β are hyper-parameters to control the edge prediction. According to \mathbf{E} , we use a post-processing sub-module, including the hole filling and connected component algorithms, to generate the candidate regions $\mathcal{R} = \{\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_T\}$, where $\mathbf{R}_i \in \{0, 1\}^{H \times W}$ represents the i -th region.

We then propose a RAML module for anomaly segmentation to classify the candidate regions \mathcal{R} . For each region $\mathbf{R}_i \in \{0, 1\}^{H \times W}$, the region-aware feature embedding is obtained as below:

$$f_{object} = \mathcal{D}\left(\frac{\sum_{j,k} \mathbf{F}^{j,k} \mathbf{R}_i^{j,k}}{\sum_{j,k} \mathbf{R}_i^{j,k}}\right) \in \mathbb{R}^{N_2} \quad (3)$$

where $\mathbf{F}^{j,k} \in \mathbb{R}^{N_1}$ is the feature vector of pixel (j, k) , $\mathcal{D}(\cdot)$ consists of two fully-connected layers to control the embedding dimension. f_{object} is compared to all the prototypes of the known classes by metric learning constrained by circle loss [Sun *et al.*, 2020]. Specifically, the prototype of l -th known class f_l can be obtained using the semantic segmentation label. Then, the region-aware anomaly probability of \mathbf{R}_i can be expressed as below,

$$\mathcal{P}(\mathbf{R}_i, \mathbf{F}) = \max_{1 \leq l \leq N} \frac{f_{object} \cdot f_l}{\|f_{object}\| \|f_l\|}. \quad (4)$$

Finally, to generate a pixel-level anomalous probability map, we combine the information from the non-normalized logit and the above region-aware anomaly probabilities. For each pixel (j, k) , uncertainty intensity $\mathbf{Q}^{j,k}$ is computed as,

$$\mathbf{Q}^{j,k} = - \max_{1 \leq l \leq N} \mathbf{U}_{(l)}^{j,k} \cdot \mathcal{P}(\mathbf{R}_i, \mathbf{F}), \quad (5)$$

where the pixel (j, k) belongs to region \mathbf{R}_i , \mathbf{F} is the feature map, $\mathcal{P}(\cdot, \cdot)$ is the region-aware anomaly probabilities. $\mathbf{U}_{(l)}^{j,k}$ is the l -th output of pixel (j, k) in the non-normalized logit \mathbf{U} . We then normalize the uncertainty intensity $\mathbf{Q}^{j,k}$ for each pixel to obtain the anomalous probability map, which is used to identify the unknown regions in the image.

3.3 Incremental Few-shot Learning via MCA

After the anomaly segmentation, open world semantic segmentation requires the model to identify all objects of M novel classes in the unknown regions. One way to realize the incremental few-shot learning is to use a few labeled images containing objects with novel classes to fine-tune the close-set segmentation model under the loss \mathcal{L}_{seg} . However, experiments show that this improvement is trivial. We thus propose an innovative MCA module for further creating sub-regions in the unknown regions from anomaly images $\tilde{\mathbf{X}}$. MCA takes the prediction of the label predictor \mathcal{G} in the close-set model as its input to output $(N + K)$ channels with softmax activation $\mathbf{C} \in [0, 1]^{(N+K) \times H \times W}$. The first N channels are the segmentation results for all in-distribution classes, while the last K ($K > M$) channels are *meta channels* to overly segment the unknown regions. Several MCA-related losses are integrated into \mathcal{L}_{seg} during the fine-tuning, and the overall loss function is,

$$\mathcal{L}_{overall} = \mathcal{L}_{seg} + \lambda_{inter} \mathcal{L}_{inter} + \lambda_{split} \mathcal{L}_{split} + \lambda_{rec} \mathcal{L}_{rec}. \quad (6)$$

The first term \mathcal{L}_{seg} is the segmentation loss for all in-distribution classes from Equation 1. The second term utilizes the negative of Dices to minimize the intersection between any pairs of output channels, which is defined as:

$$\mathcal{L}_{inter} = \sum_{1 \leq i < j \leq N+K} (1 - \ell_{dice}(\mathbf{C}_i, \mathbf{C}_j)) \quad (7)$$

where $\ell_{dice}(\cdot, \cdot)$ indicates the dice loss and $\mathbf{C}_i, \mathbf{C}_j$ are the i -th and j -th channels of the segmentation output.

The third term aims to avoid the sub-regions (candidates of OOD objects) gathering in a few certain channels:

$$\mathcal{L}_{split} = \sum_{i=N+1}^{N+K} -\log(\max(\eta \sum_{j,k} \mathbf{C}_i^{j,k}, 1)) \quad (8)$$

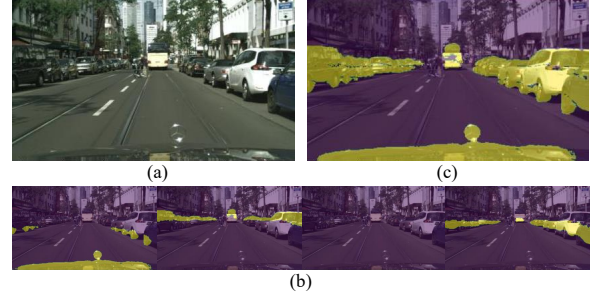


Figure 4: Visualization results of MCA. (a) Input image; (b) Meta-channel response ($K = 4$); (c) Aggregated Meta-channel.

where $\mathbf{C}_i^{j,k}$ represent (j, k) pixel output of i -th channel and η is a hyper-parameter to control the separation. \mathcal{L}_{split} reaches the minimum when the sub-regions scatter across the output channels according to Jensen's inequality.

The last term encourages the outputs of all channels to reconstruct the entire image, further avoiding loss of information:

$$\mathcal{L}_{rec} = \|\mathbf{X} \odot (\sum_{i=1}^{N+K} \mathbf{C}_i - \mathbf{1}_{H \times W})\|^2 \quad (9)$$

where \odot is the element-wise multiplication operator and $\mathbf{1}_{H \times W}$ is a matrix with all ones.

As shown in Figure 4, we observe that MCA tends to segment objects based on local semantic information. One unknown object may be segmented into more than one channel and lose completeness. (e.g., The windows and wheels of cars may be divided into different channels.) Therefore, we aggregate the sub-regions from certain meta channels according to few-shot (here L -shot) labeled images, which generates the candidate regions $\mathcal{R} = \{\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_T\}$ for the final RAML module of incremental few-shot learning.

Similar to Equation 3, the region-aware feature embedding f_{object} for each region \mathbf{R}_i could be computed. The prototype of i -th unknown class ($1 \leq i \leq M$) from L -shot newly labeled images is defined as:

$$c_i = \frac{1}{L} \sum_{j=1}^L f_i^{(j)} \quad (10)$$

where $f_i^{(j)}$ represents the feature embedding of i -th unknown class in j -th annotated image. For each region-aware feature embedding f_{object} , we use cosine similarity to measure the distance between this candidate region and every unknown class:

$$s_{object}^i = \frac{f_{object} \cdot c_i}{\|f_{object}\| \|c_i\|}, i = 1, 2, \dots, M \quad (11)$$

The candidate region can be classified as the i -th novel class $C_{out,i}$ only if the cosine similarities satisfy the following two criteria:

$$\begin{cases} s_{object}^i > \theta_{novel} \\ s_{object}^i > s_{object}^{i'} \quad \forall i' \neq i \end{cases} \quad (12)$$

where θ_{novel} is a hyper-parameter to control classification.

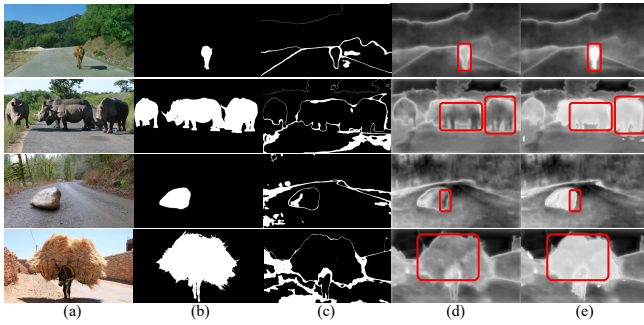


Figure 5: Visualization results of anomaly segmentation on *Road Anomaly*. (a) input image; (b) ground truth; (c) edge prediction; (d) results of MaxLogit [Hendrycks *et al.*, 2019]. (e) results of our proposed RAML method. For (d) and (e), higher value represents greater probability of anomaly. The red bounding boxes indicate that RAML ensures the integrity of the anomaly regions.

Dataset	<i>Lost and Found</i>			<i>Road Anomaly</i>		
	AUPR \uparrow	AUROC \uparrow	FPR95 \downarrow	AUPR \uparrow	AUROC \uparrow	FPR95 \downarrow
Ensemble	-	57	-	-	67	-
RBM	-	86	-	-	59	-
MSP	21	83	31	19	70	61
MaxLogit	37	91	21	32	78	49
DUIR	-	93	-	-	83	-
DML	45	97	10	37	84	37
RAML(Ours)	46	97	8	42	86	32

Table 1: Results of anomaly segmentation on *Lost and Found* and *Road Anomaly*.

4 Experiments

Our experiments include three parts: (1) experimental results of anomaly segmentation in subsection 4.1; (2) experimental results of incremental few-shot learning results in subsection 4.2; (3) ablation studies in subsection 4.3 and Appendix.

4.1 Anomaly Segmentation

Datasets. 7000 full-frame annotated driving scenes from *BDD100k* [Yu *et al.*, 2020] are used to train the close-set segmentation model, containing 19 categories of objects as in-distribution objects. For anomaly segmentation, we use another two road scene datasets, *Lost and Found* [Pinggera *et al.*, 2016] and *Road Anomaly* [Lis *et al.*, 2019], with anomalous objects other than ones in *BDD100k*.

Implementation details. We follow [Hendrycks *et al.*, 2019; Cen *et al.*, 2021] to use PSPNet as the network backbone of our close-set segmentation module and apply two fully connected layers for RAML. We follow [Hendrycks and Gimpel, 2017] to use three metrics to evaluate the performance of anomaly segmentation, including area under ROC curve (AUROC), area under the precision-recall curve (AUPR), and the false-positive rate at 95% recall (FPR95).

Results. As shown in Table 1, our proposed RAML module achieves the SOTA performance on *Lost and Found* and *Road Anomaly* for anomaly segmentation. Figure 5 presents some visual examples to compare RAML and the pixel-wise method. The proposed RAML module produces higher response values and better integrity within the anomalous objects, significantly reducing the false-negative cases.

4.2 Incremental Few-shot Learning

Datasets. we use *Cityscapes* dataset to train and evaluate our RAML module in the incremental few-shot learning step. *Cityscapes* consists of 2975 real-world images in the training set and 500 in the validation set with a resolution of 2048×1024 . The division of training set and test set in our experiments is consistent with this division.

Implementation details. We follow [Cen *et al.*, 2021] to train a DeeplabV3+ model as the close-set model, which is followed by two fully connected layers for RAML and use mean Intersection-over-Union (mIoU) to evaluate the performance of segmentation results. Specifically, $mIoU_{old}$ and $mIoU_{novel}$ are the mIoUs of known and unknown classes, respectively. The metric $mIoU_{harm}$ is a comprehensive index [Xian *et al.*, 2019] that balances $mIoU_{old}$ and $mIoU_{novel}$.

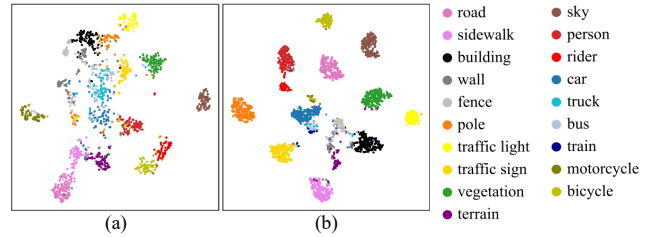


Figure 6: t-SNE visualization for (a) pixel-wise NPM method and (b) our proposed RAML method. All learned metrics of 19 classes of the *Cityscapes* dataset are included, where *car*, *truck* and *bus* are OOD classes.

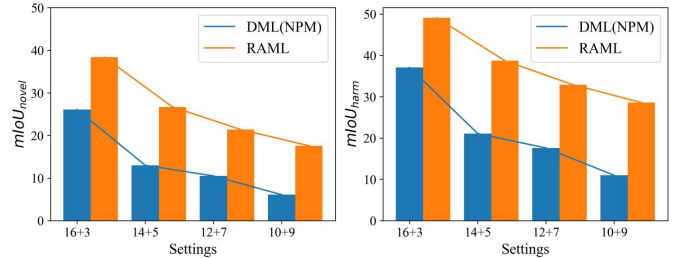


Figure 7: Ablation study of the ratio of unknown classes to known classes. We compare our method to NPM and report results with $mIoU_{novel}$ and $mIoU_{harm}$.

Results. We test our method on *CityScapes* and compare our method to pixel-wise NPM and PLM proposed by [Cen *et al.*, 2021]. In our experiment, *car*, *truck*, and *bus* are 3 OOD classes not involved in the training stage while the other 16 classes are regarded as in-distribution classes. As shown in Table 2, our proposed RAML module outperforms the previous methods with a relatively large margin. According to Figure 8, pixel-wise metric learning shows erroneous broken segmentation results on OOD objects, while the proposed RAML demonstrates a remarkable ability to maintain the integrity of these results. In addition, Figure 6 shows that the feature embeddings produced by the proposed RAML maintain a reasonable inter-class distance and their intra-class distributions

16+1 setting		road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	train	motorcycle	bicycle	car	truck	bus	mIoU _{all}	mIoU _{novel}	mIoU _{old}	mIoU _{ham}
Baseline	All 17	97.8	82.4	91.8	52.3	57.5	59.9	64.1	74.2	91.9	61.4	94.6	79.4	58.8	75.6	61.7	74.9	94.8	-	-	74.9	-	-	-
	First 16	98.0	82.1	91.4	43.6	56.4	58.9	61.4	72.6	91.6	60.5	94.4	79.1	57.6	67.9	61.1	75.1	-	-	-	72.0	-	-	-
	FT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.6	-	-	0.4	6.6	0.0	0.0
5 shot	PLM	97.1	79.3	89.2	41.9	55.3	57.5	60.8	71.0	91.1	59.4	93.9	73.3	49.2	34.2	14.3	51.8	75.7	-	-	64.4	75.7	63.7	69.2
	NPM	96.2	79.3	89.2	41.6	52.0	56.3	61.1	69.4	90.4	58.8	94.1	74.4	55.3	53.4	39.2	70.3	64.6	-	-	67.4	64.6	67.6	66.1
	RAML(Ours)	97.3	82.6	91.4	51.0	57.2	59.2	65.5	74.4	91.7	63.9	94.7	79.1	59.1	23.7	52.1	72.3	85.2	-	-	70.6	85.2	69.7	76.7
1 shot	PLM	96.8	77.1	89.6	41.4	48.7	53.2	60.3	64.5	90.3	55.6	94.3	59.1	43.6	39.5	12.0	35.7	64.5	-	-	60.4	64.5	60.1	62.2
	NPM	95.9	79.2	88.8	41.3	50.5	56.0	61.0	69.1	90.2	58.6	94.1	73.6	55.1	49.7	37.4	69.6	60.1	-	-	66.5	60.1	66.9	63.3
	RAML(Ours)	97.4	82.6	91.5	51.0	57.3	59.3	65.5	74.4	91.8	64.0	94.7	79.2	59.1	11.5	52.2	72.4	85.5	-	-	70.0	85.5	69.0	76.4

16+3 setting		road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	train	motorcycle	bicycle	car	truck	bus	mIoU _{all}	mIoU _{novel}	mIoU _{old}	mIoU _{ham}
Baseline	All 19	97.9	83.0	91.7	51.5	58.3	59.8	64.2	74.2	92.0	61.2	94.6	79.7	59.1	63.9	61.5	75.0	94.2	78.5	81.4	74.8	-	-	-
	First 16	98.0	82.1	91.4	43.6	56.4	58.9	61.4	72.6	91.6	60.5	94.4	79.1	57.6	67.9	61.1	75.1	-	-	-	72.0	-	-	-
	FT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.1	0.0	0.0
5 shot	PLM	97.1	79.2	84.8	38.1	46.4	56.8	58.8	61.0	91.0	59.3	92.9	63.6	47.5	3.4	13.8	47.5	67.0	5.7	12.0	54.0	28.2	58.8	38.1
	NPM	96.1	79.3	88.7	41.5	51.5	56.3	60.7	69.0	90.4	58.8	94.1	74.3	55.1	32.0	39.1	70.2	55.7	1.6	21.0	58.2	26.1	64.2	37.1
	RAML(Ours)	97.3	82.6	91.1	50.6	57.2	59.1	65.5	74.1	91.7	64.0	94.7	79.0	58.9	3.7	52.2	72.3	79.3	9.7	26.0	63.6	38.4	68.4	49.1
1 shot	PLM	96.8	75.2	49.0	33.1	31.4	48.0	33.2	44.6	89.7	55.3	23.0	42.1	32.8	5.3	8.0	27.7	30.4	0.7	9.5	38.7	13.5	43.4	20.6
	NPM	95.8	79.2	44.6	41.2	50.2	56.0	60.5	67.5	90.1	58.6	94.0	73.5	54.9	24.9	37.2	69.6	54.5	1.1	22.0	56.6	25.9	62.3	36.5
	RAML(Ours)	97.4	82.6	91.3	50.3	56.0	59.2	65.5	74.1	91.7	63.9	94.7	79.1	58.9	3.9	52.2	72.4	80.9	5.5	23.0	63.2	36.5	68.3	47.5

Table 2: Incremental few-shot learning results on *Cityscapes* for 16+1 setting (OOD class is *car*) and 16+3 setting (OOD classes are *car*, *truck*, *bus*). The unknown classes are in blue. Finetune (FT) is the baseline with catastrophic forgetting.

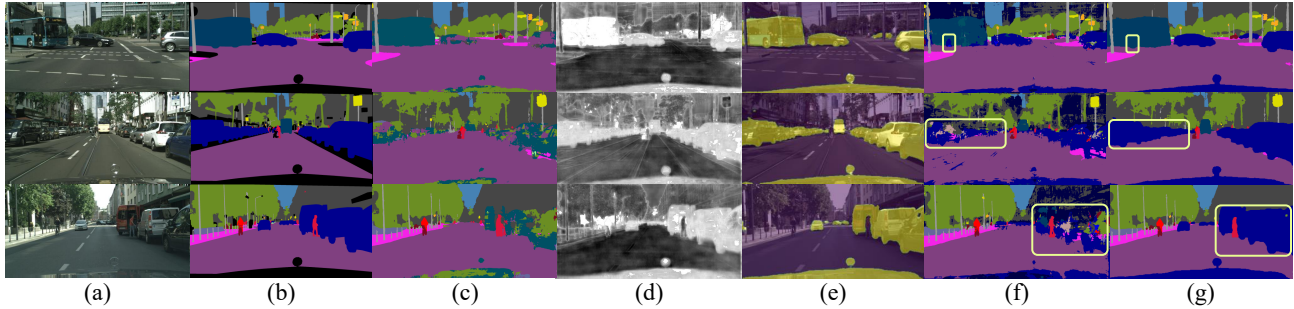


Figure 8: Visual examples of RAML for **open world semantic segmentation**: (a) input images. (b) ground truth. (c) close-set outputs. (d) anomaly segmentation outputs. (e) MCA outputs. (f) results of pixel-wise NPM [Cen *et al.*, 2021]. (g) results of our RAML module. Yellow boxes indicate that RAML method can better ensure the integrity of the OOD objects. For example, in the first row, the pixel-wise method mistakenly divides the wheels of the bus into cars, while RAML can correctly segment the entire bus. (Best view in color and zoom in.)

are also more concentrated. Such feature distribution could foster the model to obtain a robust decision boundary.

4.3 Ablation Study

Ratio of unknown classes to known classes. The performance of the trained segmentation model has highly correlated with the amount of training information. We compare our proposed RAML method with the current SOTA method, NPM [Cen *et al.*, 2021], under the different ratios of unknown classes to known classes. As shown in Figure 7, although our RAML method has a decline in performance as the ratio increases, it outperforms NPM in all ratio settings.

Method	mIoU _{all}	mIoU _{novel}	mIoU _{old}	mIoU _{ham}
Baseline	49.1	1.5	58.0	2.9
+ L_{rec}	61.8	33.6	67.1	43.2
+ $L_{rec} + L_{split}$	62.6	37.6	67.3	48.3
+ $L_{rec} + L_{split} + L_{inter}$	63.6	38.4	68.3	49.1

Table 3: Ablation study of losses used in MCA Module. Baseline is using Close-set Module directly.

Losses in MCA. This section evaluates the losses of our MCA module. As shown in Table 3, the reconstruction loss

ensures that our model obtains all information for the unknown classes, significantly improving the validity of MCA. The intersection loss and split loss also bring relatively smaller gains by improving the distribution of candidate regions in meta channels.

5 Conclusion

We have proposed RAML to enhance the performance of open world semantic segmentation. The main reason is that the region-aware feature outperforms the pixel-wise feature on maintaining the semantic integrity of the segmented OOD objects. Effective region separation methods are needed to realize RAML on anomaly segmentation and incremental few-shot learning. We, therefore, adopt the classic uncertainty-based methods to extract candidate regions for anomaly segmentation and propose an MCA module to further separate the anomaly regions for incremental few-shot learning. Experimental results show that our proposed method achieves the SOTA performance on the anomaly segmentation and the overall open world semantic segmentation. Our method has the potential to boost the use of open world semantic segmentation in practical applications.

6 Acknowledgments

This work is supported by the Grants under the National Natural Science Foundation of China (NSFC) under Grants 12090022, 11831002, 71704023, and Beijing Natural Science Foundation (Z180001).

References

- [Baur *et al.*, 2018] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *MICCAI Brainlesion Workshop*, pages 161–169, 2018.
- [Bendale and Boulton, 2015] Abhijit Bendale and Terrance Boulton. Towards open world recognition. In *CVPR*, pages 1893–1902, 2015.
- [Cen *et al.*, 2021] Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. In *ICCV*, pages 15333–15342, 2021.
- [Chan *et al.*, 2021] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *ICCV*, pages 5128–5137, 2021.
- [Chen *et al.*, 2020] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *ECCV*, pages 507–522, 2020.
- [Gal and Ghahramani, 2016] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016.
- [Hendrycks and Gimpel, 2017] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.
- [Hendrycks *et al.*, 2019] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019.
- [Hu *et al.*, 2021] Hanzhe Hu, Jinshi Cui, and Liwei Wang. Region-aware contrastive learning for semantic segmentation. In *ICCV*, pages 16291–16301, 2021.
- [Joseph *et al.*, 2021] K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, 2021.
- [Lis *et al.*, 2019] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *ICCV*, pages 2152–2161, 2019.
- [Oreshkin *et al.*, 2018] Boris N Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, 2018.
- [Pinggera *et al.*, 2016] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In *IROS*, pages 1099–1106, 2016.
- [Saito *et al.*, 2021] Kuniaki Saito, Ping Hu, Trevor Darrell, and Kate Saenko. Learning to detect every thing in an open world. *arXiv preprint arXiv:2112.01698*, 2021.
- [Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.
- [Sun *et al.*, 2020] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, 2020.
- [Wang *et al.*, 2021] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, 2021.
- [Wen *et al.*, 2016] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515, 2016.
- [Xia *et al.*, 2020] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *ECCV*, pages 145–161, 2020.
- [Xian *et al.*, 2019] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata. Semantic projection network for zero- and few-label semantic segmentation. In *CVPR*, 2019.
- [Yu *et al.*, 2020] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pages 2636–2645, 2020.
- [Yuan *et al.*, 2020] Yuhui Yuan, Xiaokang Chen, Xilin Chen, and Jingdong Wang. Segmentation transformer: Object-contextual representations for semantic segmentation. In *ECCV*, 2020.
- [Zhong *et al.*, 2021] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *CVPR*, pages 9462–9470, 2021.

DAST: Unsupervised Domain Adaptation in Semantic Segmentation Based on Discriminator Attention and Self-Training

Fei Yu¹, Mo Zhang^{1,2}, Hexin Dong¹, Sheng Hu¹, Bin Dong^{1,2,3}, Li Zhang^{1,2}

¹ Center for Data Science, Peking University, Beijing, China

² Center for Data Science in Health and Medicine, Peking University, Beijing, China

³ Beijing International Center for Mathematical Research(BICMR), Peking University, Beijing, China
{yufei1900, zhangmo007, donghexin, hs95, zhangli_pku}@pku.edu.cn dongbin@math.pku.edu.cn

Abstract

Unsupervised domain adaption has recently been used to reduce the domain shift, which would ultimately improve the performance of semantic segmentation on unlabeled real-world data. In this paper, we follow the trend to propose a novel method to reduce the domain shift using strategies of discriminator attention and self-training. The discriminator attention strategy contains a two-stage adversarial learning process, which *explicitly* distinguishes the well-aligned (domain-invariant) and poorly-aligned (domain-specific) features, and then guides the model to focus on the latter. The self-training strategy adaptively improves the decision boundary of the model for target domain, which *implicitly* facilitates the extraction of domain-invariant features. By combining the two strategies, we find a more effective way to reduce the domain shift. Extensive experiments demonstrate the effectiveness of our proposed method on numerous benchmark datasets.

Introduction

Semantic segmentation is a classic computer vision task that aims to predict a semantic label for each pixel in an image. Despite the notable progress in this field driven by the rapid development of deep learning (Chen et al. 2018; Long, Shelhamer, and Darrell 2015; Ronneberger, Fischer, and Brox 2015; Wei et al. 2018), it remains challenging to apply segmentation model trained on the labeled source data to the unlabeled target/real-world data, which vary substantially in their illumination, style, and the context in different domains. One possible solution draws on supervised learning to retrain or fine-tune the pre-trained model, which however requires expensive and time-consuming pixel-level manual annotations. An alternative way is to use unsupervised domain adaptation (UDA) to reduce the domain shift, so as to train a model that is able to segment the target images without labels.

The key component of semantic segmentation using UDA methods is to align the features from different domains (Chen et al. 2019; Hoffman et al. 2018; Tsai et al. 2018; Vu et al. 2019). Although the main idea is straightforward — matching the overall feature-level distributions of the

source and the target domains, the difficulty of implementation varies with adapting features for different regions in an image. For example, the adaption is easier in the case of sky than in the cases of buildings, traffic lights, and sidewalks, because the regions of the sky are similar regardless of the images while the latter group is characterized with different architectural styles or traffic rules. (Luo et al. 2019b) believes that aligning the source domain and the target domain globally leads to negative transfer of information and undermines the performance of the model in the originally well-aligned regions. Therefore, they propose to generate a local alignment score map and allow different weights for regions with different local alignment scores.

Following the spirit of (Luo et al. 2019b), in this work, we propose a strategy called discriminator attention (DA), to directly evaluate whether the local features are hard-adapted. The proposed DA strategy includes two stages of adversarial learning — discovering and correcting. In the discovering phase, a discriminator network (also known as the discoverer, D) aligns the intermediate features of the segmentation network and uses the confidence of local alignment to form an attention map that reweights the feature maps for label prediction. In the correcting stage, another discriminator network (called corrector, C) further aligns the output of the segmentation network based on the previous attention map. As illustrated in Figure 1 (b), the model pays more attention to hard-adapted regions for domain alignment.

Considering that the distribution of *real-world* data (target domain) is over-complex, we further introduce a self-training strategy to guarantee that the decision boundary of the model is suitable for the target domain. As shown in Figure 1 (c), the decision boundary of the segmentation network after UDA still tends to favor the distribution of source domain data, but the tendency is corrected after we apply the self-training strategy. Specifically, we adaptively improve the model’s decision boundary by training the segmentation network with pseudo labels generated from the previous predictions.

In summary, we propose an effective and intuitive unsupervised domain adaptation method for semantic segmentation, combining the strategies of discriminator attention and self-training (DAST). The main contributions can be summarized as follows:

- We propose a novel two-stage adversarial learning (DA),

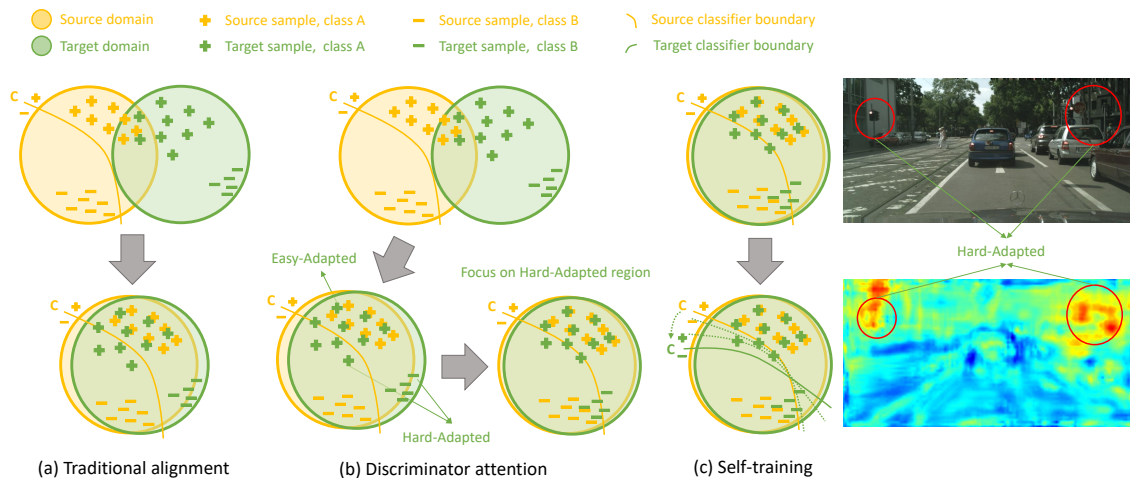


Figure 1: Illustration of the traditional adversarial learning methods and ours. (a) Traditional adversarial learning methods equally align the distribution among the entire images. (b) Our proposed discriminator attention can focus on the hard-adapted regions and achieve a better alignment. (c) We introduce a self-training module to learn an adaptive classifier boundary for the target domain, which can further improve the performance. (Best view in color)

which utilizes attention mechanism to attach higher weights to hard-adapted regions and simultaneously align the feature space and the output space.

- Our method is complementary to existing domain adaptation techniques, such as self-training.
- Our method achieves superior performance on the adaptation from SYNTHIA (Ros et al. 2016)/GTA5 (Richter et al. 2016) dataset to *real-world* dataset, Cityscapes (Cordts et al. 2016).

Related Works

In this section, we briefly summarize the methods related to the key ideas of the proposed DAST, including adversarial learning, discriminator confidence and self-training.

Adversarial Learning. Being the most explored approach in the field of unsupervised domain adaptation for semantic segmentation, adversarial learning mostly aligns information either on feature level or pixel level to reduce the domain shift¹.

For the feature-level alignment, Hoffman et al. (Hoffman et al. 2016) first apply adversarial learning to align the feature distributions between different domains to train a semantic segmentation model for the real-world images. Later on, Tsai et al. (Tsai et al. 2018) find that to align the output space distribution is more effective than to align the distribution of the intermediate feature space. Luo et al. (Luo et al. 2019b) utilize co-training to keep semantic consistency from multiple views of the features, which encourages the category-level alignment of different domains. Vu et al. (Vu et al. 2019) use adversarial learning to match the entropy of output predictions in source and target domains, which provides an alternative way of output space alignment. Tsai et

al. (Tsai et al. 2019) construct different modes of images through patch-level clustering, and obtain a discriminator that pays more attention to high-level patterns, so as to optimize the domain alignment.

The pixel-level alignment is also known as image-to-image translation or style transfer. Benefiting from generative adversarial network (GAN), pixel-level alignment translates source images to the target domain, or vice versa, to facilitate learning a segmentation model across different domains, such as classic CycleGAN (Zhu et al. 2017). Recently, researchers have added the flavor of feature-level alignment to the pixel-level alignment, in order to achieve more accurate segmentation. Hoffman et al. (Hoffman et al. 2018) and Chen et al. (Chen et al. 2019) align the intermediate features to optimize image-to-image translation. Li et al. (Li, Yuan, and Vasconcelos 2019) construct a two-way learning process, which iteratively improves the segmentation and image-translation. Chang et al. (Chang et al. 2019) use pixel-level adversarial learning to disentangle the image features and train the segmentation model with content-only information. Choi et al. (Choi, Kim, and Kim 2019) adopt AdaIN (Huang and Belongie 2017) to embed the image information of unlabeled target images into the training process, making the model suitable to segment the target images.

Discriminator Confidence. Discriminator confidence is the output of the discriminator network in a fully convolutional manner. Hung et al. (Hung et al. 2019) use discriminator confidence to select the regions with small differences between the segmentation prediction and the label to form a pseudo label for the model training. (Kurmi, Kumar, and Namboodiri 2019; Wang et al. 2019) explore it in the intermediate layer as an attention map for domain adaptation in image classification. Inspired by these works, we implement the attention map in the proposed DA module using the dis-

¹We take output-level alignment as a special case of feature-level alignment

criminator confidence scores from the intermediate features.

Self-Training. Self-training or self-distillation has shown impressive results in recent years (Dong et al. 2019; Zhai et al. 2019; Zhang et al. 2019). In the field of UDA for semantic segmentation, CNN based self-training methods mainly fine-tune a trained segmentation model using the target images and the pseudo labels, which implicitly forces the model to extract the domain-invariant features. Zou et al. (Zou et al. 2018) perform self-training by adjusting class weights to generate more accurate pseudo labels to train the segmentation model. French et al. (French, Mackiewicz, and Fisher 2018) adopt the mean-teacher framework, which introduces a consistency regularization to realize domain adaptation between the mean-teacher (target domain) and the student (source domain). In the proposed method, we find that self-training could be combined with the DA module to further improve the decision boundary of the segmentation model for unlabeled target images.

Method

In this section, we first provide an overview of our method. Then, we describe the overall objective function. Finally, we discuss the proposed discriminator attention in more detail.

Method Overview

In this work, we focus on the problem of unsupervised domain adaptation for semantic segmentation, where we have the access to the labeled source dataset $\{\mathbf{x}_s, \mathbf{y}_s\}$ and unlabeled target dataset $\{\mathbf{x}_t\}$. As shown in Figure 2, the overall network architecture is mainly composed of a segmentation network (segmentor S), and two discriminator networks (discoverer D and corrector C). The network backbone of the segmentor S can be any fully-convolutional network for semantic segmentation. For better description and discussion, S is divided into a feature extractor E and a label predictor P , where $S = E \circ P$. Discriminators (D and C) are CNN-based classifiers with a fully convolutional output, which could provide confidence scores for all output locations to evaluate the local alignment of the different domains.

In the source flow, E extracts a feature map \mathbf{f}_s from a source domain image \mathbf{x}_s , where $\mathbf{f}_s = E(\mathbf{x}_s)$. The predictor P then takes \mathbf{f}_s as an input to form a pixel-level semantic segmentation \mathbf{p}_s , where $\mathbf{p}_s = P(\mathbf{f}_s)$, which will be used to calculate a segmentation loss \mathcal{L}_{seg} under the supervision of the source label \mathbf{y}_s . On the other hand, \mathbf{f}_s and \mathbf{p}_s will be input into the discoverer D and the corrector C for feature-level and output-level adversarial learning, respectively.

In the target flow, for a given image \mathbf{x}_t , E outputs a feature map \mathbf{f}_t which is first input to the discoverer D . By optimizing the adversarial loss \mathcal{L}_{adv}^D , D aligns the feature distribution of \mathbf{f}_t and \mathbf{f}_s and provides a confidence score of alignment for each location in \mathbf{f}_t to form an attention map α , where $\alpha = |D(\mathbf{f}_t)|$. α reweights \mathbf{f}_t into a new feature map $\hat{\mathbf{f}}_t = \alpha(\mathbf{f}_t)$, which is input to P to yield the pixel-level prediction $\hat{\mathbf{p}}_t$ with more focus on poorly-aligned regions, where $\hat{\mathbf{p}}_t = P(\hat{\mathbf{f}}_t)$. The corrector C is then introduced to perform

an adversarial learning between \mathbf{p}_t and \mathbf{p}_s . To further enhance the adaptation of poorly-aligned regions, we reweight the adversarial loss \mathcal{L}_{adv}^C with the attention map α .

In addition, we apply a self-training strategy to improve the decision boundary of the segmentation model. Similar to (Li, Yuan, and Vasconcelos 2019), we introduce a super parameter q of the pixel portion. We generate the pseudo label $\hat{\mathbf{p}}_t$ using the top q of pixels in \mathbf{p}_t with higher probability values and mask out other pixels which will not participate in gradient back-propagation.

The training process of the proposed method is summarized in **Algorithm 1**. In practice, we set the initial q to 50% and the maximum iteration K of self-training to 3 (the performance converges).

Algorithm 1 Training process of proposed method

Input:

The source domain sample, (x_s, y_s)
The target domain sample, x_t
The initial network, $S(\text{segmenter}), D(\text{discoverer}), C(\text{corrector})$

Output:

The trained network, $S'_{K+1}, D'_{K+1}, C'_{K+1}$
1: train $S'_0 \leftarrow S, D'_0 \leftarrow D, C'_0 \leftarrow C$ with loss \mathcal{L}_{seg} and \mathcal{L}_{adv}
2: **for** $k = 0$ to K **do**
3: input x_t into S'_k and generate pseudo label p_t with a fixed portion q_k
4: train $S'_{k+1} \leftarrow S'_k, D'_{k+1} \leftarrow D'_k, C'_{k+1} \leftarrow C'_k$ with loss $\mathcal{L}_{seg}, \mathcal{L}_{adv}$ and \mathcal{L}_{p-seg}
5: **end for**
6: return $S'_{K+1}, D'_{K+1}, C'_{K+1}$

Objective Functions

The overall loss function mainly consists of four loss terms:

$$\mathcal{L}_{overall}(E, P, D, C) = L_{seg}(E, P) + \lambda_d \mathcal{L}_{adv}^D(E, D) + \lambda_c \mathcal{L}_{adv}^C(E, P, C) + \mu \mathcal{L}_{p-seg}(E, P).$$

The first term $\mathcal{L}_{seg}(E, P)$ guides the segmenter S ($S = P \circ E$) to perform a dense prediction of the segmentation in the source domain,

$$\begin{aligned} \min_{E, P} \mathcal{L}_{seg}(E, P) \\ = \min_{E, P} \mathbb{E}_{\mathbf{x}_s, \mathbf{y}_s \sim p(\mathbf{X}_s, \mathbf{Y}_s)} (\ell_{ce}(P \circ E(\mathbf{x}_s), \mathbf{y}_s)) \end{aligned}$$

where $\ell_{ce}(\cdot, \cdot)$ indicates the multi-class cross entropy loss used in this work.

The second and third terms are the adversarial losses of the discoverer D and the corrector C , respectively. Following LSGAN (Mao et al. 2017), we use the least square loss to replace the sigmoid cross entropy in the vanilla GAN, because the sigmoid-based loss usually stops updating when the discriminator reaches the optimum (Hong et al. 2019). $\mathcal{L}_{adv}^D(E, D)$ and $\mathcal{L}_{adv}^C(E, P, C)$ correspond to the two-stage adversarial learning in our discriminator attention module.

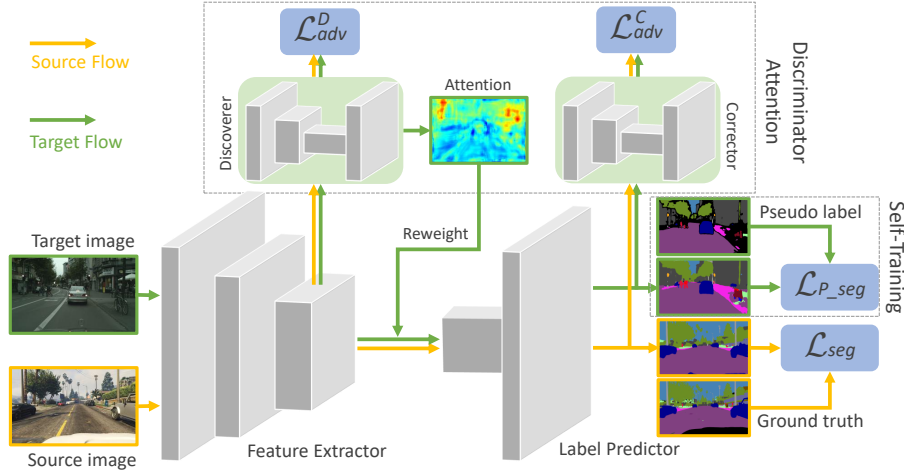


Figure 2: Overview of our proposed method. The randomly selected images from the source and target domain are used to train a cross-domain segmentation network by adversarial training. Two fully convolutional discriminator network named discoverer and corrector are used to obtain a better alignment. The discoverer could align the intermediate features of the segmentation network and form an attention map using the confidence of local alignment. The corrector can focus on the hard-adapted regions and optimize the overall adversarial learning based on the previous attention map. In addition, we utilize self-training to train an adaptive classifier boundary for the target domain. (Best view in color)

In the first stage, domain-invariant features extracted by E are expected to confuse the discoverer D , which aims to minimize the loss $\mathcal{L}_{adv}^D(E, D)$ by alternatively optimizing D and E ,

$$\min_D \mathcal{L}_{adv}^D(D) = \min_D \mathbb{E}_{\mathbf{x}_s \sim p(\mathbf{x}_s)} [(D(\mathbf{f}_s) - 0)^2] + \mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{x}_t)} [(D(\mathbf{f}_t) - 1)^2]$$

$$\min_E \mathcal{L}_{adv}^D(E) = \min_E \mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{x}_t)} [(D(E(\mathbf{x}_t)) - 0)^2]$$

After D is optimized, for a given target image \mathbf{x}_t , an attention map is generated to distinguish the easy-adapted and hard-adapted regions, $\alpha = |D(\mathbf{f}_t)|$.

In the second stage, we expect that $P \circ E$ outputs segmentation predictions that are able to confuse C .

$$\min_C \mathcal{L}_{adv}^C(C) = \min_C \mathbb{E}_{\mathbf{x}_s \sim p(\mathbf{x}_s)} [(C(\mathbf{p}_s) - 0)^2] + \mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{x}_t)} [(C(\mathbf{p}_t) - 1)^2]$$

$$\min_{E, P} \mathcal{L}_{adv}^C(E, P) = \min_{E, P} \mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{x}_t)} [(C(P \circ E(\mathbf{x}_t)) - 0)^2]$$

The fourth loss term is related to self-training strategy, which adaptively improves the decision boundary of the segmenter S ($S = P \circ E$) to fit the target distribution,

$$\min_{E, P} \mathcal{L}_{p_seg}(E, P) = \min_{E, P} \mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{x}_t)} (\ell_{ce}(\mathbf{p}_t, \hat{\mathbf{p}}_t)).$$

In the overall loss, $\lambda_d, \lambda_c, \mu$ are the hyper-parameters used to balance the relative importance of different terms. During training, we set $\lambda_d = 0.01, \lambda_c = 0.01, \mu = 1$.

Design of Attention Mechanism

For the target image feature \mathbf{f}_t , the confidence scores of the discoverer $\alpha = |D(\mathbf{f}_t)|$ show whether \mathbf{f}_t locally matches the distribution of \mathbf{f}_s . A low α_{ij} represents a well-aligned region in \mathbf{x}_t and a high α_{ij} represents a poorly-aligned region. Therefore, we use α as an attention map of \mathbf{f}_t to encourage the model to focus on matching features of those poorly-aligned regions. Moreover, to prevent the gradient explosion at the early stage of the experiment, we add a \tanh activation to α as a normalization layer. Finally, we expand $\tanh(\alpha)$ to fit the dimension of \mathbf{f}_t for the subsequent element-wise multiplication,

$$\alpha' = \text{expand}(\tanh(\alpha))$$

Since the magnitude of $\tanh(\alpha)$ is less than 1, its gradient may disappear in the late stage of the training process. We thus adopt the residual attention mechanism (Wang et al. 2017) to calculate the new feature map,

$$\mathbf{f}'_t = \mathbf{f}_t + \mathbf{f}_t \odot \alpha'$$

Experiments and Results

In this section, we will present our experiments and results. We first describe the benchmark datasets and experimental setups. Then, we report our main results and compare them with the state-of-the-art methods on the benchmark datasets.

Datasets

We evaluate the proposed DAST method on the challenging *synthetic-2-real* setups: SYNTHIA (Ros et al. 2016) and GTA5 (Richter et al. 2016) datasets are used as the source domain dataset and Cityscapes (Cordts et al. 2016) is used

Methods	Arch.	Mech.	road	sidewalk	building	wall	fence	pole	light	sign	vege.	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
Source only	V	-	60.7	13.7	56.9	12.9	20.1	19.0	15.4	6.5	77.7	16.2	56.8	40.0	3.3	63.6	15.3	9.5	0.0	8.1	0.1	26.1
CLAN	V	A	88.0	30.6	79.2	23.4	20.5	26.1	23.0	14.8	81.6	34.5	72.0	45.8	7.9	80.5	26.6	29.9	0.0	10.7	0.0	36.6
BDL	V	RSA	89.2	40.9	81.2	29.1	19.2	14.2	29.0	19.6	83.7	35.9	80.7	54.7	23.3	82.7	25.8	28.0	2.3	25.7	19.9	41.3
FDA-MBT	V	RSA	86.1	35.1	80.6	30.8	20.4	27.5	30.0	26.0	82.1	30.3	73.6	52.5	21.7	81.7	24.0	30.5	29.9	14.6	24.0	42.2
Baseline	V	A	87.3	29.8	78.6	21.1	18.2	22.5	21.5	11.0	79.7	29.6	71.3	46.8	6.5	80.1	23.0	26.9	0.0	10.6	0.3	35.0
Ours(DA)	V	A	89.3	40.4	79.2	34.9	22.8	23.1	24.0	16.8	79.9	28.7	67.9	45.1	17.8	82.1	25.7	31.9	4.1	19.7	3.1	38.8
Ours(DAST)	V	SA	90.5	49.2	81.9	34.0	27.0	26.5	26.6	21.5	83.0	37.3	76.3	52.0	23.1	83.5	29.9	42.0	12.1	19.8	25.8	44.3
Source only	R	-	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
CLAN	R	A	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
SIBAN	R	A	88.5	35.4	79.5	26.3	24.3	28.5	32.5	18.3	81.2	40.0	76.5	58.1	25.8	82.6	30.3	34.4	3.4	21.6	21.5	42.6
BDL	R	RSA	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
DPR	R	RSA	92.3	51.9	82.1	29.2	25.1	24.5	33.8	33.0	82.4	32.8	82.2	58.6	27.2	84.3	33.4	46.3	2.2	29.5	32.3	46.5
UIDA	R	SA	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0.0	30.2	35.8	46.3
DTST	R	RSA	90.6	44.7	84.8	34.3	28.7	31.6	35.0	37.6	84.7	43.3	85.3	57.0	31.5	83.8	42.6	48.5	1.9	30.4	39.0	49.2
FDA-MBT	R	RSA	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
Baseline	R	A	86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
Ours(DA)	R	A	92.3	54.2	81.9	27.3	25.3	33.4	39.1	23.2	84.0	34.2	71.1	58.7	29.7	85.2	28.1	34.7	4.8	25.6	19.6	44.8
Ours(DAST)	R	SA	92.2	49.0	84.3	36.5	28.9	33.9	38.8	28.4	84.9	41.6	83.2	60.0	28.7	87.2	45.0	45.3	7.4	33.8	32.8	49.6

Table 1: Experimental results for GTA5 \rightarrow Cityscapes. "Source only" denotes the model only trained on source data without adaptation. The architecture "V" and "R" represent the VGG-16 and ResNet-101 backbones, respectively. The mechanism "R", "S", and "A" means image-to-image translation, self-training, and adversarial training, respectively. Our baseline model is AdaptSegNet (Tsai et al. 2018). Other previous state-of-the-art methods include CLAN (Luo et al. 2019b), SIBAN (Luo et al. 2019a), BDL (Li, Yuan, and Vasconcelos 2019), DPR (Tsai et al. 2019), UIDA (Pan et al. 2020), DTST (Wang et al. 2020), FDA-MBT (Yang and Soatto 2020).

as the target domain dataset. The details of these datasets are described as follows: **1) GTA5 dataset.** GTA5 dataset consists of 24966 labeled urban scene images with a resolution of 1914×1052 collected from the video game Grand Theft Auto V. **2) SYNTHIA dataset.** SYNTHIA is another more challenging synthetic image dataset. For this dataset, we only use the SYNTHIA-RAND-CITYSCAPES subset, which has 16 common categories with Cityscapes. It contains 9400 labeled urban scene images with a resolution of 1280×760 . **3) Cityscapes dataset.** Cityscapes consists of 2975 real-world images in the training set and 500 in the validation set with a resolution of 2048×1024 . In all the experiments of this work, we use the 2975 images in the Cityscapes training set as the unlabeled target images and test the model with mean Intersection-over-Union (mIoU) on the 500 validation images. We only used the labels of the Cityscapes images to evaluate the segmentation performance and not for the training process.

Network Architecture

Inspired by (Choi, Kim, and Kim 2019; Tsai et al. 2018), we adopt the DeepLab (Chen et al. 2017) framework with VGG-16 (Simonyan and Zisserman 2014) and ResNet-101 (He et al. 2016) backbone as our segmentation network. The initial weight is pretrained on ImageNet (Deng et al. 2009). After the last convolutional layer, the Atrous Spatial Pyramid Pooling (ASPP) module is applied with the sampling rates of $\{6, 12, 18, 24\}$. Finally, we utilize an upsampling layer to rescale the final segmentation output to match the dimension of the input image.

The discriminators (the discoverer D and the corrector C) are fully convolutional networks that retain the spatial information. Furthermore, D consists of 4 convolutional layers with channel numbers of $\{256, 128, 64, 1\}$ and the values of kernel size, padding size, and stride are 3, 1, and 1, respectively. C aligns the semantic predictions of different domains. Following (Tsai et al. 2018), it consists of 5 convolutional layers with kernel size, padding size, and stride of 4, 1, 2, respectively, and its channel numbers are $\{64, 128, 256, 512, 1\}$. Instead of regular ReLU, C uses Leaky ReLU as the activation with a fixed negative slope of 0.2.

Implementation Details

The model is implemented using the PyTorch toolbox and runs on a single Titan V GPU with 12 GB memory². We introduce different settings for the segmentation network and the discriminators. 1) For the segmentation network, we use the stochastic gradient descent (SGD) algorithm as the optimizer. The initial learning rate is set as 2.5×10^{-4} , momentum is 0.9 and weight decay is 5×10^{-4} . 2) For the discriminators, we use the Adam algorithm as the optimizer. The initial learning rate is set as 10^{-4} and $\beta_1 = 0.9, \beta_2 = 0.99$. We also adopt the same polynomial decay with a power of 0.9 to update learning rate as mentioned in (Tsai et al. 2018).

We train the discriminator attention module only for 150k iterations and after that, we add self-training to fine-tune the model with pseudo labels for an additional 20k in several

²Code: https://github.com/yufei1900/DAST_segmentation

Methods	Arch.	Mech.	road	sidewalk	building	wall	fence	pole	light	sign	vege.	sky	person	rider	car	bus	mbike	bike	mIoU	mIoU*
Source Only	V	-	6.4	16.1	47.5	6.2	0.2	18.9	0.3	8.1	68.6	75.2	46.6	7.1	57.2	12.9	2.9	7.1	23.8	27.4
CLAN	V	A	80.4	30.7	74.7	-	-	-	1.4	8.0	77.1	79.0	46.5	8.9	73.8	18.2	2.2	9.9	-	39.3
BDL	V	RSA	72.0	30.3	74.5	0.1	0.3	24.6	10.2	25.2	80.5	80.0	54.7	23.2	72.7	24.0	7.5	44.9	39.0	46.1
FDA-MBT	V	RSA	84.2	35.1	78.0	6.1	0.44	27.0	8.5	22.1	77.2	79.6	55.5	19.9	74.8	24.9	14.3	40.7	40.5	47.3
Baseline	V	A	78.9	29.2	75.5	-	-	-	0.1	4.8	72.6	76.7	43.4	8.8	71.1	16.0	3.6	8.4	-	37.6
Ours(DA)	V	A	82.5	33.3	76.8	2.1	0.9	20.2	1.9	8.1	76.4	77.9	42.9	13.6	69.9	17.5	7.7	15.1	34.2	40.3
Ours(DAST)	V	SA	86.1	35.7	79.9	5.2	0.8	23.1	0.0	6.9	80.9	82.5	50.6	19.8	79.7	21.9	21.3	38.8	39.6	46.5
Source Only	R	-	55.6	23.8	74.6	9.2	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5	38.6
CLAN	R	A	81.3	37.0	80.1	-	-	-	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	22.6	30.7	-	47.8
SIBAN	R	A	82.5	24.0	79.4	-	-	-	16.5	12.7	79.2	82.8	58.3	18.0	79.3	25.3	17.6	25.9	-	46.3
BDL	R	RSA	86.0	46.7	80.3	-	-	-	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	-	51.4
DPR	R	RSA	82.4	38.0	78.6	8.7	0.6	26.0	3.9	11.1	75.5	84.6	53.5	21.6	71.4	32.6	19.3	31.7	40.0	46.5
UIDA	R	SA	84.3	37.7	79.5	5.3	0.4	24.9	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	41.7	48.9
DTST	R	RSA	83.0	44.0	80.3	-	-	-	17.1	15.8	80.5	81.8	59.9	33.1	70.2	37.3	28.5	45.8	-	52.1
FDA-MBT	R	RSA	79.3	35.0	73.2	-	-	-	19.9	24.0	61.7	82.6	61.4	31.1	83.9	40.8	38.4	51.1	-	52.5
Baseline	R	A	79.2	37.2	78.8	10.5	0.3	25.1	9.9	10.5	78.2	80.5	53.5	19.6	67.0	29.5	21.6	31.3	39.5	45.9
Ours(DA)	R	A	83.2	40.6	80.3	10.2	0.3	27.5	7.9	11.2	79.4	84.6	54.1	20.9	73.4	33.2	18.1	27.3	40.8	47.2
Ours(DAST)	R	SA	87.1	44.5	82.3	10.7	0.8	29.9	13.9	13.1	81.6	86.0	60.3	25.1	83.1	40.1	24.4	40.5	45.2	52.5

Table 2: Experimental results for SYNTHIA \rightarrow Cityscapes. The mIoU* denotes the mean IoU of 13 common classes. "Source only" denotes the model only trained on source data without adaptation. The architecture "V" and "R" represent the VGG-16 and ResNet-101 backbones, respectively. The mechanism "R", "S", and "A" means image-to-image translation, self-training, and adversarial training, respectively. Our baseline model is AdaptSegNet (Tsai et al. 2018).

Methods	mIoU(VGG16)	mIoU(Res101)
Baseline	35.0	41.4
+MSE	35.9	43.4
+MSE+2D	37.2	43.9
+MSE+2D+Atten.	38.8	44.8

Table 3: The ablation study results of discriminator attention adapted from GTA5 to Cityscapes. "MSE" denotes the mean square error loss function. "2D" denotes the feature distribution is aligned by 2 discriminators without attention.

rounds until the performance converges. During training, the images from GTA5 are resized to 1280×720 resolution, the images from SYNTHIA are resized to 1280×760 resolution, and the images from Cityscapes are resized to 1024×512 . During validation, we upsample the segmentation predictions to 2048×1024 to calculate evaluation metrics. We train our model without any data augmentation steps.

Experimental Results

The previous methods of UDA for semantic segmentation can be roughly divided into image-to-image translation (R), self-training (S), adversarial training (A), and their combinations. The experimental results compared with these methods are shown in Table 1 (GTA5 to Cityscapes) and Table 2 (SYNTHIA to Cityscapes). The proposed DAST achieves the superior performance on two benchmark datasets.

GTA5 \rightarrow Cityscapes. As shown in Table 1, our proposed method with DA only significantly outperforms the baseline by 3.8% and 3.4% in the mean IoU for two architectures and

exceed all the other models using a single strategy. Although CLAN also aims to tackle the equally global alignment problem, our method is more effective and outperforms them by 2.2% and 1.6% for two architectures. Compared with the composite methods, our DAST also achieves the state-of-the-art performance.

SYNTHIA \rightarrow Cityscapes. In the SYNTHIA dataset, the spatial layout or local context differs substantially from that in Cityscapes dataset. Despite the large domain shift between SYNTHIA and Cityscapes, our proposed method with DA only outperforms other adversarial learning-based methods and brings 2.7% and 1.3% improvement compared to the baseline over the 13 common classes for two architectures. Among the composite methods, our DAST also achieves the state-of-the-art performance.

Analysis

Ablation Studies

To verify the effectiveness of each part in discriminator attention, we perform 3 ablation tests. As shown in Table 3, the MSE loss function brings 0.9% and 2.0% improvement compared to the baseline. The combination of 2 discriminators brings another 1.3% and 0.5% improvement. The introduction of attention contributes 1.6% and 0.9% mIoU gain. Our final model exceeds the baseline for most of the categories in terms of segmentation accuracy.

Hyper-parameters Analysis

As shown in Table 4, when $q = 50\%$, the model achieves 49.6 of mIoU as the best performance on Cityscapes validation set.

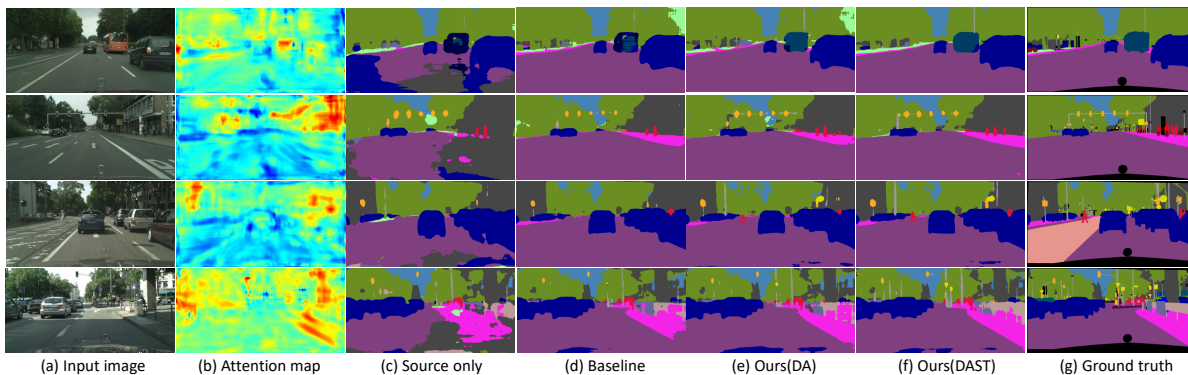


Figure 3: Visual results of segmentation map and attention map on VGG-16 backbone. (a) Input images. (b) Attention map generated by our discriminator attention module. (c) Segmentation map predicted by source only model. (d) Segmentation map predicted by baseline model. (e) Segmentation map predicted by our DA model. (f) Segmentation map predicted by our DAST model. (g) Ground truth.

q	40%	50%	60%
mIoU	49.0	49.6	48.7

Table 4: The results of hyper-parameter q from 40% to 60% adapted from GTA5 to Cityscapes on ResNet-101 backbone.

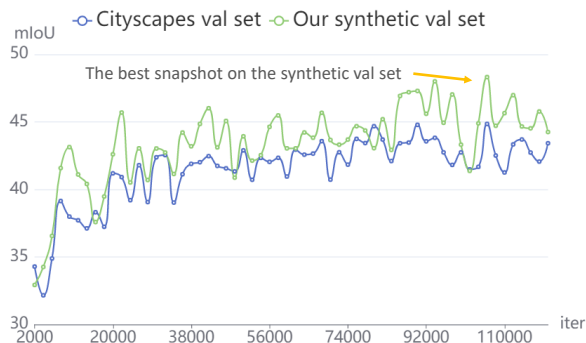


Figure 4: mIoU comparison on the Cityscapes val set and our synthetic val set by adapting from GTA5 to Cityscapes.

When to Start the Self-Training

Most previous studies suffer from biased models and overly-optimistic estimates, because they often select the best result from all evaluations of the intermediate snapshots on the Cityscapes val set. To address the overoptimistic issue, we establish a synthetic val set which consists of 500 randomly selected GTA5 images transferred to Cityscapes style by the method of (Yang and Soatto 2020). The performance of the model on the synthetic val set can guide the choice of the best snapshot and when to start the self-training. We use the mean of the two styles ($\beta = 0.5$ and $\beta = 0.9$) as the final results of the synthetic val set. As shown in Figure 4, the synthetic val set can approximately fit the Cityscapes val set.

Visualization

We use four qualitative examples in Figure 3 to illustrate that the attention map serves as a good indicator of hard-adapted regions. For example, the hard-adapted regions—the “bus”, “sidewalk” and “traffic light” that are wrongly predicted in the source-only models (col (c))—are assigned with higher confidence values in the attention map. Whereas the easy-adapted regions—the “road”, “sky” and “car” regions that are correctly predicted in the source-only models—are characterized with lower confidence values in the attention map. This consistency proves that the attention map successfully distinguishes hard-adapted regions and easy-adapted regions.

Figure 3 also reveals that the attention mechanism can improve the segmentation results. Cols (d) and (e) demonstrate the segmentation map predicted by the baseline model and our model that introduces the attention mechanism, respectively. It is obvious that our results improved by the attention mechanism are visually closer to the ground truth (col (g)) than the baseline’s. Col (f) demonstrates that self-training could be combined with the DA module to further improve the performance.

Conclusions

In this paper, we have proposed a novel method that combines the discriminator attention and the self-training to realize the unsupervised domain adaptation for semantic segmentation. The discriminator attention module includes two stages of adversarial learning, which utilize the attention map to attach higher weights to hard-adapted regions and performs the feature-level and the output-level alignments between different domains. The self-training module dynamically generates pseudo labels to adapt the decision boundary of the segmentation network to fit the distribution of unlabeled target images. The experimental results and the qualitative examples prove that our method outperforms the previous state-of-the-art methods on the benchmark datasets.

Acknowledgments

This work was supported in part by National Key R&D Program of China (No. 2018YFC0910700); National Natural Science Foundation of China (NSFC) grant No. 11831002, 81801778, 71704024; Beijing Natural Science Foundation (No. 180001) and Beijing Academy of Artificial Intelligence (BAAI).

References

- Chang, W.-L.; Wang, H.-P.; Peng, W.-H.; and Chiu, W.-C. 2019. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1900–1909.
- Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40(4): 834–848.
- Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.
- Chen, Y.-C.; Lin, Y.-Y.; Yang, M.-H.; and Huang, J.-B. 2019. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1791–1800.
- Choi, J.; Kim, T.; and Kim, C. 2019. Self-Ensembling with GAN-based Data Augmentation for Domain Adaptation in Semantic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 6830–6840.
- Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- Dong, B.; Hou, J.; Lu, Y.; and Zhang, Z. 2019. Distillation \approx Early Stopping? Harvesting Dark Knowledge Utilizing Anisotropic Information Retrieval For Overparameterized Neural Network. *arXiv preprint arXiv:1910.01255*.
- French, G.; Mackiewicz, M.; and Fisher, M. 2018. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 6.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *International Conference on Machine Learning*, 1989–1998.
- Hoffman, J.; Wang, D.; Yu, F.; and Darrell, T. 2016. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*.
- Hong, Y.; Hwang, U.; Yoo, J.; and Yoon, S. 2019. How generative adversarial networks and their variants work: An overview. *ACM Computing Surveys (CSUR)* 52(1): 1–43.
- Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.
- Hung, W. C.; Tsai, Y. H.; Liou, Y. T.; Lin, Y. Y.; and Yang, M. H. 2019. Adversarial learning for semi-supervised semantic segmentation. In *29th British Machine Vision Conference, BMVC 2018*.
- Kurmi, V. K.; Kumar, S.; and Namboodiri, V. P. 2019. Attending to discriminative certainty for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 491–500.
- Li, Y.; Yuan, L.; and Vasconcelos, N. 2019. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6936–6945.
- Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.
- Luo, Y.; Liu, P.; Guan, T.; Yu, J.; and Yang, Y. 2019a. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 6778–6787.
- Luo, Y.; Zheng, L.; Guan, T.; Yu, J.; and Yang, Y. 2019b. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2507–2516.
- Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802.
- Pan, F.; Shin, I.; Rameau, F.; Lee, S.; and Kweon, I. S. 2020. Unsupervised Intra-domain Adaptation for Semantic Segmentation through Self-Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3764–3773.
- Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *European conference on computer vision*, 102–118. Springer.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

- Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3234–3243.
- Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Tsai, Y.-H.; Hung, W.-C.; Schuler, S.; Sohn, K.; Yang, M.-H.; and Chandraker, M. 2018. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7472–7481.
- Tsai, Y.-H.; Sohn, K.; Schuler, S.; and Chandraker, M. 2019. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE International Conference on Computer Vision*, 1456–1465.
- Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2517–2526.
- Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.
- Wang, X.; Li, L.; Ye, W.; Long, M.; and Wang, J. 2019. Transferable attention for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5345–5352.
- Wang, Z.; Yu, M.; Wei, Y.; Feris, R.; Xiong, J.; Hwu, W.-m.; Huang, T. S.; and Shi, H. 2020. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12635–12644.
- Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; and Huang, T. S. 2018. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7268–7277.
- Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4085–4095.
- Zhai, X.; Oliver, A.; Kolesnikov, A.; and Beyer, L. 2019. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, 1476–1485.
- Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; and Ma, K. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, 3713–3722.
- Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.
- Zou, Y.; Yu, Z.; Vijaya Kumar, B.; and Wang, J. 2018. Un-supervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, 289–305.