AUTOMATIC ENDOSCOPY CLASSIFICATION BY FUSING DEPTH ESTIMATIONS AND IMAGE INFORMATION

Diego Bravo[a*], *Josué Ruano*[a*], *María Jaramillo*[a], *Sebastian Medina*[b], *Martín Gómez*[c], *Fabio A. González*[b], *and Eduardo Romero*[a 1]

[a] Computer Imaging and Medical Applications Laboratory (CIM@LAB)
[b] Machine Learning, Perception and Discovery Lab (MindLab)
[c] Hospital Universitario Nacional de Colombia, Unidad de Gastroenterología, Bogotá, Colombia
Universidad Nacional de Colombia

ABSTRACT

Gastric cancer ranks as the fifth leading cause of cancer mortality worldwide. The quality of upper gastrointestinal endoscopy is crucial towards early identification of premalignant conditions and relies on the endoscopist's skill and thorough examination of stomach landmarks. Unfortunately, it has been observed that existing cancerous lesions may go undetected during examination. To standardize the quality of this procedure, meticulous protocols have been proposed. To support this process, we focused on developing a model to identify the anatomical locations in esophagogastroduodenoscopy images. This study advances endoscopic image classification by incorporating depth map estimation, essential for measuring distances to specific landmarks. This method, analyzing 2,054 images from 96 patients across 13 gastric regions using the ConvNeXT architecture with information fusion techniques, achieved an 87% F1 macro score. This approach suggests that depth map integration can improve stomach region classification, boosting prediction accuracy and potentially reducing missed gastric lesions.

Index Terms— Endoscopy, Sub-anatomical region, Classification, Fusion Information, Depth map.

1. INTRODUCTION

Gastric cancer remains a significant global health concern, with an estimated 968,784 new cases diagnosed in 2022, making it the fifth leading cause of cancer mortality worldwide, with 660,175 deaths recorded in the same period [1]. Early detection is key to improving treatment outcomes for gastric cancer or precursor lesions, and adopting effective screening strategies, including minimally invasive screening and endoscopy [2]. Recently, the advances achieved in endoscopic technology, particularly in esophagogastroduodenoscopy (EGD) or upper gastrointestinal endoscopy (UGIE), have positioned it as the gold standard for diagnosis of upper gastrointestinal (UGI) diseases. Widely used as a screening test for early gastric cancer (EGC), EGD offers comprehensive visualization of the esophagus, stomach, and proximal duodenum. Nevertheless, endoscopy is a highly difficult procedure, with cognitive and technical factors complicating the risk of misdiagnosis, a fact documented since between 20%–25% of lesions are missed for EGC [3] while 11.3% of UGI cancers are not detected [4].

Several gastroenterology associations have developed protocols to enhance the efficiency and quality of these procedures [5][6] [7]. The Japan Gastroenterological Endoscopy Society developed a guideline for endoscopic diagnosis of EGC [8], mainly focused on the technical skills to examine the upper gastrointestinal tract. During the endoscopy, to avoid blind spots, K. Yao proposed a systematic screening protocol for the stomach (SSS) [9]. Overall, the SSS comprises a series of endoscopic photos of four quadrants of the gastric antrum, body, and middle–upper body. In practice, guidelines to map the entire stomach do exist but they are often partially followed, especially in developing countries [10].

Therefore, reliable methodologies to alert endoscopists about blind spots are needed. By incorporating computational methods into medical practice, gastroenterologists can establish a framework for ongoing quality monitoring and improvement, including thorough examination of anatomical regions and minimum recommended procedure times [11, 12]. These methods should serve as a supplementary reader, offering guidance about the current anatomical location within the patient's upper gastrointestinal system and providing quality indicators such as time spent exploring specific organs or sub-anatomical regions.

Researchers have explored two main approaches: single and multi-frame algorithms. In the literature, methods for anatomical landmark detection in UGIE have been introduced with deep learning techniques. These methods involve frame classification by fine-tuning of pre-trained image classification models: deep convolutional neural networks [13]. They include the VGG models, the Inception series, ResNet, and the contemporary convolutional neural network benchmark. For instance, Takiyama et al. [14] used a GoogleNet architecture to accurately recognize 4 anatomical locations (larynx, esophagus, stomach, and duodenum), as well as 3 subsequent sub-classifications specifically for stomach images. Wu et al. [15] trained a VGG-16 network to classify gastric locations into ten categories, and further refined the classification into 26 anatomical parts (22 for the stomach, 2 for the esophagus, and 2 for the duodenum). Additionally, Chang et al. [16] trained a ResNet architecture to classify EGD images into 8 anatomical locations, with an additional location for the pharynx and Bravo et al [17] performed a comparison of 23 networks to classify six gastric locations. Including temporal information, Li et al. [18] trained a combination of Inception-V3 and Long short-term memory (LSTM) models with adjacent frames to classify EGD images into 31 sites, ranging from the hypopharynx to the duodenum. Bravo et al. [19] compared the performance of a Gated Recurrent Unit (GRU) and a transformer encoder for classifying six gastric locations in video frame sequences and organ detection in EGD videos.

[1] Corresponding author, E-mail: edromero@unal.edu.co (Dr. E. Romero).
* Equal contribution

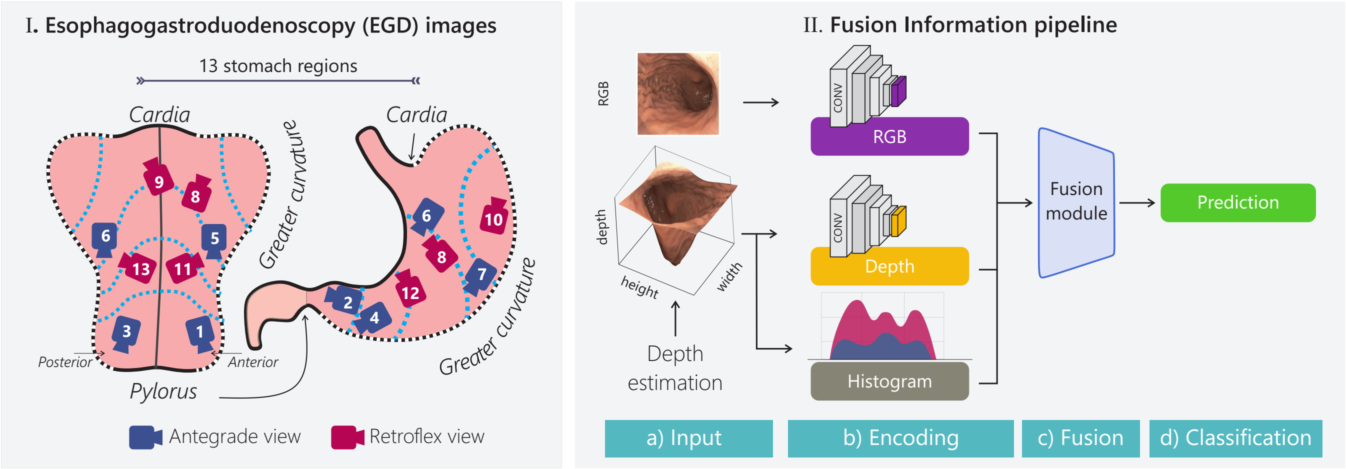

Fig. 1. Overview of the proposed methodology: (I) anatomical EGD sub-categories, (II) illustrates the workflow with : (a) initial data in RGB and the estimated depth maps (Section 2.1), (b-c) data encoding and integration (Section 2.2), and (d) sub-category classification of stomach regions (Section 3.2).

Despite sensitivity metrics generally showing positive results, specific areas, notably the upper middle body including the lesser curvature and posterior wall, continue to demonstrate suboptimal performance, with sensitivity rates of 78% and 83% respectively, as highlighted in [18]. According to Hosokawa's studies [20], [21] have identified a significant risk of missed EGC in these specific zones compared to other areas. Improving performance could involve integrating information about the distance between the endoscope and anatomical landmarks. This detail is vital so that experts can not only accurately navigate specific structures, but also carry out effective photographic documentation, ensuring that they are located at a distance that allows clear observation of the structure.

This paper significantly contributes by integrating depth map estimations with the automatic identification of gastric landmarks. This integration significantly improves the detection accuracy in vital body areas, especially in the lesser curvature and posterior wall. Our study offers more reliable predictions in these critical areas.

2. METHODOLOGY

This work explores the potential of combining standard color images (RGB) and depth maps for identifying crucial stomach regions during endoscopy. Two fusion approaches, early and late, are employed to integrate information from these static images. As depicted in Figure 1-I, the stomach is segmented into 13 distinct areas based on their anatomical location and orientation. The proposed methodology, outlined in Figure 1 (II), utilizes an independent Convolutional Neural Network (CNN) trained on both RGB and depth maps through transfer learning (step b). Two fusion strategies are then employed: early fusion, which combines RGB and depth data in the input layers of the CNN (step c), and late fusion, which combines the extracted features from separate models trained on each data type (step b). Finally, the combined information is used in step (d) to classify the sample into one of the 13 stomach region classes through a dedicated classification layer.

2.1. Depth estimation

Endoscopic images capture the inner surface of the stomach, where the shape, the 3D structure, is a function of the shading variations concerning the depth (distance) and orientation of the camera-light source. Light intensity diminishes with depth according to an inverse square law, where light intensity (I) is proportional to $I \propto \frac{1}{r^2}$ (r being the radial depth from the light source). Then, a depth map in endoscopy represents the distance between the camera-light source and the stomach wall, a map that may be estimated using a technique known as shape-from-shading. However, estimating depth from shading is challenging because the projection of a 2D image to a 3D scene leads to multiple possible solutions. To constrain the solution a set of assumptions were made: (a) the camera behaves like a pinhole with a constant focal length, (b) the light source position is nearly identical to the camera's center, and (c) the tissue surface (mucosa) is assumed to be Lambertian reflectance. This study used a Shape-from-Shading Network (SfSNet) to estimate depth maps from single RGB images via pixel-level regression [22]. This is a convolutional neural network with an encoder-decoder layout employing: EfficientNetB0 as backbone, long skip connections, and a custom loss function for gastrointestinal surfaces. The loss function balanced depth map reconstruction integrating three components: L_z for pixel-wise depth, L_e for edge detail (gradients), and L_c for curvature (Hessian derivatives for curved edge as gastric folds):

$$\mathbf{L}(d,\hat{d}) = w_1 L_z(d,\hat{d}) + w_2 L_e(\nabla(d),\nabla(\hat{d})) + w_3 L_c(H(d),H(\hat{d}))$$

A public collection of synthetic endoscopy videos with depth maps was used to train and test the SfSNet [23]. 80% was set for training and 20% for testing. During the training, five hyperparameters were tuned in 40 trials. The network achieved a threshold accuracy of 99.73% (decision threshold set in the ratio between the ground-truth and estimated depth maps, herein set to 1.25, a commonly used strict value in the literature) and a root-mean-square error of 3.66 mm. Finally, the trained network estimates depth maps from real endoscopy images.

2.2. Encoding and merging data representations

This methodology leverages a pre-trained ConvNeXT model developed by Liu et al. (2022) [24], selected for its performance in prior benchmarks related to anatomical endoscopy classification [17]. The approach is structured into two main stages. Initially, two independent models are trained, each specializing in a different data modality: one model is dedicated to processing RGB (color) images, while

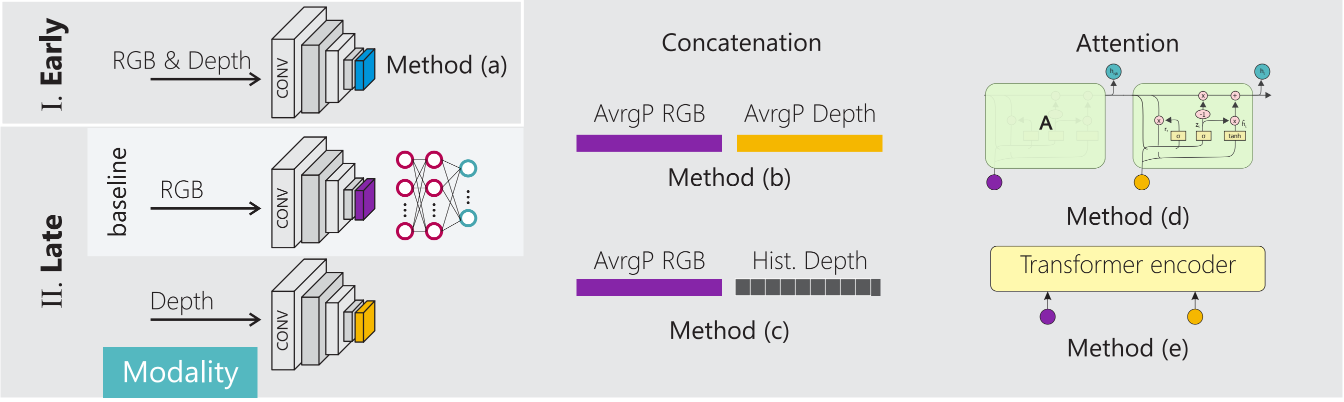

Fig. 2. Overview of information fusion techniques: Early fusion integrates RGB and depth data at the input stage, Late fusion merges the model embeddings for decision-making

the other specializes in analyzing depth maps, as illustrated in Figure 2-Modality. The subsequent phase involves a fusion module, wherein two separate strategies for merging information are examined. The early fusion strategy integrates RGB and depth data at the input level, as depicted in Figure 2-I. Alternatively, late fusion integrates the individual embeddings of the trained RGB and depth models, as shown in Figure 2-II. The RGB and depth models serve as feature extractors, generating 768 unique features from the final average pooling layer. Additionally, for depth maps (method c), the model employs 49 histograms derived from spatial grids, capturing distance relationships. The fusion of information is achieved through two techniques: the concatenation of embeddings and the implementation of an attention layer. Attention mechanisms, employ a Gated Recurrent Unit (GRU) and a Transformer encoder to learn a shared feature space for both color and depth modalities, enabling information integration. The obtained features through methods b-e are then fed into a neural network to carry out classification tasks.

2.3. Dataset

The dataset includes data from 96 patients who underwent EGD procedures. The age of the participants averaged 62 ± 15.5 years, with 50.6% being female. A total of 2,054 frames from recorded white light videos were manually categorized into 13 anatomical groups by an expert, following the systematic stomach screening protocol outlined by Yao et al. (2013) [9]. These frames were captured at a spatial resolution of 1,350×1,080 pixels. The study adhered to the principles of the Declaration of Helsinki, and ethical approval was granted by the Ethics Committee of the Hospital Universitario Nacional (approval number: CEI-2019-06-10).

2.4. ConvNeXT training

The model undergoes a two-phase training process, following a conventional transfer learning and fine-tuning approach. In the initial phase, known as "warmup", only the fully connected layers are unfrozen for training, while the feature extraction layers remain frozen. This strategic choice enables the model to rapidly acquire knowledge of how to transform features in a manner that optimally captures the patterns within the regional images. In the second phase, called "fine-tuning," a set of layers are unfrozen. This allows the entire model to adapt to the specific nuances of the regional image classification problem.

3. EVALUATION AND RESULTS

This section outlines the evaluation and findings of the suggested methods for integrating data representations, using the dataset detailed in section 2.3. The objective is to offer a detailed assessment of how effectively these methods fuse information from RGB and depth maps to categorize anatomical regions in endoscopic images.

3.1. Experimental Setup

The models were evaluated using a dataset partitioned into 70% for training and validation (58 cases with 1,185 frames for training and 9 cases with 198 frames for validation) and 30% for testing (comprising 29 cases with 671 frames) across all patients in the dataset. The CNN architectures received input data in the form of RGB images with dimensions $3 \times H \times W$, depth maps sized $1 \times H \times W$, and early fusion RGB-Depth images sized $4 \times H \times W$. Here, H and W denote the height and width, respectively, each standardized at 224 pixels according to the model's requirements.

3.1.1. CNN configuration

To ensure a straightforward comparison, the training of the CNNs was structured into two distinct phases: (a) Initially, there was a "warm-up" phase focused on training the classification layers. During this phase, these layers were trained for 10 epochs with a constant learning rate. (b) Following the warm-up phase, a fine-tuning phase commenced, targeting the final 20% of the feature layers. This fine-tuning was conducted over 100 epochs to optimize the model's performance. Details of CNNs and training configuration are presented below:

- **Pre-trainned weights:** ImageNet
- **Optimizer**: Adam
- **Loss function:** Weighted cross-entropy for class imbalance.
- **Dense layers:** 2 dense layers with dropout, and 13 output neurons for sub-anatomical categories (see Figure1-I)

The warm-up and fine-tuning phases were part of a hyperparameter optimization process conducted across 200 trials. During this optimization, we monitored the F-measure to identify the optimal batch size, initial learning rate, and learning rate schedule (gamma and step size). This meticulous approach enabled the selection of the most effective model as a feature extractor, guided by its performance in the validation phase.

3.1.2. Early and Late Fusion

In the early fusion strategy, depicted in Figure 2-I method(a), the ConvNeXT model was modified to handle the additional depth information. This adaptation involved expanding the first convolutional layer from handling three RGB channels to four channels (including depth). This aimed to enhance the model's ability to process multi-dimensional data while preserving its original architecture. To maintain consistency, the original layer's weights and biases were replicated and incorporated into the expanded layer. Furthermore, as detailed in Section 3.1.1, a comprehensive hyperparameter optimization process was conducted over 200 trials during training.

In late fusion, as illustrated in Figure 2-II, we explore various methods (b-e) for integrating color and depth information. For method (d), a GRU model is configured with a single layer containing 128 hidden units. The method (e) Transformer model is set up with one attention head and one transformer layer, processing input feature vectors linearly reduced to a dimensionality of 256. Furthermore, each method (b-e) incorporates a fully connected layer, structured by the specifications detailed in subsection 3.1.1.

3.2. Results

For each architecture, the model with the highest validation f1-score across the trials was challenged with the testing set. The results were provided in two scenarios using the classification metrics.

A. General results

To assess the models' performance, classification metrics and their corresponding 95% confidence intervals for accuracy were calculated, as detailed in Table 1. A normal approximation method was employed for the confidence interval calculations.

Method - %	Recall	Precision	F1_score	CI
RGB	85.53	84.73	84.86	85.99± 2.48
Depth	79.30	79.26	79.02	80.63± 2.85
Method (a)	81.31	81.84	81.18	81.97± 2.77
Method (b)	83.22	84.27	83.43	84.50± 2.60
Method (c)	85.54	85.38	85.34	86.59± 2.40
Method (d)	**88.09**	**86.97**	**87.42**	**88.38± 2.25**
Method (e)	84.36	84.77	84.40	85.39± 2.55

Table 1. Performance metrics for the different evaluated methods. Accuracy values are reported along with the confidence interval (CI) with a 95% confidence level. The table presents macro recall, precision, and F1 score.

The fusion method denoted as (d), which employs an attention layer with a Gated Recurrent Unit (GRU), yielded the most favorable results, achieving a macro-F1 score of 87.42%. This score surpasses the 84.86% score obtained from the trained RGB CNN by 2.56 percentage points. Moreover, from a clinical standpoint, method (d) stands out as the more reliable choice. This method capitalizes on a critical distance concept that directly correlates with photodocumentation. This correlation arises due to the necessity of calculating distances from the endoscope to specific anatomical landmarks. In certain scenarios, these landmarks encompass the same gastric tissue but require the consideration of different distances to observe it more comprehensively or in finer detail. Such an approach can offer invaluable support for estimations within the clinical context.

B. Stomach region (sub-anatomical) results

Table 2 illustrates the effectiveness of method (d) in 13 different anatomical areas, emphasizing a set of classification metrics for each anatomical landmark. Significant improvements are noted in areas L1-2, L4-11, and L13, corresponding to different stomach regions (see Figure1-I), when compared to a standard CNN trained on RGB images.

	Sensitivity		**Precision**		**F1 Score**	
Label	**RGB**	**M-d**	**RGB**	**M-d**	**RGB**	**M-d**
L1	82.76	82.76	75.00	**77.42**	78.69	**80.00**
L2	78.38	**81.08**	80.56	**81.08**	79.45	**81.08**
L3	**80.00**	76.67	82.76	**85.19**	**81.36**	80.70
L4	81.82	**84.85**	84.38	**84.85**	83.08	**84.85**
L5	89.66	**91.38**	80.00	**82.81**	84.55	**86.89**
L6	80.99	**84.30**	94.23	**94.44**	87.11	**89.08**
L7	77.59	**79.31**	70.31	**73.02**	73.77	**76.03**
L8	94.17	**95.00**	94.17	**96.61**	94.17	**95.80**
L9	89.66	**91.38**	88.14	**91.38**	88.89	**91.38**
L10	93.94	**96.97**	91.18	**91.43**	92.54	**94.12**
L11	80.65	**90.32**	**100.00**	93.33	89.29	**91.80**
L12	82.35	**91.18**	82.35	**91.18**	82.35	**91.18**
L13	100.00	100.00	78.38	**87.88**	87.88	**93.55**

Table 2. Comparative performance of sensitivity, precision, and F1 score for the RGB baseline model versus fusion information approach Method-d (M-d).

This advancement is particularly significant considering the high risk of missing premalignant lesions in L6, and L8-9 regions, underscoring the need for more thorough examinations of these specific zones.

4. CONCLUSIONS AND DISCUSSION

This research showcases advanced information fusion techniques, incorporating crucial aspects such as distance measurement in endoluminal scenes, vital for the photodocumentation protocol. This integration significantly elevates the accuracy and trustworthiness of estimations. Looking ahead, the future work aims to include multiscale fused information, broadening the scope and depth of the analysis. This improvement could enhance endoluminal analysis and photodocumentation, providing a more thorough and detailed approach in this field.

The next phase of our project involves the release of a public database containing patient data from the Hospital Universitario Nacional de Colombia. This initiative will also focus on enhancing depth estimation and implementing an automated system for encoding spatiotemporal information. We will address significant challenges, such as the prevalent noise in esophagogastroduodenoscopy procedures, and integrate quality indicators, thereby creating a more cohesive and effective framework for endoscopic analysis.

5. ACKNOWLEDGMENT

This work was supported by project with code 110192092354 and entitle "Program for the Early Detection of Premalignant Lesions and Gastric Cancer in urban, rural and dispersed areas in the Department of Nariño" of call No. 920 of 2022 of MinCiencias.

6. REFERENCES

[1] J Ferlay, M Ervik, F Lam, M Laversanne, M Colombet, L Mery, M Piñeros, A Znaor, I Soerjomataram, and F Bray, "Global cancer observatory: Cancer today," 2024.

[2] Adrian Săftoiu, Cesare Hassan, Miguel Areia, Manoop S Bhutani, Raf Bisschops, Erwan Bories, Irina M Cazacu, Evelien Dekker, Pierre H Deprez, Stephen P Pereira, et al., "Role of gastrointestinal endoscopy in the screening of digestive tract cancers in europe: European society of gastrointestinal endoscopy (esge) position statement," *Endoscopy*, vol. 52, no. 04, pp. 293–304, 2020.

[3] Mitsuru Kaise, "Advanced endoscopic imaging for early gastric cancer," *Best Practice & Research Clinical Gastroenterology*, vol. 29, no. 4, pp. 575–587, 2015.

[4] Shyam Menon and Nigel Trudgill, "How commonly is upper gastrointestinal cancer missed at endoscopy? a meta-analysis," *Endoscopy international open*, vol. 2, no. 02, pp. E46–E50, 2014.

[5] Glenn M Eisen, Todd H Baron, Jason A Dominitz, Douglas O Faigel, Jay L Goldstein, John F Johanson, J Shawn Mallery, Hareth M Raddawi, John J Vargo II, J Patrick Waring, et al., "Methods of granting hospital privileges to perform gastrointestinal endoscopy," *Gastrointestinal endoscopy*, vol. 55, no. 7, pp. 780–783, 2002.

[6] MJ Farthing, RP Walt, RN Allan, CH Swan, IT Gilmore, CN Mallinson, JR Bennett, CJ Hawkey, WR Burnham, AI Morris, et al., "A national training programme for gastroenterology and hepatology.," *Gut*, vol. 38, no. 3, pp. 459, 1996.

[7] Alistair D Beattie, Michel Greff, Vincent Lamy, and Christopher N Mallinson, "The european diploma of gastroenterology: progress towards harmonization of standards," *European journal of gastroenterology & hepatology*, vol. 8, no. 4, pp. 403–406, 1996.

[8] Kenshi Yao, Noriya Uedo, et al., "Guidelines for endoscopic diagnosis of early gastric cancer," *Digestive Endoscopy*, vol. 32, no. 5, pp. 663–698, 2020.

[9] Kenshi Yao, "The endoscopic diagnosis of early gastric cancer," *Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology*, vol. 26, no. 1, pp. 11, 2013.

[10] Gwang Ha Kim, Sung Jo Bang, et al., "Is screening and surveillance for early detection of gastric cancer needed in korean americans?," *The Korean Journal of Internal Medicine*, vol. 30, no. 6, pp. 747, 2015.

[11] Andrew M Veitch, Noriya Uedo, Kenshi Yao, and James E East, "Optimizing early upper gastrointestinal cancer detection at endoscopy," *Nature reviews Gastroenterology & hepatology*, vol. 12, no. 11, pp. 660–667, 2015.

[12] Wladyslaw Januszewicz and Michal F Kaminski, "Quality indicators in diagnostic upper gastrointestinal endoscopy," *Therapeutic Advances in Gastroenterology*, vol. 13, pp. 1756284820916693, 2020.

[13] Francesco Renna, Miguel Martins, Alexandre Neto, António Cunha, Diogo Libânio, Mário Dinis-Ribeiro, and Miguel Coimbra, "Artificial intelligence for upper gastrointestinal endoscopy: a roadmap from technology development to clinical practice," *Diagnostics*, vol. 12, no. 5, pp. 1278, 2022.

[14] Hirotoshi Takiyama, Tsuyoshi Ozawa, et al., "Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks," *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018.

[15] Lianlian Wu, Wei Zhou, et al., "A deep neural network improves endoscopic detection of early gastric cancer without blind spots," *Endoscopy*, vol. 51, no. 06, pp. 522–531, 2019.

[16] Yuan-Yen Chang, Pai-Chi Li, Ruey-Feng Chang, Chih-Da Yao, Yang-Yuan Chen, Wen-Yen Chang, and Hsu-Heng Yen, "Deep learning-based endoscopic anatomy classification: an accelerated approach for data preparation and model validation," *Surgical Endoscopy*, pp. 1–11, 2021.

[17] Diego Bravo, Josué Ruano, María Jaramillo, Daniel Gallego, Martín Gómez, Fabio A González, and Eduardo Romero, "Automatic classification of esophagogastroduodenoscopy subanatomical regions," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–5.

[18] Yan-Dong Li, Shu-Wen Zhu, et al., "Intelligent detection endoscopic assistant: An artificial intelligence-based system for monitoring blind spots during esophagogastroduodenoscopy in real-time," *Digestive and Liver Disease*, vol. 53, no. 2, pp. 216–223, 2021.

[19] Diego Bravo, Sebastian Medina, Josué Ruano, María Jaramillo, Daniel Gallego, Martín Gómez, Fabio A González, and Eduardo Romero, "Automated anatomical classification and quality assessment of endoscopy by temporal-spatial analysis," in *2023 IEEE 19TH International Symposium on Medical Information Processing and Analysis (SIPAIM)*. IEEE, 2023, pp. 1–5.

[20] O Hosokawa, S Tsuda, E Kidani, K Watanabe, Y Tanigawa, S Shirasaki, H Hayashi, and T Hinoshita, "Diagnosis of gastric cancer up to three years after negative upper gastrointestinal endoscopy," *Endoscopy*, vol. 30, no. 08, pp. 669–674, 1998.

[21] Osamu Hosokawa, Masakazu Hattori, Kenji Douden, Hiroyuki Hayashi, Kouji Ohta, and Yasuharu Kaizaki, "Difference in accuracy between gastroscopy and colonoscopy for detection of cancer.," *Hepato-gastroenterology*, vol. 54, no. 74, pp. 442–444, 2007.

[22] Josué Ruano, Martín Gómez, Eduardo Romero, and Antoine Manzanera, "Leveraging a realistic synthetic database to learn shape-from-shading for estimating the colon depth in colonoscopy images," *arXiv preprint arXiv:2311.05021*, 2023.

[23] Josué Ruano, Diego Bravo, María Jaramillo, Martín Gómez, Javier Pascau, Fabio A González, and Eduardo Romero, "Generating synthetic endoscopy videos following a systematic screening protocol," in *2023 IEEE 19TH International Symposium on Medical Information Processing and Analysis (SIPAIM)*. IEEE, 2023, pp. 1–5.

[24] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11976–11986.