Real-Time Multi-Label Upper Gastrointestinal Anatomy Recognition from Gastroscope Videos

Tao Yu, Huiyi Hu, Xinsen Zhang, Honglin Lei, Jiquan Liu, Weiling Hu, Huilong Duan and Jianmin Si

Abstract: Esophagogastroduodenoscopy (EGD) is a critical step in the diagnosis of upper gastrointestinal disorders. However, due to inexperience or high workload, there is a wide variation in EGD performance by endoscopists. Variations in performance may result in exams that do not completely cover all anatomical locations of the stomach, leading to a potential risk of missed diagnosis of gastric diseases. Numerous guidelines or expert consensus have been proposed to assess and optimize the quality of endoscopy. However, there is a lack of mature and robust methods to accurately apply to real clinical real-time video environments. In this paper, we innovatively define the problem of recognizing anatomical locations in videos as a multi-label recognition task. This can be more consistent with the model learning of image-to-label mapping relationships. We propose a combined structure of a deep learning model (GL-Net) that combines a graph convolutional network (GCN) with long short-term memory (LSTM) networks to both extract label features and correlate temporal dependencies for accurate real-time anatomical locations identification in gastroscopy videos. Our methodological evaluation dataset is based on complete videos of real clinical examinations. A total of 29,269 images from 49 videos were collected as a dataset for model training and validation. Another 1736 clinical videos were retrospectively analyzed and evaluated for the application of the proposed model. Our method achieves 97.1% mean accuracy (mAP), 95.5% mean per-class accuracy and 93.7% average overall accuracy in a multi-label classification task, and is able to process these videos in real-time at 29.9 FPS. In addition, based on our approach, we designed a system to monitor routine EGD videos in detail and perform statistical analysis of the operating habits of endoscopists, which can be a useful tool to improve the quality of clinical endoscopy.

Keywords: anatomy recognition; deep learning; endoscopy; multi-label; video analysis

1. Introduction

Gastric cancer [1] is the second leading cause of cancer-related deaths [2]. In clinical practice, Esophagogastroduodenoscopy (EGD) is a key step in the diagnosis of upper gastrointestinal tract disease. However, the rate of misdiagnosis and underdiagnosis of gastric diseases is high, reducing the detection of precancerous lesions and gastric cancer. This is because there is a great variation in EGD performed by endoscopists with different qualifications. On one hand, some inexperienced physicians may miss some critical areas and blind corners during the examination. On the other hand, physicians in densely populated areas face long examinations every day, which may lead to missed examinations and errors due to subjective mental or physical fatigue. This may result in the endoscopist not being able to comprehensively cover all anatomical locations throughout the stomach.
during the examination. Studies have shown that high-quality endoscopy can lead to more accurate diagnostic results [3], and it is crucial to further expand endoscopic techniques and improve routine endoscopy coverage and examination quality. Many authorities have now proposed clinical examination guidelines with corresponding expert consensus to evaluate and optimize the quality of endoscopy. The American Society of Gastrointestinal Endoscopy (ASGE) and the American College of Gastroenterology (ACG) have developed and published quality metrics common to all endoscopic procedures in EGD. The European Society of Gastroenterology (ESGE) systematically surveyed the available evidence and developed the first evidence-based performance measures for EGD (procedural integrity, examination time, etc.) in 2015 [4,5]. However, the lack of practical tools for rigorous monitoring and evaluation makes it difficult to apply many quantitative quality control indicators [6] (e.g., whether comprehensive coverage of anatomical location examination is achieved) in practice, which is a major constraint to quality control efforts.

The quality standard of GI endoscopy can be defined as: when doctors do endoscopy, they need to ensure that all key parts of the GI tract are within the scope of the examination, and maintain an appropriate observation duration, leaving no blind spots and avoiding the lens moving too fast or missing the observation of key areas. In recent years, advances in deep learning-based artificial intelligence technologies have continued to soar, with significant progress in the field of medical image recognition. Quality control of gastrointestinal endoscopy is the basis for the application of AI technology in endoscopic imaging and the prerequisite for applying artificial intelligence technology to disease screening and supplementary diagnosis. Advances have been made in the identification of gastric diseases [7,8], precancerous lesions [9–14] and gastric cancer [15–21]. It is important to use artificial intelligence systems to monitor the indicators in the quality control of gastrointestinal endoscopy in real time. However, previous studies have mainly focused on the intelligent auxiliary diagnosis of GI lesions. Due to the lack of relevant datasets for anatomical structures and the more complex and large data annotation efforts for this type of task, only a few studies were devoted to quality monitoring of routine endoscopy. Wu et al. [22] divided the anatomical location of the stomach into 10 and subdivided it into 26. DCNN was applied for anatomy classification. The final accuracy rates were 90% and 65.9%, respectively. Based on DCNN and reinforcement learning, 26 gastric anatomical locations were classified [23], and blind spots in EGD videos were monitored with an accuracy of 90.02%, which served to monitor the quality of real-time examinations. Ting et al. [24] proposed a deep ensemble feature network to combine the features extracted by multiple CNNs, to boost the recognition of three anatomical sites and two image modals with an accuracy of 96.9% and 23.8 frames per second (FPS). He et al. [25] divided the anatomical structure of endoscopy to 11 sites, and achieved 91.11% accuracy by using DenseNet121 [26]. The model was used to assist physicians in avoiding examination blind spots during examinations and to achieve comprehensive coverage of endoscopy.

Despite the good results of the above-mentioned studies on quality control of gastrointestinal endoscopy, some problems and challenges remain. First, all the above-mentioned work on anatomical location identification models is based on single-label multi-class classification, which deviates from the reality of actual clinical examinations. Multiple related anatomical sites are usually present simultaneously in the same image. When the ratio of multiple anatomical locations in the field of view is equal, a single label is not sufficient to accurately describe the currently examined location, which, in turn, increases the bias of model feature learning. Multi-label classification learning is more accurate in this application compared to multi-class image single-label classification [27], but it is challenging to further exploit this a priori relationship to improve the model accuracy due to the spatial correlation between anatomical locations, which leads to dependencies between labels. Second, all of the above work is based on anatomical location recognition models, which are trained based on static image data rather than real-time video data. It is not sufficient to identify anatomical locations under videos based on static image datasets alone. While consecutive video frames are highly similar, the dynamics of the scene cannot
be expressed in static images, and this dynamically changing data is important for the application of the model in real video scenes. Although the dynamic ones can produce severe scene blurring [28] and thus affect judgment due to camera motion and gases generated during surgery, etc., the impact of such blurred data can be mitigated by some means. In conclusion, spatial and temporal factors are strong priors for the anatomical relationships within the endoscope and between consecutive frames, and are key to further improving the performance of the recognition model.

In this paper, we present a novel combined structure of deep learning models to process EGD videos to accurately identify anatomical structures of the gastrointestinal tract on real-time white-light upper gastrointestinal tract. The task consists of classifying each single frame of an EGD image sequence into a number of anatomical structures in 25 sites. Our model is built on a combination of a graph convolutional network (GCN) and a long short-term memory (LSTM) network, where the GCN is used to capture label dependencies, and the LSTM is used to extract inter-frame temporal dependencies. Specifically, we train them jointly in an end-to-end way to relate coded label interdependencies and extract high-level features of visual and temporal information of consecutive video frames. The combined features learned by our method can correlate different anatomical structures under endoscopy and are sensitive to camera movements in the video, allowing accurate identification of all anatomical structures contained in each frame of a continuous video, especially the transition frames between different anatomical locations.

The main contributions of this paper are summarized as follows: (1) Unlike previous single-label multi-class studies, we define anatomical recognition as a multi-label classification task. This setting is more in line with clinical needs and real-time video-based examination. (2) GCN-based multi-label classification algorithm. In this paper, graph structure is introduced to learn domain prior knowledge, i.e., topological interdependencies between anatomical structure labels. A ResNet-GCN model is then constructed to implement multi-label classification. (3) Fusion of ResNet-GCN and LSTM modules. Due to the complexity of EGD endoscopy scenes, it is very difficult to classify the anatomical structures of each frame accurately. Considering that EGD videos have temporal continuity and anatomical structures have spatial continuity in the video sequence, we use LSTM to learn the temporal information and spatial continuity features of anatomical structures in EGD videos based on the ResNet-GCN model. Then, we fuse the ResNet-GCN module and the LSTM module to implement an end-to-end framework, called GL-Net, for the accurate identification of UGI anatomical structures. The model fully reflects the topological dependence of labels and the continuity of anatomical structures in time and space. (4) Retrospective analysis of EGD video quality based on the GL-Net model. The quality of 1736 real EGD videos was statistically analyzed in terms of the coverage of 25 anatomy sites observed, the total examination time generated by the endoscopists, the examination time of each specific site, and the ratio of valid to invalid frames according to the endoscopic guidelines and expert consensus. The statistical analysis of the indicators gives a quantitative evaluation of the quality of the endoscopists, indicating the practical feasibility of using AI technology to ensure the quality of EGD following clinical guidelines.

The rest of this paper is organized as follows. Section 1 describes the datasets and introduces our proposed method in detail. Section 2 demonstrates the experimental results, which are discussed in Section 3. Section 4 is the conclusion of our work.

2. Materials and Methods

An overview of our proposed approach is presented in Figure 1. We used a backbone CNN model to extract visual features from static images and a GCN classification network to learn the relationship between the labels. The LSTM structure was used to model the temporal association of consecutive frames and focus on the invariant target features in the spatio-temporal information to obtain more accurate recognition.

To better exploit the correlation between labels, recurrent neural networks [29,30], attention mechanisms [31], and probabilistic graphical models [32,33] are widely used.
Wang et al. [30] used RNNs to convert labels into embedding vectors to model the correlation between each label. Zhu et al. [34] proposed a spatial regularization network (SRN) with only image-level supervised learning of the spatial regularization between labels. Recently, Chen et al. [35] proposed a multi-label image recognition model based on GCN which can capture global correlations between labels and infer knowledge from beyond a single image, and achieved good results. Inspired by Chen’s work, we use graph structures to explore the dependencies between labels. Specifically, GCNs are used to disperse information from multiple labels so as to learn associative and dependent classifiers for each anatomically located label. These classifiers are further fused to image features to predict the correct outcome with label associations.

In the work of incorporating time series into deep learning models, many approaches based on dynamic time warping [36], conditional random fields [37], and hidden Markov models (HMMs) [38] have been proposed. However, the existing methods have some problems and challenges. For example, when exploring temporal correlation, these methods mostly focus on linear statistical models, which cannot accurately represent the complex temporal information during endoscopy. Second, it is difficult for these methods to accurately analyze transitional video frames where multiple targets are present at the same time, which are important for the accurate identification of anatomical locations. Several methods have been proposed to process sequential data by nonlinear modeling of temporal dependencies, such as LSTM, and have been successfully applied to many challenging tasks [28,39,40]. To address the problem of surgical procedure identification similar to EGD inspection, Jin Y et al. [28] introduced LSTM to learn temporal dependencies, and trained it in combination with convolutional neural networks. The learned temporal features are very sensitive to the changes of the surgical procedure, and can accurately identify the phase transition frames. Receiving inspiration from this approach, we proposed an LSTM fused with a GCN-based multi-label classification model for end-to-end training.

![Figure 1. Overall framework of our ResNet-GCN model for multi-label image recognition. D denotes the dimension of the feature maps, h and w denote height and width. c denotes the number of category classes. The blue dots indicate all class labels.](image)

2.1. Datasets

Following the guidance of ESGE [41] and the Japanese systematic screening protocol [42], three experts were invited to label the EGD images into 25 different anatomy sites. Representative images are shown in Figure 2. Since real endoscopy is performed under videos, severe noise (e.g., blood, bubbles, defocusing, artifacts, etc.) is generated. It is challenging to identify each image frame purely using video scenes alone. To improve the generalization ability of the dataset, 49 endoscopy videos were collected from Sir Run Run Shaw Hospital in this study. These videos were divided into a training set (39 videos) and a test set (10 videos), ensuring that images of the same case were not divided into both the training and test sets. We then split the videos into frames based on a sampling rate of 5 Hz,
ensuring that the video clips contained temporal information while introducing as little redundant information as possible. This is the offset adjusted according to experience [28]. The larger the span of frames, the greater the temporal variation, so adapting the model to this variation facilitates the establishment of inter-frame relationships and the removal of invalid frames (see Figure 3). After splitting and labeling, we have 23,471 training images and 5798 test images with multi-label annotations. In the training phase, we divided the training process into two stages. In the first ResNet-GCN phase, we put all qualified images in the video together to train the gastric part classification network (GCN). In the second training data preparation phase, we took ten consecutive frames as a segment and input them together into the LSTM network. All EGD videos were captured in white light endoscopy with an OLYMPUS EVIS LUCERA ELITE CLV-290SL at FPS of 25 and resolution of 1920 × 1080 per frame. Inspection of personal information (such as date of inspection and patient name) is removed to ensure privacy and security.

Figure 2. The representative images of UGI anatomical structures predicted by our model for multi-label classification in this paper. Specially, antrum, lower body, middle-upper body, and r-middle-upper body are further divided into four parts. Clockwise from the left side are the Anterior wall (A), Lesser curvature (L), Posterior wall (P), and Greater curvature (G). In addition, r stands for retroflex view.

Figure 3. Invalid frame diagram in EGD video scene. The upper left figure shows the defocusing caused by the lens being too close to the gastric mucosa, the upper right figure shows the motion artifact caused by the rapid movement of the lens, the lower left figure shows the light reflection of gastric mucus under light, and the lower right figure shows the large area of blood covering the gastric mucosa.

2.2. Backbone Structure

Many innovative model design and training techniques have emerged at this stage, including the attention mechanism [43], Transformer [44], and the excellent NAS-based EfficientNet [45]. Considering the universality, stability and generality of the methods, we selected ResNet [46] as the backbone network. The residual structure allows the model
capacity to vary in a flexible range, so that the model with the ResNet block as the basic unit can be built deep enough to complete the model convergence well. We use ResNet-50 [46], which was pre-trained on ImageNet [47], as the backbone network for feature extraction. Generally, the deeper the layers in the model, the larger the sensory field of the feature map and the higher the level of abstraction of the image features. Therefore, the proposed model structure attempts to extract features from one of the deepest convolutional layers of the backbone network for the construction of an attention map combining feature maps and association labels.

Let $I$ denote an input static image or one of consecutive frames with ground-truth multi labels $y = [y_1, y_2, \ldots, y_C]$, where $C$ is the count of all anatomical locations. The feature extraction process is expressed as

$$x = f_{\text{GMP}}(f_{\text{Backbone}}(I; \theta_{\text{Backbone}})) \in \mathbb{R}^{2048 \times 7 \times 7},$$

where $f_{\text{GAP}}(\cdot)$ present the operation of global max pooling, $f_{\text{Backbone}}(\cdot)$ denotes the feature extraction from backbone structure. $x$ is the compressed feature and contains the feature expressions in the image associated with the classification labels, which will be fused with the correlations between the labels in a matrix multiplication manner.

2.3. GCN Structure

In multi-label classification, multiple recognition targets usually appear together in an image. In some cases they must appear simultaneously, and in some cases they absolutely cannot appear at the same time. We need to efficiently establish the dependencies between targets to accurately establish feature representations in images, and correlations between multiple anatomical locations.

Since objects usually appear simultaneously in video scenes, the key to multi-label image recognition is to model the label dependencies, as shown in Figure 4. Inspired by Chen et al. [35], we model the interdependencies between anatomical locations using a graph structure where each node is a word embedding of an anatomical location, and that embedding feature is mapped to a set of classifiers constructed using GCN for image feature attention feature combinations. Thus, the approach preserves the semantic structure in the feature space and models label dependencies.

![Figure 4](image_url)

Figure 4. Multi-label images and their directed graphs. Within a single image, multiple anatomical locations appear and the image is divided into multiple regions, (A) Angulus, (B) Antrum L, (C) Antrum A and (D) Pylorus (the region indicated by the white arrow). At the right site, directed graphs are used to model label dependencies. The bidirectional arrows indicate that they are related and are likely to appear within the same view at the same time.

GCN is the operation on the graph structure. The structure uses the feature map and the corresponding correlation matrix as the input, and then updates the node features. The GCN structure can be written as follows:
\[ H^{l+1} = h\left(\hat{A}H^lW^l\right), \]  

where \( h(\cdot) \) represents a non-linear mapping, \( \hat{A} \) is the normalized matrix \( A \) and \( W^l \) is the transformation weight. \( H^{l+1} \) and \( H^l \) present the updated and current graph node representation.

The graph node presentation is then incorporated into the model output feature expression in the form of matrix multiplication, and the information is combined so that feature representations and labels are weighted and associated. The loss function, multi-label classification loss (e.g., binary cross entropy loss), is defined as follows:

\[
L = \sum_{c=1}^{C} y^c \log(\sigma(\hat{y}^c)) + (1 - y^c) \log(1 - \sigma(\hat{y}^c)),
\]

where \( \sigma(\cdot) \) is the activation function of sigmoid.

### 2.4. LSTM Structure

After the above structure is trained to process video frames based on static images, the final prediction results may fluctuate due to the presence of some poor quality frames in the video. Due to the continuity of video data, temporal information provides background information for each frame identification. At the same time, individual frames may have similar appearance under the same endoscopic anatomy and scene, or they may be slightly blurred, making it difficult to distinguish them purely by their visual appearance. In contrast, the phase identification of the current frame would be more accurate if we could take into account the dependence of the current frame on the adjacent past frames. Therefore, time series information is introduced in this study to improve the stability of the model.

Temporal information modeling. In our GL-Net, we input the image features extracted from the ResNet backbone network into the LSTM network, and use the memory units of the LSTM network to correlate current frame and past frame information for improved identification using temporal dependence.

Figure 5 demonstrates the fundamental LSTM [48] units used in GL-Net. Each LSTM cell is equipped with three gates: \( i_t \) denotes input gate, \( f_t \) denotes forget gate and \( o_t \) denotes output gate. Three units are used to regulate the interaction between memory cells \( c_t \). At timestep \( t \), given input \( x_t \), hidden state before \( h_{t-1} \), and memory cell before \( c_{t-1} \), LSTM structural units are learned and updated in the following manner:

\[
i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i),
\]

\[
f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f),
\]

\[
o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o),
\]

\[
g_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c),
\]

\[
c_t = f_t \odot c_{t-1} + i \odot g_t,
\]

\[
h_t = o_t \odot \tanh(c_t),
\]
In order to fully exploit both the label association and temporal information, we propose a new recursive convolutional network, GL-Net, as shown in Figure 6. GL-Net integrates ResNet-GCN for visual descriptor extraction with label-dependent association, and the LSTM network for temporal dynamic modeling. It outperforms existing methods for independent learning of visual and temporal features. We train GL-Net end-to-end, where the parameters of the ResNet structure and the LSTM structure are co-optimized to achieve better anatomical location recognition.

Figure 5. The structure of LSTM storage unit [49]. The arrows indicate the path of data forward propagation.

Figure 6. Overall framework of GL-Net model.
In detail, to identify the frames at time $t$, we extract the video clip containing a set of current frames. The sequence of frames in the video clip is represented by $x = \{x_{t'}, \ldots, x_{t-1}, x_t\}$. We use $f_j$ to denote the representative image features of each single frame $x_j$. The image features $f = \{f_{t'}, \ldots, f_{t-1}, f_t\}$ of the video clips are sequentially put into an LSTM network, which is denoted by $U_\theta$ with parameters $\theta$. With the input $x_t$ and the previous hidden state $h_{t-1}$, the LSTM calculates the output $o_t$ and the updated current hidden state $h_t$ as $o_t = h_t = U_\theta(x_t, h_{t-1})$. Finally, the prediction of frame $x_t$ is generated by feeding the output $o_t$ into the softmax:

$$\hat{p}_i = \text{sigmoid}(W_z o_t + b_z),$$

where $W_z$ and $b_z$ respectively denote the weight and bias term, $\hat{p}_i \notin \mathbb{R}^C$ is the predicted vector and $C$ denotes the number of classes.

Let $\hat{p}_i^t$ be the $i$-th element of $\hat{P}_t$, which denotes the prediction probability of frame $x_t$ and it belongs to the class $i$, $l_t$ denotes the ground truth of frame $x_t$, the negative log-likelihood of the frame of time $t$ can be calculated as:

$$\ell(x_t) = -\log \hat{p}_i^{l_t}(U_\theta(x_t)).$$

### 2.5. Experimental Setups

To efficiently train the proposed model structure, we train the ResNet-GCN network first in order to subsequently initialize the entire network, considering that the parameter size of the ResNet-GCN network is larger than that of the LSTM structural units. During the training process, the images are augmented with random horizontal flips.

After training the ResNet model, we trained the GL-Net, integrating visual, labeling, and temporal information to converge. At this point, the pre-trained ResNet parameters were initialized as the parameters of its backbone model, and the parameters of the LSTM structural unit were initialized using xavier, and, empirically, the learning rate of the LSTM was set to 10 times that of the ResNet-GCN.

Our proposed model is implemented based on the Pytorch [50] framework, using a TITAN V GPU. For the first stage, our proposed structure uses two connected GCN modules with dimensions of 1024 and 2048, respectively. In the image representation learning branch, we adopt ResNet-50 as the backbone of feature extraction, which is pretrained on ImageNet. For label representations, 25-dim one-hot word embedding is adopted. SGD is employed for training, with a batch size of 16 and momentum of 0.9, with a weight decay of $5 \times 10^{-3}$. The initial learning rate is set to 0.01, and decreased to $1/10$ every 10 epochs, until $1 \times 10^{-5}$.

In the end-to-end training stage, we use three LSTM layers. SGD is used as optimizer, with a batch size of 8, a momentum of 0.9 with weight decay of $1 \times 10^{-2}$, a dropout rate of 0.5, and we adopt LeakyReLU [51] as activation function. The learning rates are initially set as $1 \times 10^{-4}$ for ResNet and $1 \times 10^{-3}$ for LSTM, and are divided by a factor of 10 every 5 epochs. A total of 100 epochs were trained in the model.

### 3. Results

#### 3.1. Evaluation Metrics

The evaluation metrics adopted in this paper are consistent with [30,52]. We compute the overall precision, recall, F1 (OP, OR, OF1) and per-class precision, recall, F1 (CP, CR, CF1). For each image, the labels are predicted as positive if their confidences are greater than the threshold (i.e., 0.5 in experience). Following [48,53], we computed the average precision (AP) for each individual class, and the average precision (mAP) for all classes.

#### 3.2. Experimental Results

#### 3.2.1. GCN Structure

The statistical results are presented in Table 1, and we compared them with related current spatial-temporal methods, including CNN-RNN [30], RNN-Attention [31], etc.
The models involved in the comparison all used the same set of training and test data. The benchmark backbone was kept uniform for a fair comparison. It is clear to see that the GCN-based approach obtains the best classification performance, which is due to capturing the dependencies between the labels. Compared to advanced methods for capturing dependencies of frames, our method achieves better performance in almost all metrics, which proves the effectiveness of GCN. Specifically, the proposed GCN scheme obtained 93.1% of mAP, which is 21.1% higher than their method. Even using the ResNet-50 model as the backbone model, we could still achieve better results (+17.1%). This suggests that there are strong dependencies and correlations between anatomical location labels in full-coverage examinations in white-light endoscopy scenarios, and the basic backbone of CNN together with a GCN structure can capture them well.

Table 1. Comparison of average assessment results of anatomy multi-label identification with other methods. Bold indicates the maximum value.

| Method          | mAP  | CP   | CR   | CF1  | OP   | OR   | OF1  |
|-----------------|------|------|------|------|------|------|------|
| CNN-RNN [30]    | 0.720| 0.772| 0.772| 0.772| 0.785| 0.770| 0.777|
| RNN-Attention [31]| 0.649| 0.665| 0.578| 0.618| 0.610| 0.660| 0.634|
| ResNet-50       | 0.891| 0.936| 0.777| 0.849| 0.912| 0.734| 0.813|
| ResNet-GCN      | 0.931| 0.941| 0.855| 0.896| 0.929| 0.820| 0.871|
| GL-Net          | 0.971| 0.955| 0.959| 0.957| 0.937| 0.954| 0.945|

We further use the heatmaps to explain the model. By weighting and summing the class activation maps [54] of the final convolutional layer, the attention map can accurately highlight the areas of the image that have a high weight on recognition, thus revealing the network’s implicit attention to the image and intercepting the learning information of the network [54]. The attention maps of models is shown in Figure 7.

![Figure 7](image-url). Visualization of the attention maps. The label on the left indicates the label for visual gradient back propagation and also the ground truth annotation. The upper label indicates the model name of the visualization. The visualization region is approximately close to red, indicating that more of the model inference weight tends to be in that region.

As illustrated in Figure 7, both in the middle and upper part of the stomach body, the GCN-based model was the best in terms of visual representation. For the other models, it is easy to randomly place weights at some locations in the image without constructing label
associations. It is not possible to correctly distinguish between these anatomical structures that do not differ much between classes (greater curvature, posterior wall, anterior wall and less curvature of the body are difficult to distinguish). In contrast, the GCN-based model can pay more attention to feature regions in the image where texture features are prominent and responsive to the class. For the gastric angulus, because the structure of the angulus is prominent in the visual field, the general model can pay attention relatively accurately at this location. Compared to other models, the GCN-based model’s weights are able to provide more comprehensive and complete coverage at this location, including the lesser curvature of the antral.

3.2.2. GCN with LSTM Structure

To demonstrate the importance of combining label association and temporal features for this task, we carried out a series of experiments by combining ResNet-50 with different modeling approaches, namely (1) ResNet-50 with GCN, and (2) ResNet-50 with GCN followed by LSTM.

The experimental results are listed in Table 1. The scheme with LSTM achieved better results, demonstrating the importance of temporal correlation for more accurate identification. A total of 97.1% of mAP, 95.7% CF1 and 94.5% OF1 can be seen from the GL net method proposed in this paper. Specifically, compared with ResNet-GCN, our end-to-end trainable GL-Net improves the mAP, CF1 and OF1 by 4.0%, 6.1% and 7.4%, respectively. Similarly, we compared the average accuracy of the two schemes on each anatomical structure (see in Table 2). Compared with the ResNet-GCN model without the LSTM module, the accuracy of GL-Net in the anterior wall of middle-upper body, the lesser curvature of the lower body, the posterior wall of the middle-upper body, the large curvature of the middle-upper body and angulus were improved by 22.5%, 17.6%, 11.2%, 8.8% and 8.3%, respectively.

Table 2. Comparison of average precision of each individual anatomy identification with other methods. Bold indicates the maximum value.

| Anatomy                  | CNN-RNN | RNN-Attention | ResNet50 | ResNet-GCN | GL-Net |
|--------------------------|---------|---------------|----------|------------|--------|
| esophagus                | 0.903   | 0.821         | 0.998    | 0.999      | 1.0    |
| Squamocolumnar juction  | 0.880   | 0.748         | 0.986    | 0.994      | 0.999  |
| Cardia I                 | 0.867   | 0.715         | 0.965    | 0.967      | 0.998  |
| Cardia O                 | 0.873   | 0.658         | 0.994    | 1.0        | 1.0    |
| Fundus                   | 0.855   | 0.757         | 0.978    | 0.979      | 0.902  |
| Middle-upper body A      | 0.604   | 0.595         | 0.809    | 0.906      | 0.994  |
| Middle-upper body L      | 0.585   | 0.594         | 0.813    | 0.872      | 0.984  |
| Middle-upper body P      | 0.546   | 0.562         | 0.680    | 0.745      | 0.970  |
| Middle-upper body G      | 0.586   | 0.518         | 0.601    | 0.885      | 0.904  |
| Lower body A             | 0.595   | 0.548         | 0.841    | 0.918      | 0.918  |
| Lower body L             | 0.575   | 0.517         | 0.713    | 0.802      | 0.821  |
| Lower body P             | 0.585   | 0.454         | 0.816    | 0.876      | 0.936  |
| Lower body G             | 0.541   | 0.568         | 0.731    | 0.812      | 0.988  |
| Antrum A                 | 0.746   | 0.727         | 0.977    | 0.979      | 0.999  |
| Antrum L                 | 0.732   | 0.710         | 0.987    | 0.984      | 0.997  |
| Antrum P                 | 0.755   | 0.722         | 0.983    | 0.985      | 0.995  |
| Antrum G                 | 0.706   | 0.730         | 0.976    | 0.981      | 0.997  |
| Angulus G                | 0.793   | 0.756         | 0.898    | 0.905      | 0.988  |
| R-middle-upper body A    | 0.621   | 0.534         | 0.855    | 0.900      | 0.980  |
| R-middle-upper body L    | 0.648   | 0.542         | 0.892    | 0.925      | 0.959  |
| R-middle-upper body P    | 0.770   | 0.651         | 0.919    | 0.951      | 0.956  |
| R-middle-upper body G    | 0.766   | 0.636         | 0.908    | 0.945      | 0.997  |
| Duodenal bulb            | 0.751   | 0.623         | 0.997    | 0.998      | 0.999  |
| Duodenal descending      | 0.906   | 0.827         | 0.998    | 1.0        | 1.0    |
| Pylorus                  | 0.753   | 0.722         | 0.969    | 0.980      | 0.997  |
By introducing label association and temporal information, our GL-Net can learn features that are more discriminative than those produced by traditional CNNs that consider only visual information. Figure 8 shows the comparison of the prediction results of the two models in the video clips. It can be seen that due to the shooting angle, bubble reflection and other reasons, the variance between some classes is small and difficult to distinguish, or sometimes features are almost completely covered. Therefore, the ResNet-GCN network, which only depends on the features of a single frame image, cannot classify correctly, while GL-Net can avoid the error by considering time dependence, and each frame is identified accurately. In addition, some frames in the video have no classification results, that is, the confidence of all the predicted results does not exceed the set threshold, which may be related to the noise in the video. GL-Net can also accurately recognize each frame in this case, which indicates that GL-Net considering the temporal information can improve the performance of UGI anatomical structure recognition in EGD video. In addition, GL-Net can process these videos in real time at 29.9 FPS, a processing speed that has great potential for application in real-time clinical scenarios.

Figure 8. The prediction results on the video clips of ResNet-GCN and GL-Net. The bottom of each line corresponds to the results of ResNet-GCN, GL-Net and Ground Truth (GT) annotations separately. Specifically, 0-esophagus, 1-squamocolumnar juction, 3-outside of cardia, 9-greater curvature of lower body, 13-greater curvature of Antrum, 14-posterior wall of Antrum, 15-anterior wall of antrum, 17-angulus, 22-duodenal bulb, 24-pylorus.

3.2.3. Retrospective Analysis of EGD Videos

Based on the methodology proposed in this paper, we designed a framework for statistical analysis of the examination quality of real EGD videos in hospitals according to quality monitoring guidelines. We collected a total of 1736 EGD videos, all of which were captured with an OLYMPUS EVIS LUCERA ELITE CLV-290SL at 25 FPS and operated by
expert physicians. In addition to the anatomical position identification model proposed in this paper, our system uses an invalid frame filtering model \cite{55} to ensure that our statistical results are performed on clear and valid images.

The outputs of the proposed system are: (1) coverage statistics of the 25 sites observed; (2) total examination time; (3) examination time for each specific site; (4) the ratio of valid frames versus invalid frames.

The average coverage of anatomical structures during the EGD produced by the endoscopist was 85.81\%, but only 19.28\% of the patients were not blinded. In addition, the rate of misses for each anatomical structure (total number of videos with undetected anatomical structures/total number of videos) can be seen in Table 3. It can be clearly seen that most of the anatomical structures had a probability of being missed, except for the esophagus. Among them, the lower body lesser curvature had the highest miss rate of 52.41\%, indicating that this area tends to be a blind area in EGD surgery. In addition, the small curvature of the middle and upper body, the descending duodenum, the posterior wall of the lower body, and the large curvature of the middle and upper body also had a blind spot rate of more than 20\% in the retrograde view.

Table 3. The average miss rate of each UGI anatomical structure during EGD. Bold indicates the maximum value.

| Anatomy                        | Miss Rate (%) |
|--------------------------------|---------------|
| esophagus                      | 0.00          |
| Squamocolumnar junction        | 0.60          |
| Cardia I                       | 10.24         |
| Cardia O                       | 0.6           |
| Fundus                         | 9.04          |
| Middle-upper body A            | 7.83          |
| Middle-upper body L            | 10.24         |
| Middle-upper body P            | 15.06         |
| Middle-upper body G            | 43.98         |
| Lower body A                   | 13.86         |
| Lower body L                   | 23.49         |
| Lower body P                   | 18.67         |
| Lower body G                   | 52.41         |
| Antrum A                       | 9.64          |
| Antrum L                       | 13.25         |
| Antrum P                       | 11.45         |
| Antrum G                       | 9.64          |
| Angulus                        | 6.05          |
| R-middle-upper body A          | 22.89         |
| R-middle-upper body L          | 13.86         |
| R-middle-upper body P          | 10.24         |
| R-middle-upper body G          | 15.06         |
| Duodenal bulb                  | 5.42          |
| Duodenal descending            | 24.70         |
| Pylorus                         | 6.63          |

As shown in Table 4, the mean examination time for all the videos was 6.572 min, but the variance is quite different, which may be because some videos take biopsies or make abnormal findings. Considering that there were some blind spots in the process of EGD, we further analysed the inspection time when 25 sites were completely observed. As can be seen, it takes 7.37 min for endoscopists to check all the anatomical structures.
Table 4. Statistics of inspection time during EGD.

| Inspection Type                  | Mean (min) |
|----------------------------------|------------|
| Regular endoscopy                | 6.57       |
| Coverage of all anatomy          | 7.37       |

Table 5 shows the examination time of each specific anatomical structure. Obviously, the most time-consuming site is the esophagus, which takes 85.8 s, far more than the other 24 sites. However, the endoscopists spend the least time in the lower gastric body, and the average observation time of the lesser curvature of lower body is only 1.8 s.

Table 5. Inspection time of each UGI anatomical structures during EGD. Bold indicates the maximum value.

| Anatomy                          | Inspection Time (s) |
|----------------------------------|---------------------|
| esophagus                        | 85.8                |
| Squamocolumnar junction          | 15.6                |
| Cardia I                         | 31.2                |
| Cardia O                         | 26.4                |
| Fundus                           | 44.4                |
| Middle-upper body A              | 39                  |
| Middle-upper body L              | 19.2                |
| Middle-upper body P              | 7.8                 |
| Middle-upper body G              | 2.4                 |
| Lower body A                     | 22.2                |
| Lower body L                     | 8.4                 |
| Lower body P                     | 10.8                |
| Lower body G                     | 1.8                 |
| Antrum A                         | 31.2                |
| Antrum L                         | 30                  |
| Antrum P                         | 22.8                |
| Antrum G                         | 42                  |
| Angulus                          | 45                  |
| R-middle-upper body A            | 15                  |
| R-middle-upper body L            | 16.8                |
| R-middle-upper body P            | 34.2                |
| R-middle-upper body G            | 25.8                |
| Duodenal bulb                    | 57.6                |
| Duodenal descending              | 13.8                |
| Pylorus                           | 39.6                |

In addition, although no studies have clearly defined the effective time of endoscopists in operation, the visibility of mucosa has become an important indicator in the quality control guidelines of colonoscopy. Therefore, we believe that the proportion of invalid frames (including blood, bubbles, defocusing or artifacts) in the process of EGD also reflects the EGD quality. Based on this, we analysed the proportion of effective frames and invalid frames in the duration. According to the results in Table 6, the average ratio is about 2:7.

Table 6. The ratio of effective frame and invalid frame during EGD.

| /                                | Ratio (%) |
|----------------------------------|-----------|
| Effective Frames                 | 22.68     |
| Invalid Frames                   | 77.32     |

4. Discussion

In this study, we used actual clinical EGD videos for real-time identification of gastric anatomical structures and quality control of computer-aided gastroscopy. We designed an
efficient algorithm that integrates ResNet, GCN and LSTM networks to form the proposed GL-Net. The model achieves 97.1% mAP. Compared with previous works [23,24,56], we have the following advantages: (1) we propose a multi-label video frame-level gastric anatomical location identification method that can more accurately describe the physician’s current examination location with considerable clinical significance. (2) Our model can accurately identify anatomical locations in video frames and transition frames by learning label associations and spatio-temporal features correlation of images. (3) We conducted a quantitative statistical analysis of real EGD videos to summarize existing physicians’ operating habits and deficiencies, and to provide a quantitative analysis tool for effective implementation of examination quality control guidelines.

4.1. Recognition Evaluation

The purpose of this study was to use artificial intelligence to alleviate the problem that EGD quality control guidelines are not easily carried out and implemented in the clinic. In cases where the level of gastroscopy varies between physicians, there is a risk of missing a diagnosis if the entire endoscopy is not covered in that particular examination. Although similar work using CNN to assist in EGD quality control has been done in previous studies, there are several shortcomings. First, the use of a single label to represent the image is inaccurate, especially when the area occupied by different anatomical locations of the image in adjacent transition frames is large. Second, previous studies have mainly trained models on discrete still image data, which is insufficient in complex continuous video scenes and prone to a high number of false positive analysis values. In this study, we propose a new framework to address these problems. First, by introducing GCN in the training task to construct label associations, which, in turn, improves the accuracy of the location recognition. Second, the temporal associations in the video frames are addressed by introducing LSTM and continuous video frame datasets.

The GCN-based model outperforms other models with a uniform backbone structure, demonstrating its effectiveness in label-dependent modeling. The GCN has more advantages in multi-label modeling compared to the natural image dataset, which also indicates a strong interdependence between gastric anatomical locations. For analogs with lower scores, such as the large curvature and posterior wall in the upper middle body, and the small curvature, anterior and posterior wall in the lower body, there was a significant improvement. This is because GCN exploits the relationship between strong and weak label features extracted by models.

More importantly, we optimize the visual performance and sequential dynamics throughout the training process by introducing label associations and spatio-temporal priors. In general, the features generated by introducing more label associations and temporal feature constraints are more discriminative than those generated by traditional CNNs that consider only spatial information. In Figure 8, GL-Net can achieve accurate recognition results that conform to label association rules and correspond to image features, especially for frames with changing locations. In addition, based on LSTM, the results are more stable with fewer jumps, so the overall performance is improved, which is crucial for this task. Although there are many novel video-based 3D CNN methods proposed, we believe that compared to LSTM methods, 3D CNNs cannot provide correlations with longer connections due to the limitation of computational volume and computational speed. Therefore, we believe that using LSTM is the appropriate method for modeling temporal correlations. For the categories with relatively low scores, this may be due to the lack of distinct features and the insufficient number of datasets. However, considering the network performance, computational resources, and training difficulty, we use a 50-layer ResNet to implement GL-Net, so that the computational resources and training time can be controlled within a satisfactory range and satisfactory results can be obtained. With sufficient computational resources, we can choose a deeper CNN network to further improve the performance, or use multi-GPU distributed training.
In recent years, deep learning techniques in computer vision have made rapid progress, and representative recognition network structures such as VGGNet [57], Inceptions [58], ResNet [46], DenseNet [26], MobileNet [59], EfficientNet [45], and RegNet [60] have been expanding the accuracy, effectiveness, scale, and real-time performance of the networks. The Transformer [44], a self-attentive mechanism structure extending from the field of natural language processing (NLP), has given a trend to unify and combine image and text data. The reliance on data-driven deep learning models makes it easy for researchers to overlook the important role played by clinical priors in the application of medical image perception techniques; clinical tasks do not exist in isolation and data distributions are not independent of each other. Relationships between lesions and data feature distribution relationships have not been applied to the model design process. The research in this paper is inspired by the combination of clinical priori knowledge and deep learning methods. The major difference between our proposed method and previous single-label static frame methods is that the correlations between anatomical locations and the spatio-temporal relationship between consecutive frames are introduced into model design as constraints. This allows us to achieve better evaluation of our model under the same feature extraction backbone structure with the relational constraints introduced by GCN and LSTM.

4.2. Clinical Retrospective Analysis

Observing the integrity of all 25 locations is of paramount importance, however, we found that only 19.28% of patients were observed at all locations and nearly five locations had a leak rate of more than 20%. This suggests that the quality of endoscopy needs to be improved.

Studies have shown that spending more time on EGD improves the diagnostic rate, so we recorded the total procedure time during EGD and counted the procedure time per anatomical location based on model analysis. This helps the endoscopist to control the duration of each examination procedure, thereby reducing variability in the level of examination due to competent factors such as experience and fatigue examinations. This study concluded that “slow” endoscopists (who take on average more than 7 min to perform a normal endoscopy) are more likely, or even up to two times more likely, to detect high-risk gastric lesions [61]. However, in a retrospective analysis of experimental data results, the total time of the procedure was lower than the recommended time. Therefore, we recommend that endoscopists be able to increase the examination time further.

Among the various sites, the esophagus was the only one that was not missed in all videos, and had the longest examination time. On the one hand, this is because the esophagus has a certain length in space and is the entrance to the EGD. On the other hand, we have patients with Barrett’s esophagus [62] in our videos. Studies have shown that the examination time of Barrett’s esophagus is related to the detection rate of the associated tumor [63]. The less time spent on the lower body curvature also contributes to its high rate of missed diagnoses. The effective examination time is only 23%, so the mucosal visibility of UGI is not high enough during most EGD examinations, which is due to invalid frames when the endoscopist performs operations such as flushing and insufflation, or when the lens shakes and fails to focus. This value can be used as a reference indicator. For endoscopists with a high percentage of invalid frames, further demands can be made on the operation level.

With these data, we can clearly see the behavioral habits of Chinese doctors in gastroscopy and the possible blind spots. It is beneficial for the system to achieve quality monitoring, improve the quality of gastroscopy, and further improve the detection rate of diseases. All the indicators mentioned in this paper can reflect the details of the gastroscopy process to some extent. These indicators prove that our model has great potential value for application to improve the quality of examination.
5. Conclusions

In this paper, we propose a novel and effective recursive convolutional neural network, GL-Net, for automatic recognition of the anatomical location of the stomach in EGD videos. GL-Net consists of two partial structures, namely GCN and LSTM, which are used to extract label-dependent and time-dependent features, respectively.

The GCN part of our method is able to extract the label dependency of multi-label image recognition, compared to the currently related study of static image based recognition methods for single-label multi-class anatomical location recognition. Meanwhile, the spatial-temporal features extracted by the LSTM part are able to identify adjacent similar frames more accurately.

In addition, we designed a real-time system based on the GL-Net method to automatically monitor detailed metrics during EGD (e.g., anatomical examination coverage, effective observation frame statistics, observation statistics of each anatomical site, etc.) and perform statistical analysis on the quality of EGD examinations. A quantitative assessment of the quality of the endoscopist’s examination is performed to demonstrate the professional operating habits of the endoscopist and the presence of potential accidents and problems. It also demonstrates the feasibility of implementing endoscopic quality control guidelines using artificial intelligence technology. It can effectively mitigate the subjective and empirical differences among endoscopists, improve the quality of routine endoscopy, and provide a reference for writing endoscopy reports and performing clinical procedures in real time with anatomical positions. In the future, the combination of anatomical position identification results and endoscopic mucosal health condition for comprehensive analysis is expected to further improve the quality control of computer-assisted endoscopy and assist in lesion diagnosis.

We believe that computer-aided detection and artificial intelligence techniques will play an increasing role. The rapid changes in model structure in recent years have allowed us to use increasingly advanced approaches to clinical data. However, the characteristics of the data distribution should be considered more in studies, such as the multi-label classification in this paper, which is more clinically realistic than single-label classification, and the potential associations within clinical prior knowledge and tasks, such as the construction of inter-label associations with the spatio-temporal association in this paper. Incorporating researchers’ or clinicians’ prior knowledge into the model training process is a more specific, accurate, and reliable solution to obtain practical solutions. We believe that in the future development of deep learning medical imaging research work, AI technology and medical knowledge will be further integrated to obtain further technical breakthroughs, as well as playing a greater role in the clinic and being more easily accepted by the public.

Author Contributions: Conceptualization, T.Y. and H.H.; methodology, T.Y. and H.H.; software, X.Z.; validation, H.L.; formal analysis, T.Y.; investigation, J.L.; resources, J.L.; data curation, W.H.; writing—original draft preparation, T.Y. and H.H.; writing—review and editing, J.L.; visualization, X.Z.; supervision, H.D. and J.S.; project administration, J.L.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of Zhejiang, China (No.2021C03111) and the National Natural Science Foundation of China (No.81827804).

Institutional Review Board Statement: Ethical review and approval were waived for this study, due to the retrospective design of the study and the fact that all data used were from existing and anonymized clinical datasets.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the patient(s) to publish this paper.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.
Abbreviations
The following abbreviations are used in this manuscript:

- EGD: Esophagogastroduodenoscopy
- GCN: graph convolutional network
- LSTM: long short-term memory
- ASGE: The American Society of Gastrointestinal Endoscopy
- ACG: The American College of Gastroenterology
- ESGE: The European Society of Gastroenterology
- HMMs: hidden Markov models

References

1. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 818–833.
2. Ang, T.L.; Fock, K.M. Clinical epidemiology of gastric cancer. Singap. Med. J. 2014, 55, 621–628. [CrossRef]
3. Rutter, M.D.; Rees, C.J. Quality in gastrointestinal endoscopy. Endoscopy 2014, 46, 526–528. [CrossRef] [PubMed]
4. Cohen, J.; Safdi, M.A.; Deal, S.E.; Baron, TH.; Chak, A.; Hoffman, B.; Jacobson, B.C.; Mergener, K.; Petersen, B.T.; Petrinj, J.L.; et al. Quality indicators for esophagogastroduodenoscopy. Gastrointest. Endosc. 2006, 63, S10–S15. [CrossRef] [PubMed]
5. Park, W.G.; Cohen, J. Quality measurement and improvement in upper endoscopy. Tech. Gastrointest. Endosc. 2012, 14, 13–20.
6. Brethauer, M.; Aabakken, L.; Dekker, E.; Kaminski, M.F.; Roesch, T.; Hultcrantz, R.; Suchanek, S.; Jover, R.; Kuipers, E.J.; Bisschops, R.; et al. Requirements and standards facilitating quality improvement for reporting systems in gastrointestinal endoscopy: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement. Endoscopy 2016, 48, 291–294. [CrossRef]
7. Nayyar, Z.; Khan, M.; Alhussein, M.; Nazir, M.; Aurangzeb, K.; Nam, Y.; Kadry, S.; Haider, S. Gastric Tract Disease Recognition Using Optimized Deep Learning Features. CMC-Comput. Mater. Contin. 2021, 68, 2041–2056. [CrossRef]
8. Zhang, X.; Chen, F.; Yu, T.; An, J.; Huang, Z.; Liu, J.; Hu, W.; Wang, L.; Duan, H.; Si, J. Real-time gastric polyp detection using convolutional neural networks. PLoS ONE 2019, 14, e0214133. [CrossRef]
9. Guimares, P.; Keller, A.; Fehlmann, T.; Lammert, F.; Casper, M. Deep-learning based detection of gastric precancerous conditions. Gut 2020, 69, 4–6. [CrossRef]
10. Wang, C.; Li, Y.; Yao, J.; Chen, B.; Song, J.; Yang, X. Localizing and Identifying Intestinal Metaplasia Based on Deep Learning in Oesophagoscope. In Proceedings of the 2019 8th International Symposium on Next Generation Electronics (ISNE), Zhengzhou, China, 9–10 October 2019; pp. 1–4. [CrossRef]
11. Yan, T.; Wong, P.K.; Choi, I.C.; Vong, C.M.; Yu, H.H. Intelligent diagnosis of gastric intestinal metaplasia based on convolutional neural network and limited number of endoscopic images. Comput. Biol. Med. 2020, 126, 104026. [CrossRef]
12. Zheng, W.; Zhang, X.; Kim, J.; Zhu, X.; Ye, G.; Ye, B.; Wang, J.; Luo, S.; Li, J.; Yu, T.; et al. High Accuracy of Convolutional Neural Network for Evaluation of Helicobacter pylori Infection Based on Endoscopic Images: Preliminary Experience. Clin. Transl. Gastroenterol. 2019, 10, e0109. [CrossRef]
13. Itoh, T.; Kawahira, H.; Nakashima, H.; Yata, N. Deep learning analyzes Helicobacter pylori infection by upper gastrointestinal endoscopy images. Endosc. Int. Open 2018, 6, E139–E144. [CrossRef]
14. Lin, N.; Yu, T.; Zheng, W.; Hu, H.; Xiang, L.; Ye, G.; Zhong, X.; Ye, B.; Wang, R.; Deng, W.; et al. Simultaneous Recognition of Atrophic Gastritis and Intestinal Metaplasia on White Light Endoscopic Images Based on Convolutional Neural Networks: A Multicenter Study. Clin. Transl. Gastroenterol. 2021, 12, e00385. [CrossRef]
15. Lee, J.H.; Kim, Y.J.; Kim, Y.W.; Park, S.; Choi, Y.J.; Kim, Y.J.; Park, D.K.; Kim, K.G.; Chung, J.W. Spotting malignancies from gastric endoscopic images using deep learning. Surg. Endosc. Other Interv. Tech. 2019, 33, 3790–3797. [CrossRef]
16. Zhu, Y.; Wang, Q.C.; Xu, M.D.; Zhang, Z.; Cheng, J.; Zhong, Y.S.; Zhang, Y.Q.; Chen, W.F.; Yao, L.Q.; Zhou, P.H.; et al. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. Gastrointest. Endosc. 2019, 89, 806–815.e1. [CrossRef]
17. Ikenoyama, Y.; Hirawasa, T.; Ishioka, M.; Namikawa, K.; Yoshimizu, S.; Horiuchi, Y.; Ishiyama, A.; Yoshio, T.; Tsuchida, T.; Takeuchi, Y.; et al. Detecting early gastric cancer: Comparison between the diagnostic ability of convolutional neural networks and endoscopists. Dig. Endosc. 2021, 33, 141–150. [CrossRef]
18. Ueyama, H.; Kato, Y.; Akazawa, Y.; Yatagai, N.; Komori, H.; Takeda, T.; Matsumoto, K.; Ueda, K.; Matsumoto, K.; Hojo, M.; et al. Application of artificial intelligence using a convolutional neural network for diagnosis of early gastric cancer based on magnifying endoscopy with narrow-band imaging. J. Gastroenterol. Hepatol. 2021, 36, 482–489. [CrossRef]
19. Ling, T.; Wu, L.; Fu, Y.; Xu, Q.; An, P.; Zhang, J.; Hu, S.; Chen, Y.; He, X.; Wang, J.; et al. A deep learning-based system for identifying differentiation status and delineating the margins of early gastric cancer in magnifying narrow-band imaging endoscopy. Endoscopy 2021, 53, 469–477. [CrossRef]
20. Saito, H.; Aoki, T.; Aoyama, K.; Kato, Y.; Tsutui, A.; Yamada, A.; Fujishiro, M.; Oka, S.; Ishihara, S.; Matsuda, T.; et al. Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network. Gastrointest. Endosc. 2020, 92, 144–151.e1. [CrossRef]
The image contains a page from a document, likely a journal or conference proceedings. The text is formatted in a list of references, each entry formatted in a standard bibliographic style. Here is a representation of the content in plain text:

21. Hu, H.; Gong, L.; Dong, D.; Zhu, L.; Wang, M.; He, J.; Shu, L.; Cai, Y.; Cai, S.; Su, W.; et al. Identifying early gastric cancer under magnifying narrow-band images with deep learning: A multicenter study. *Gastroint. Endosc.* 2022, 103, 1333–1341.e3. [CrossRef]

22. Wu, L.; Zhou, W.; Wan, X.; Zhang, J.; Shen, L.; Hu, S.; Ding, Q.; Mu, G.; Yin, A.; Huang, X.; et al. A deep neural network improves endoscopic detection of early gastric cancer without blind spots. *Endoscopy 2019*, 51, 522–531. [CrossRef]

23. Wu, L.; Zhang, J.; Zhou, W.; An, P.; Shen, L.; Liu, J.; Jiang, X.; Huang, X.; Mu, G.; Wan, X.; et al. Randomised Controlled Trial of WISENSE, a Real-Time Quality Improving System for Monitoring Blind Spots during Esophagogastrroduodenoscopy. *Gut* 2019, 68, 2161–2169. [CrossRef] [PubMed]

24. Lin, T.H.; Jiang, J.Y.; Huang, C.R.; Tsai, Y.C.; Cheng, H.C.; Sheu, B.S. Deep Ensemble Feature Network for Gastric Section Classification. *IEEE J. Biomed. Health Inform.* 2021, 25, 77–87. [CrossRef] [PubMed]

25. He, Q.; Bao, S.; Ahmad, O.F.; Yang, B.; Chen, X.; Valdastri, P.; Lovat, L.B.; Stoyanov, D.; Zuo, S. Deep learning-based anatomical site classification for upper gastrointestinal endoscopy. *Int. Comput. Assist. Radiol. Surg.* 2020, 15, 1085–1094. [CrossRef]

26. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* 2016, arXiv:1608.06993.

27. Liu, L.; Wang, P.; Shen, C.; Wang, L.; Van Den Hengel, A.; Wang, C.; Shen, H.T. Compositional Model Based Fisher Vector Coding for Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* 2017, 39, 2335–2348. [CrossRef] [PubMed]

28. Jin, Y.; Dou, Q.; Chen, H.; Yu, L.; Qin, J.; Fu, C.; Heng, P. SV-RCNet: Workflow Recognition From Surgical Videos Using Recurrent Convolutional Network. *IEEE Trans. Med. Imaging* 2018, 37, 1114–1126. [CrossRef] [PubMed]

29. Chen, S.F.; Chen, Y.C.; Yeh, C.K.; Wang, Y.C.F. Order-Free RNN with Visual Attention for Multi-Label Classification. *arXiv* 2017, arXiv:1707.05495.

30. Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; Xu, W. CNN-RNN: A Unified Framework for Multi-label Image Classification. *arXiv* 2016, arXiv:1604.04573.

31. Wang, Z.; Chen, T.; Li, G.; Xu, R.; Lin, L. Multi-label Image Recognition by Recurrently Discovering Attentional Regions. *arXiv* 2017, arXiv:1711.02816.

32. Li, Q.; Qiao, M.; Ban, W.; Tao, D. Conditional Graphical Lasso for Multi-label Image Classification. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2977–2986.

33. Li, X.; Zhao, F.; Guo, Y. Multi-Label Image Classification with a Probabilistic Label Enhancement Model. In Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence, UAI’14, Quebec City, QC, Canada, 23–27 July 2014; AUAI Press: Arlington, VA, USA, 2014; pp. 430–439.

34. Zhu, F.; Li, H.; Ouyang, W.; Yu, N.; Wang, X. Learning Spatial Regularization with Image-Level Supervisions for Multi-label Image Classification. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Los Alamitos, CA, USA, 2017; pp. 2027–2036. [CrossRef]

35. Chen, Z.M.; Wei, X.S.; Wang, P.; Guo, Y. Multi-Label Image Recognition with Graph Convolutional Networks. *arXiv* 2019, arXiv:1904.03582.

36. Padoy, N.; Blum, T.; Ahmadi, S.A.; Feussner, H.; Berger, M.O.; Navab, N. Statistical modeling and recognition of surgical workflow. *Med. Image Anal.* 2012, 16, 632–641. [CrossRef]

37. Tao, L.; Zappella, L.; Hager, G.D.; Vidal, R. Surgical Gesture Segmentation and Recognition. *Med. Image Comput. Comput. Assist. Interv.* 2013, 16, 339–346. [CrossRef]

38. Lalys, F.; Riffraud, L.; Morandi, X.; Jannin, P. Surgical Phases Detection from Microscope Videos by Combining SVM and HMM. In *Medical Computer Vision. Techniques and Applications in Medical Imaging*; Menze, B., Langs, G., Tu, Z., Criminisi, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 54–62.

39. Gers, F.A.; Eck, D.; Schmidhuber, J. Applying LSTM to Time Series Predictable through Time-Windows Approaches. In Proceedings of the Artificial Neural Networks—ICANN 2001, Vienna, Austria, 21–25 August 2001; Dorffner, G., Bischof, H., Hornik, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2001; pp. 669–676.

40. Zeng, T.; Wu, B.; Zhou, J.; Davidson, I.; Ji, S. Recurrent Encoder-Decoder Networks for Time-Varying Dense Prediction. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 18–21 November 2017; pp. 1165–1170. [CrossRef]

41. Bisschops, R.; Areia, M.; Coron, E.; Dobre, D.; Kaskas, B.; Kuvarev, R.; Pech, O.; Ragunath, K.; Weusten, B.; Familiari, P.; et al. Performance measures for upper gastrointestinal endoscopy: A European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative. *Endoscopy 2016*, 48, 843–864. [CrossRef]

42. Yao, K.; Uedo, N.; Kamada, T.; Hirasa, T.; Nagahama, T.; Yoshinaga, S.; Oka, M.; Inoue, K.; Mabe, K.; Yao, T.; et al. Guidelines for endoscopic diagnosis of early gastric cancer. *Dig. Endosc.* 2020, 32, 663–698. [CrossRef]

43. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *arXiv* 2017, arXiv:1709.01507.

44. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* 2020, arXiv:2010.11929.

45. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* 2019, arXiv:1905.11946.

46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* 2015, arXiv:1512.03385.

47. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM 2017*, 60, 84–90. [CrossRef]
48. Dong, J.; Xia, W.; Chen, Q.; Feng, J.; Huang, Z.; Yan, S. Subcategory-Aware Object Classification. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 827–834. [CrossRef]
49. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* 1997, 9, 1735–1780. [CrossRef]
50. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv* 2019, arXiv:1912.01703.
51. Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical Evaluation of Rectified Activations in Convolutional Network. *arXiv* 2015, arXiv:1505.00853.
52. Ge, W.; Yang, S.; Yu, Y. Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning. *arXiv* 2018, arXiv:1802.09129.
53. Wei, Y.; Xia, W.; Lin, M.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; Yan, S. HCP: A Flexible CNN Framework for Multi-Label Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* 2016, 38, 1901–1907. [CrossRef]
54. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. *arXiv* 2015, arXiv:1512.04150.
55. Xu, Z.; Tao, Y.; Wenfang, Z.; Ne, L.; Zhengxing, H.; Jiquan, L.; Weiling, H.; Huilong, D.; Jianmin, S. Upper gastrointestinal anatomy detection with multi-task convolutional neural networks. *Healthc. Technol. Lett.* 2019, 6, 176–180.
56. Chang, Y.Y.; Li, P.C.; Chang, R.F.; Yao, C.D.; Chen, Y.Y.; Chang, W.Y.; Yen, H.H. Deep learning-based endoscopic anatomy classification: An accelerated approach for data preparation and model validation. *Surg. Endosc.* 2021, 1–11. [CrossRef]
57. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. In Proceedings of the Neural Information Processing Systems (NIPS’14), Montreal, QC, Canada, 8–13 December 2014; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; MIT Press: Cambridge, MA, USA, 2014; Volume 1, pp. 568–576.
58. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *arXiv* 2014, arXiv:1409.4842.
59. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* 2017, arXiv:1704.04861.
60. Radosavovic, I.; Prateek Kosaraju, R.; Girshick, R.; He, K.; Dollár, P. Designing Network Design Spaces. *arXiv* 2020, arXiv:2003.13678.
61. Teh, J.L.; Tan, J.R.; Lau, I.J.F.; Saxena, N.; Salim, A.; Tay, A.; Shabbir, A.; Chung, S.; Hartman, M.; Bok-Yan So, J. Longer Examination Time Improves Detection of Gastric Cancer During Diagnostic Upper Gastrointestinal Endoscopy. *Clin. Gastroenterol. Hepatol.* 2015, 13, 480–487.e2. [CrossRef] [PubMed]
62. Conio, M.; Filiberti, R.; Blanchi, S.; Ferraris, R.; Marchi, S.; Ravelli, P.; Lapertosa, G.; Iaquito, G.; Sablich, R.; Gusmaroli, R.; et al. Risk factors for Barrett’s esophagus: a case-control study. *Int. J. Cancer* 2002, 97, 225–229. [CrossRef] [PubMed]
63. Gupta, N.; Gaddam, S.; Wani, S.B.; Bansal, A.; Rastogi, A.; Sharma, P. Longer inspection time is associated with increased detection of high-grade dysplasia and esophageal adenocarcinoma in Barrett’s esophagus. *Gastrointest. Endosc.* 2012, 76, 531–538. [CrossRef] [PubMed]