ISBN: 978-981-18-1791-5
2021 the 11th International Workshop on Computer Science and Engineering (WCSE 2021)
doi: 10.18178/wcse.2021.06.013

Classification of Gastric Cancer based on Teacher-attention Distillation and Improved Dual-path Network

Chenchao Huang [1], Wei Peng [2], Kun Yu [1 +] and Wenzhi Bao [1]

[1] School of Data Science Engineering, East China Normal University, Shanghai, China

[2] Information Technology Services, East China Normal University, Shanghai, China

Abstract. Early discovery of gastric cancer plays an important role in the clinical prognosis of gastric cancer. In recent years, using deep learning to classify CT medical images effectively improves the accuracy and efficiency of classification. But traditional deep learning models lack the ability to learn rich contextual information from CT data. This paper aims to explore the way that use weakly label area image to improve the accuracy of classification of gastric CT (Computer Tomography, CT) image differentiation. We propose a new model structure that combines improved dual-path network (DPN) to reuse and mine new image features，and uses teacher attention distillation to encode rich contextual information. Combining our contributions, we are able to achieve most 0.8135 AUC. The overall results show better AUC performance than the state-of-the-art.

Keywords: Differentiation status of gastric cancer, CT image classification, DPN, TAD

1. Introduction

According to the latest National Cancer Report [1] released by the National Cancer Centre in 2019, the incidence of gastric cancer ranks second among all cancers, which greatly threatens the health of the human. The status of differentiation of gastric cancer is closely related to the epidemiology and prognosis of patients. If the status of differentiation of gastric cancer can be accurately assessed, it will help guide clinical precision treatment, formulate individualized treatment plans and evaluate the prognosis of patients [2-3].

Gastroscopy biopsy is the way to diagnose early gastric cancer with high accurate [4]. But there are still errors in the judgment of the status of pathological differentiation of gastric cancer, mainly because of the selection of specimens, which affects the results of the biopsy [5]. In addition, gastric cancer detection methods include radioactivity detection method [6], ultrasound detection method [7], serum gastric function detection method [8], but they often require multiple sampling and testing to get accurate results. CT images have good clinical value in the diagnosis of gastric cancer [9]. Sun et al [10] have introduced a gastric ulcer differentiation system based on deep convolutional neural network (CNN), but their network lacks the exploration of contextual information. Wang et al [11] have used pre-trained CNN to speed up the model convergence, but because of the difference between the pre-training data (ImageNet) and the target data (CT data), it is difficult to say that the use of transfer learning can bring substantial effect to the task. However, few people currently study the classification of gastric cancer based on deep-learning method. The main reason is the lack of public data sets of gastric tumours. If the deep learning methods of other cancers are directly used [12-14], the difference in tumour morphology will be ignored. Therefore, it is necessary to design a network to classify the status of differentiation of gastric cancer.

We propose a new model structure for classification of the differentiation status of gastric cancer, which combines Improved-DPN and teacher-attention distillation (TAD) module. The purpose of TAD is to guide the network to pay more attention to the teacher area (Labelled area）during the learning process. Improved dual-path network [15] is added to reuse and mine new image features. The major contributions of our work are:

[+] Kun Yu. Tel.: 13917230427; fax: 021-62576192
E-mail address: kyu@cc.ecnu.edu.cn

- We propose a novel model based on ResNeXt [16]. We add TAD and Improved-DPN into our model, which can encode rich contextual information and are helpful for classification effect.
- We integrate TAD with model. By adding TAD, the deeper block to mimic the feature maps of labelled area or a preceding block so that our network can focus more on labelled area.
- We improve the DPN and add it into our model, which solves the defect that DPN cannot add a self-attention feature map and improves the model effect by introducing the attention map generated by its own layer.
- We expend the data to satisfy the deep learning training without changing the characteristics of the CT image.

The rest of this paper is organized as follows. In section 2, there is a brief overview of related work. In section 3, we display our network architecture including each module and illustrate the focal loss function. Experiments and discussions are conducted in section 4. Finally, we conclude with a summary of our main contributions and results.

2. Related Work

In this section we will briefly review two different methods for gastric cancer research, which are traditional detection methods and deep learning methods.

2.1. Traditional Method

Gastroscopy biopsy has high accuracy in the detection, pathological typing and differentiation of gastric cancer, which has good clinical value. Huang Yan [5] proved that preoperative gastroscopy biopsy was used for pathological diagnosis of gastric cancer. Jiang [17] studied the relationship between the expression of NEUROG3 gene and the status of differentiation of gastric cancer tissues. Hu et al. [18] used tumour markers to diagnose gastric cancer. They found that the detection of Ang-2, Vav1, CEA, and CA724 had certain clinical application value in the diagnosis of gastric cancer and could be used as one of the laboratory evaluation indicators for gastric cancer. However, most of the traditional methods require sampling, repeated inspections, and waiting for the test results.

2.2. Deep-learning Method

Sun et al [10] introduced an objective and precise gastric ulcer differentiation system based on deep convolutional neural network which could support the specialists by improving the diagnostic accuracy of the endoscopic examination of gastric ulcers. Wang et al. [11] used deep convolutional neural networks and transfer learning to classify CT images of lung nodules, which overcame the problem of overfitting under limited training data, and the classification accuracy rate reached 71.88%. Zhu et al [12] proposed a weakly supervised model to classify and localize the gastric cancer region in the pathological image with image-level labels. Sajja et al [13] proposed a deep neural network that was designed based on Google Net, which was a pre-trained CNN for CT Scan Images. They achieved better classification accuracy than the contrastive networks.

The existing deep learning model cannot introduce external weak supervised learning annotation to give more attention to the tumour region. It cannot extract the context information of CT image to better acquire the context information of tumour shape, location and domain.

3. Methodology

The main architecture of our method is shown in Fig. 1. There are two branches, the main branch and the auxiliary training branch. First, we put preprocessed CT image data into the backbone to obtain feature map. The feature map passes the Improved-DPN. The output of Improved-DPN passes through the fully connected layer and softmax layer to get the result of classification. The auxiliary training branch is to guide network pay more attention to label area. We add the losses of the above two parts to obtain the total loss of the model. Specific information about each module will be provided in the following subsections.

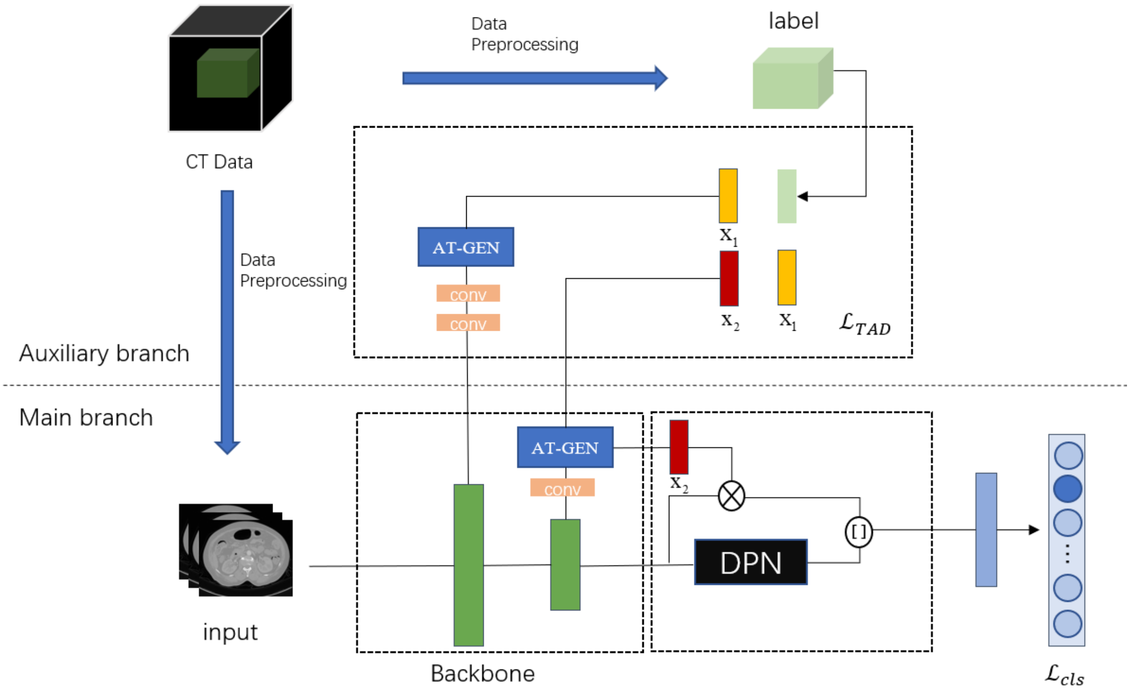

Fig. 1: The main architecture of our methodology.

3.1. Attention Generator

We introduce attention generator proposed by Hou Y et.al [19] into our work to generate attention feature maps for follow-up. Attention have great effects on the human visual experience. In addition, it has recently been demonstrated that attention can also play an important role in the context of applying artificial neural networks to various tasks in the fields of computer vision and NLP [20]. Attention maps mainly include activation-based attention maps and gradient-based attention maps [19]. AT-GEN is inspired by activation-based attention distillation. The structure of the AT-GEN is shown in Fig. 2. We first enlarge the size of the input image through bilinear interpolation. The process of reshape is to use the following method to compress the number of channels. Generation method is to superimpose the multi-channel feature maps into a single-channel map, where the superposition is the superposition after the n power of itself. We use $A_m \in R^{C_m \times H_m \times W_m}$ to denote the activation output of the m layer of the network, where C_m, H_m and W_m represent the number of channels, height and width of the image respectively. Note the generation mapping function $G: R^{C_m \times H_m \times W_m} \to R^{H_m \times W_m}$. After cross-validation, it is found that the effect is best when n is 2, that is $g_{sum}^2(A_m) = \sum_{i=1}^{C_m} |A_m|^2$. We take the value of n equals 2 for the experiment when superimposing, Finally, we get the attention map after normalization.

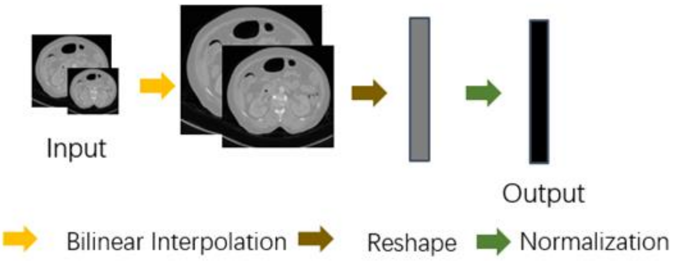

Fig. 2: The architecture of AT-GEN.

3.2. Improved Dual-path Network

We improve the dual-path network [10] by adding a branch of activation-based attention. As is known to us, DPN is a dual-path network that combines the advantages of the Resnet [11] and DenseNet [21], which is a typical hybrid network. But DPN cannot add a self-attention feature map. We put the output of backbone to pass through the AT-GEN to get the attention feature map. The obtained attention feature map is merged with the output of DPN to get the final output. We consider that by adding attention map into DPN can not only redevelop the features, but also discover new features better. The structure of the Improved-DPN is shown in Fig. 3.

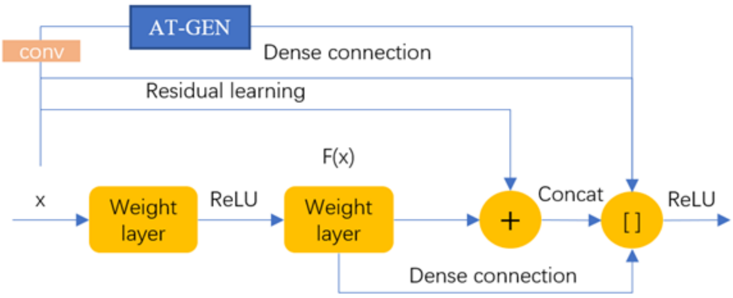

Fig. 3: The architecture of Improved-DPN.

3.3. Teacher Attention Distillation

We designed an auxiliary training branch to make better use of doctor's weak supervision annotation. The structure of the TAD is shown in Fig. 4. We put the fourth and fifth feature maps of the residual network into AT-GEN to generate attention maps x_1 and x_2, respectively. Then we let x_1 learn new features from the teacher (labelled area) and then x_2 can learn from x_1. Through this improvement, our model is able to make the deeper block to mimic the feature maps of labelled area or a preceding block so that it can learn more from the rich contextual information. We hope that TAD can teach the network and make it pay more attention to the ROI area.

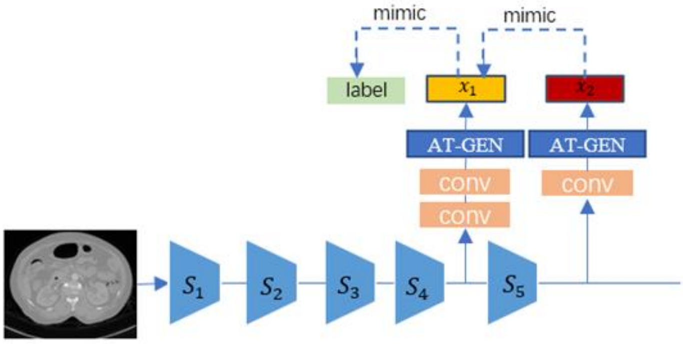

Fig. 4: An instantiation of using TAD. S1 ~ S5 comprise the stage of ResNeXt [16].

3.4. Loss Function

The total loss is comprised of two terms：

$$\mathrm{L} = \mathrm{L}_{cls} + \lambda \mathrm{L}_{TAD} \tag{1}$$

Here，the first term is classification loss of the main branch while the second term is the loss of the auxiliary training branch. The parameters λ balance the influence of classification losses and distillation loss on the final task.

The classification loss is the balanced cross entropy loss [22], which is a common method to solve the imbalance of the category. $p \in [0,1]$ is the estimated probability of the class with $y = 1$. For convenience, define p_t:

$$p_t = \begin{cases} p & if\ y = 1 \\ 1 - p & otherwise \end{cases} \tag{2}$$

The loss introduces the weighting factor $\alpha \in [0,1]$ for the class 1 and the weighting factor $1 - \alpha$ for the class -1. Similarly, we define α_t analogously to how we defined p_t. Finally, write the α-balanced CE loss as:

$$CE(p_t) = -\alpha_t \log(p_t) \tag{3}$$

In the training process, if the large-scale imbalance encountered overwhelms the cross entropy loss, negative values that are easy to classify account for most of the loss and dominate the gradient. Although α balances positive and negative samples, it cannot handle difficult-to-differentiate samples. It is recommended to add a modulation factor $(1 - p_t)^\gamma$ to the cross-entropy loss, and to adjust the focal parameter $\gamma \geq 0$. The focal loss is defined as:

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t) \tag{4}$$

The loss function in practice adopts the α-balanced variant of focal loss:

$$\mathrm{L}_{cls} = -\alpha_t (1 - p_t)^\gamma \log(p_t) \tag{5}$$

Based on experience, we find that the experimental effect is best when $\gamma = 2$. The loss of auxiliary training branch is defined as:

$$\mathrm{L}_{TAD} = \mathrm{L}_d(x_1, t) + \beta \mathrm{L}_d(x_1, x_2) \tag{6}$$

Where $\mathcal{L}_d$ is usually defined as $\mathcal{L}_2$ loss, β is a hyperparameter and *t* represent labelled feature map.

4. Experiment

Our experiment is designed to get better classification effect and validate whether the Improved-DPN and TAD is helpful for classification. In the following, we first introduce the dataset and evaluation metrics of the experimental data set. Regarding the experimental results, we compare our method with the different CNN networks and existing methods. In addition, we also show the effects of improved DPN and TAD with basic network.

4.1. Dataset

The CT image data set is provided by a cooperative scientific research project of a tertiary hospital in Shanghai, which contains gastric cancer diagnosis data from 2009 to 2019. We first eliminate sensitive patient information and screen the data. We get total 115 cases and diagnosis data, which contain cancer type, pathological diagnosis, CT sequence and other data. Among them, the status of tumor differentiation information can be obtained from the pathological diagnosis information. We acquire two labels of undifferentiated and differentiated information from pathological information.

The raw CT data and doctor's weakly supervised learning annotations can be seen by using the 3D Slicer software. The doctor roughly labels the corresponding tumor area with a cube through simple operations of the software. Fig. 5 shows the ROI (region of interest) information of two cases of gastric cancer marked by doctors. The left, middle, and right represent the horizontal plane, sagittal plane, and coronal plane of the CT image in the software.

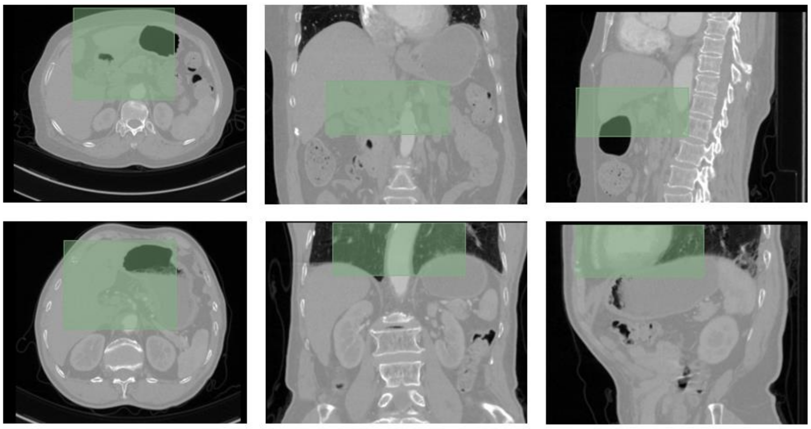

Fig. 5: ROI labelling of CT data of gastric cancer.

4.2. Data Preprocessing

We first analyse the CT data, and slice the CT image data of the horizontal plane according to the voxel interval to obtain 2D images. Counting the pixel values of each image, it is found that most of the pixel values are concentrated between -1024 and 500. Therefore, we regard the excess pixel value as a noise point and remove it. Finally, normalize the image to get our final data. Convert world coordinates to physical actual coordinates to get the labelled area, and crop to obtain the annotated images. We divide the effective data set into training set, validation set, and test set according to the ratio of 0.6, 0.2, 0.2. For the problem of class imbalance, we have expanded the data set, and the detailed information is shown in Table 1 and Table 2.

Table 1: Dataset information

Number	Train	Test	Validation
Case Number	70	23	22
Image Number	17920	5888	5632

Table 2: Training set information. 0 for undifferentiated and 1 for differentiated.

Label	Origin		Data augmentation	
	Case Number	Image Number	Case Number	Image Number
0	47	12032	47	12032
1	23	5888	47	12032
Total	70	17920	94	24064

4.3. Evaluation Metrics

We mainly use AUC, IMG_ACC and CASE_ACC as evaluation metrics. AUC is the area under the receiver operating characteristic curve (ROC), which is generally regarded as a standard method for evaluating the accuracy of a classification model. It avoids the assumption of subjectivity in the threshold selection process. When the scores obtained by continuous probability are converted into binary labels, by summing up the overall model performance, it measures the model’s performance in distinguishing positive and negative samples better than the threshold to judge. IMG_ACC represents the accuracy of prediction for all CT images. CASE_ACC represents the accuracy rate based on the case, that is, we predict all the CT images in a case, and finally use the voting principle to get the final prediction result of the case.

4.4. Results and Analysis

To verify the applicability of our proposed method, we used the test set to evaluate the model. We compared several traditional classification methods with our method. We used VGG-19 [23] and ResNeXt-50 to experiment on our dataset and got the results. In addition, we also compared the models with other existing cancer classification methods and used our data to obtain results in the models which were proposed by Tulasi Krishna Sajja et al [13] and Seokmin Han et al [14]. It can be seen from Table 3 that our proposed net has a better classification effect. It confirms that our work improves the existed methods for the problem of insufficient use of contextual information. At the same time, the introduction of annotation information can make the network pay more attention to the gastric tumour area so as to improve the effect of the model.

Table 3: Performance on different CNN networks

Methods	AUC
VGG-19	0.6417
ResNeXt-50	0.7015
Sajja et al [13]	0.7847
Han et al [14]	0.7937
Ours	**0.8135**

Table 4: Performance on test set

Methods	CASE_ACC (%)	IMG_ACC (%)	AUC
R-50-DPN	59.34	59.32	0.7467
R-50-Imp-DPN	60.73	59.62	0.7834
R-50-TAD	64.75	60.14	0.7635
R-50- Imp-DPN-TAD	**68.50**	**61.23**	**0.8135**

Finally, we conducted experiments to add improved DPN and TAD to prove their effectiveness. Table 4 show the results of the experiment. Here" R-50-DPN" denotes ResNeXt-50 + DPN and we use the same abbreviation in the following sections. The value of IMG_ACC is obtained when AUC achieves the best result. From the results in Table 4, our improved DPN structure has improved results compared to the DPN structure. AUC is increased from 0.7467 to 0.7834, which proves that the improved DPN is indeed beneficial to improve the effect of the classifier. In addition, the model with TAD has significantly increased the accuracy of the case, image accuracy and AUC, which has increased from 59.34% to 64.75%, 59.32% to 60.14% and 0.7467 to 0.7635 respectively. The model combined with improvement DPN and TAD achieve the best result in all evaluation metrics.

5. Conclusion

In this paper, we propose a model to classify the status of differentiation of gastric cancer. The proposed model uses improved DPN to add a self-attention feature map. Use TAD as an auxiliary training to make full use of the information marked by the doctor and encode rich contextual information. The experimental results show that the model has an improvement in the classification of gastric cancer differentiation. However, due to the three-dimensional characteristics of CT image data, it is easy to cause the loss of spatial feature information when we turn it into 2D images. In the future, we will consider using a 3D convolutional network for research.

6. Acknowledgments.

This research is funded by National Natural Science Foundation of China Youth Project (8180032537). The computation is supported by ECNU Multifunctional Platform for Innovation (001).

7. References

[1] National Cancer Report.National cancer report 2019[EB/OL]. *https://www.cn-healthcare.com/article/20190623/content-520594.html*, 2019-06-23/2020-03-31.

[2] Standards for Diagnosis and Treatment of Gastric Cancer (2018 Edition) [J]. *Chinese Journal of Digestive Diseases and Imaging (electronic version)*, 20! 9, 9 (03): 118-44.

[3] Hu S B, Liu C H, Wang X, et al. Pathological evaluation of neoadjuvant chemotherapy in advanced gastric cancer [J] *World J Surg Oncol*,19,17(1):3.

[4] Zhang Hejun, Jin Zhu, Cui Rongli, etc. Application of OLGA staging and grading evaluation system in histopathological evaluation of gastroscopic biopsy [J]. *Chinese Journal of Digestive Endoscopy*, 2014, 31(3): 121-125.

[5] Huang Yan. Analysis of the clinical value of preoperative gastroscopy biopsy in pathological diagnosis of gastric cancer[J]. *Forum on Primary Medicine 2020*, Volume 24, Issue 17, Page 2463-2465, 2020.

[6] A. G. Archer, D.C. Grand. Advances in radio diagnostics of primary and recurrent gastric cancer [J]. *Cancer Tratres*, 1991, 55(2):107-131.

[7] Luo Fucheng. Discussion on the classification of gastric cancer sonography [J]. *Chinese Journal of Ultrasound Medicine*. 1987, 3(1): 27.

[8] Zhang Yong, Liu Fang. The clinical application of serum gastric function detection in early gastric cancer screening[J]. *Journal of Clinical Laboratory Science (Electronic Edition)*. 2019, 8(04): 230-231.

[9] Liang Pan, Zhao Xitong, Zhao Huiping, et al. The value of CT in the diagnosis and clinical application of gastric cancer[J]. *Chinese Journal of Radiology 2020*, Volume 54 Issue 11, Page 1141-1144, ISTIC PKU CSCD, 2020.

[10] Sun J Y , Lee S W , Kang M C , et al. A Novel Gastric Ulcer Differentiation System Using Convolutional Neural Networks[C]// *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2018.

[11] Wang R, Sun H D, Zhang J L, Zhao Z J. A Transfer Learning Method for CT Image Classification of Pulmonary Nodules. *WISATS*, 2019, (2): 159-166.

[12] Zhu Z, Ding X, Zhang D , et al. Weakly-Supervised Balanced Attention Network for Gastric Pathology Image Localization and Classification[C]// *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020.

[13] Sajja T , Retz M , Kalluri H K . Lung Cancer Detection Based on CT Scan Images by Using Deep Transfer Learning[J]. *Traitement du Signal*, 2019, 36(4):339-344.

[14] Han S , Hwang S I , Lee H J . The Classification of Renal Cancer in 3-Phase CT Images Using a Deep Learning Method[J]. *Journal of Digital Imaging,* 2019, 32(4):638-643.

[15] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng. Dual path networks. arXiv preprint arXiv:1707.01629, 2017.

[16] Xie S , Girshick R , Dollar P , et al. [IEEE 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Honolulu, HI (2017.7.21-2017.7.26)] *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* - Aggregated Residual Transformations for Deep Neural Networks[J]. 2017:5987-5995.

[17] Jiang Dandan. The relationship between NEUROG3 expression and gastric cancer cell differentiation and tumor clinical staging [D]. *Chongqing Medical University*, 2020.

[18] Hu Yanshi, Pan Qing. The diagnostic value of Ang-2 and Vav1 protein combined with tumor marker detection in gastric cancer[J]. *Laboratory Medicine and Clinics*, 2021, 18(03): 333-335.

[19] Hou Y , Ma Z , Liu C , et al. Learning Lightweight Lane Detection CNNs by Self Attention Distillation[J]. 2019.

[20] S. Zagoruyko and N. Komodakis. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. *In International Conference on Learning Representations*

[21] Huang G , Liu Z , Laurens V D M , et al. Densely Connected Convolutional Networks[J]. 2016.

[22] Lin T Y , Goyal P , Girshick R , et al. Focal Loss for Dense Object Detection[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, PP(99):2999-3007..

[23] Simonyan K , Zisserman A . Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. *Computer Science*, 2014.