Multi-scale Deformable Transformer for the Classification of Gastric Glands: The IMGL Dataset

Panagiotis Barmpoutis[1,2], Jing Yuan[3], William Waddingham[2], Christopher Ross[2], Kayhanian Hamzeh[2], Tania Stathaki[3], Daniel C. Alexander[1], Marnix Jansen[2]

[1] Department of Computer Science, University College London, UK
[2] Department of Pathology, University College London, UK
p.barmpoutis@ucl.ac.uk
[3] Department of Electrical and Electronic Engineering, Imperial College London, UK

Abstract. Gastric cancer is one of the most common cancers and a leading cause of cancer-related death worldwide. Among the risk factors of gastric cancer, the gastric intestinal metaplasia (IM) has been found to increase the risk of gastric cancer and is considered as one of the precancerous lesions. Therefore, early detection of IM could allow risk stratification regarding the possibility of progression to cancer. To this end, accurate classification of gastric glands from the histological images plays an important role in the diagnostic confirmation of IM. To date, although many gland segmentation approaches have been proposed, no general model has been proposed for the identification of IM glands. Thus, in this paper, we propose a model for gastric glands' classification. More specifically, we propose a multi-scale deformable transformer-based network for glands' classification into normal and IM gastric glands. To evaluate the efficiency of the proposed methodology we created the IMGL dataset consisting of 1000 gland images, including both intestinal metaplasia and normal cases received from 20 Whole Slide Images (WSI). The results showed that the proposed approach achieves an F1 score equal to 0.94, showing great potential for the gastric glands' classification.

Keywords: Medical image classification, vision transformers, gastric cancer, intestinal metaplasia

1 Introduction

Gastric cancer is one of the most frequent causes of cancer-related deaths worldwide. As reported by the WHO in 2020 [1], it is the sixth most frequent type of cancer and it is the fourth leading cause of cancer-related deaths mainly due to its often-late stage of diagnosis [2]. The risk factors of gastric cancer include Helicobacter pylori infection, salt intake, tobacco smoking, alcohol consumption, family history of gastric cancer, gastric atrophy and intestinal metaplasia (IM) [2], [3]. More specifically, the IM of the mucosa of the stomach is a major precursor lesion that is associated with an increased risk of dysplasia and cancer [4], [5]. For this reason, early and effective diagnosis of IM is a crucial step to prevent gastric cancer. In the IM, the native gastric glands are replaced by metaplastic glands and gastric mucinous epithelial cells are replaced by goblet cells, enterocytes and colonocytes. Widely used diagnostic methods for IM include endoscopic and histological diagnosis. Endoscopic diagnosis of severe cases of IM is effortless, but there are difficulties in making the diagnosis of mild IM cases. Therefore, a biopsy of suspected cases of IM is suggested. Then, based on the Sydney protocol [6], IM is histologically diagnosed using hematoxylin and eosin (H&E) stain.

However, the visual assessment of glands by histopathologists is a laborious and time-consuming task [7]. Thus, the automated precise segmentation and classification of glands from the histological images plays an important role in the morphological analysis of glands, which is a crucial criterion for effective IM detection and management. Numerous methods have been proposed in literature for gland segmentation. However, to date, no generally applicable digital pathology approach has been proposed and applied for gastric glands' classification and more specifically for the identification and analysis of gastric intestinal metaplastic glands. Towards this end, in this paper, we propose a new methodology for gastric glands' classification based on H&E -stained images. More specifically, this paper makes the following contributions:

- We propose the IMGL-VTNet (Intestinal Metaplasia gastric GLands-Vision Transformer Net) that integrates a multi-scale deformable transformer model and a focal loss function for the gastric glands' classification.
- We publish the annotated IMGL dataset (Intestinal Metaplasia gastric GLands) that consists of normal and IM cases that we used for the training and testing of the proposed model. As a small number of research studies of gastric tissues use public data [8], we anticipate this dataset will provide the foundation for advanced studies of IM gastric glands and biopsies.

The rest of this paper is organized as follows: First, details of the proposed methodology are presented, followed by experimental results using the IMGL dataset. Finally, some conclusions are drawn and future extensions are discussed.

2 Related works

The digital medical image classification field receives growing attention and has become increasingly popular. Thus, various techniques and methods, based on either hand-crafted or deep learning features, have been developed for histopathological image classification tasks. Hand-crafted developed classification approaches for digital pathology tasks are based on grayscale density, color, texture and shape information [9], [10], [11]. After the extraction of low-level or mid-level set of features, post-processing methods such as dimensionality reduction and a classifier are usually used aiming to assign a classification label to each image [12]. On the other hand, more sophisticated classification methods such as deep-learning techniques [13] and higher-order dynamical systems [14], [15] have been developed aiming to address medical and histopathological image classification problems by extracting high-level features and knowledge directly from the data.

More recently, vision transformers inspired by the deep learning model that developed for the Natural Language Processing (NLP) [16] have been utilized for medical image segmentation [17], classification [18] and various computer vision tasks. Vision transformers apply attention mechanism to quantify pairwise long-range entity interactions [19]. These can be used as the form of self-attention layers or encoder-decoder pairs. More specifically, an adaptation of the BoTNet [19] has been proposed for image classification replacing the spatial convolutional layers with multi-head self-attention (MHSA) layers in the last stage of ResNet. In contrast, i-ViT [20] uses the transformer encoder to extract and aggregate features of instance patches for the papillary renal cell carcinoma subtyping task. The deformable DETR [21] is a fast-converging and memory-saving vision transformer with six encoder-decoder pairs, which facilitates high resolution feature maps from multiple scales. Owing to the efficiency, DT-MIL [22] applies it to high-level bag representation for multi-instance learning on histopathological images. Inspired by the deformable DETR, we propose a model that adopts a vision transformer in the glands' classification task aiming to exploit the local and global visual dependencies utilizing multi-scale deformable self-attention and a novel scale-aware feature extraction module.

3 Materials and methods

The framework of the proposed methodology for the gastric glands' classification into normal and IM cases is shown in Figure 1. Initially, the manually annotated IMGL dataset based on 20 WSI was created. Then, the segmented glands were fed to the proposed IMGL-VTNet for the classification of gastric glands.

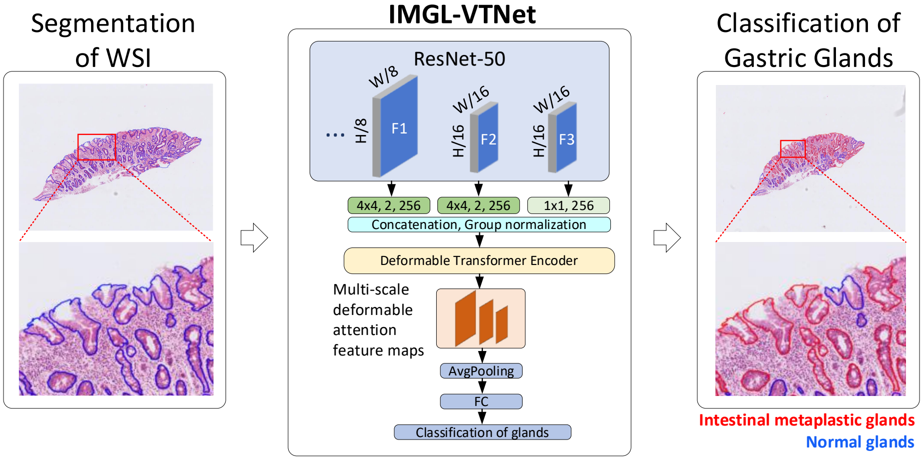
Segmentation of WSI
IMGL-VTNet
ResNet-50
W/8
H/8
F1
W/16
H/16
F2
W/16
H/16
F3
...
4x4, 2, 256
4x4, 2, 256
1x1, 256
Concatenation, Group normalization
Deformable Transformer Encoder
Multi-scale deformable attention feature maps
AvgPooling
FC
Classification of glands
Classification of Gastric Glands
Intestinal metaplastic glands
Normal glands

Fig. 1. The proposed methodology. The IMGL-VTNet takes the advantage of the deformable transformer encoder to extract multi-scale features.

3.1 IMGL dataset description

To evaluate the efficiency of the proposed methodology we created the IMGL dataset consisting of gastric glands (Figure 2). More specifically, the dataset includes 500 normal and 500 IM gastric glands. Gastric tissues were collected at University College London Hospital NHS trust, with ethical approval (research ethics committee (REC) reference: 15/YH/0311, & 19/LO/0089) with informed consent taken for prospective tissue collection. The tissues underwent routine H&E staining. For the evaluation of the IMGL-VTNet model we used five-fold cross validation selecting 800 gland images for the training and 200 images for the testing. It is worth mentioning that, as our aim is to develop a methodology for the early detection and diagnosis of IM to prevent gastric cancer, in this dataset we included mild and moderate IM cases. The dataset is available at the following link: 10.5281/zenodo.6908133.

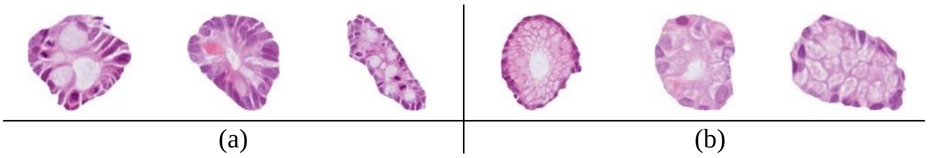

Fig. 2. Dataset images including (a) IM gastric glands and (b) normal gastric glands.

3.2 The proposed IMGL-VTNet architecture

The proposed model uses the ResNet-50 as the backbone, followed by the deformable transformer encoder-based feature extraction module. More specifically, in order to extract higher-level semantic information preserving the resolution, the stride and dilation of the last stage of the backbone are set as 1 and 2 respectively. Then, feature maps F_1 and F_2 were upsampled by two, while F_3 was encoded with a convolutional layer. Different kernel sizes were applied to each feature map as shown in Figure 1. Then, the multi-scale feature maps were concatenated, and group normalized and were fed into a deformable transformer encoder for the extraction of multi-scale features and the exploitation of local and global dependencies. Moreover, the extracted multi-level features were used, and an average pooling was considered followed by a fully connected layer for the classification of gastric glands into normal and IM.

To further enhance the model performance, a modulation term was applied to the binary Cross-Entropy loss function. The resulted focal loss [24] focuses on a set of hard examples improving the precision for these cases. More specifically, we defined the following loss function FL_i for the i-th image:

$$FL_i = w_{focal} \cdot Loss \tag{1}$$

$$w_{focal} = \begin{cases} (1-s)^{\gamma} & p = 1 \\ s^{\gamma} & p = 0 \end{cases} \tag{2}$$

$$Loss = p\log(s) + (1-p)\log(1-s) \tag{3}$$

where p is the ground truth (0 or 1) that represents the two categories (normal and IM), s is the predicted score and γ is the predesigned hyperparameter (we set $\gamma = 2$). It is worth mentioning that as the two categories of the IMGL dataset have the same number of training images, no additional balance was needed.

The input images were first resized and padded to the fixed shape of (224, 224). In addition, an augmentation method was utilized to further increase the variability of the training dataset and to avoid overfitting of the network. In particular, we included translation, rotation and flipping transformations. The Adam optimizer and mean teacher method [23] were used to get better and more robust performance. The network was trained on a single NVIDIA GeForce RTX 3090 GPU with batch size 16 for 80 epochs.

3.3 Multi-scale deformable transformer encoder

The deformable transformer encoder inputs three multi-scale feature maps with height H_l and width W_l $(l = 1, 2, 3)$. The input feature maps are first embedded with fixed positional encodings and level information to produce the query z_q. The query, input feature maps and reference points are fed into the Multi-Scale Deformable Attention Module (MSDAM) to extract the multi-scale deformable attention feature map. Then the deformable attention feature map is added to the input feature maps, followed by a Feed-Forward Network (FFN).

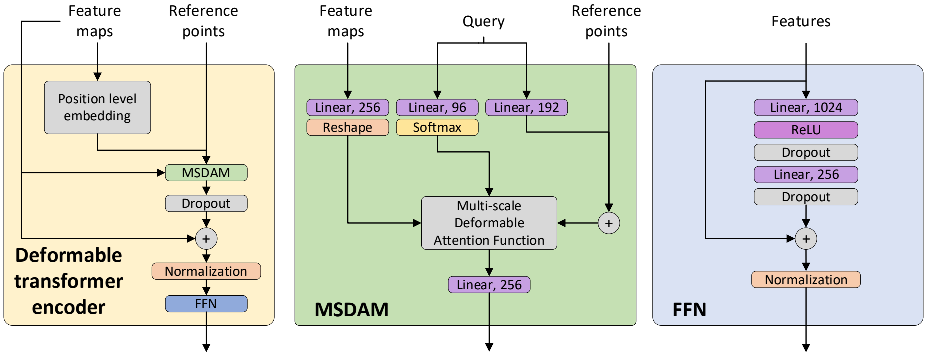

Fig. 3. Deformable transformer encoder consisting of a Multi-scale Deformable Attention Module (MSDAM) and a Feed-Forward Network (FFN).

In the MSDAM, value, weight and location tensors are first computed and applied to the multi-scale deformable attention function to produce the multi-scale deformable attention feature map z_o via weighted average. As shown in Figure 3, the value tensor v is produced by embedding the input features via a linear layer. The weight W and sampling offsets Δp are produced by embedding the query via two linear layers respectively. The weight is further normalized by a softmax operator along the scale and sampling point dimensions. The sampling location is the element-wise addition of sampling offset Δp and the reference points p. More specifically, the q-th element of the separate deformable attention feature $z' \in \mathbb{R}^{N_q \times c_v}$ $(N_q = \sum_{l=1}^{3} H_l W_l)$ at a single head is expressed as follows:

$$z'_q = \sum_p^{N_p} \sum_{l=1}^{3} W_{plhq} v_{p_{ql}+\Delta p_{qhlp}} \tag{4}$$

where q, h and p denote the elements of the deformable attention feature z_o, the attention head, and the sampling offsets respectively. W_{plhq} is an entity of $W \in \mathbb{R}^{N_q \times N_h \times 3 \times N_p}$. Furthermore, p_{ql} and Δp_{qhlp} denote the position of a reference point and one of the $N_p$corresponding sampling offset of $p \in \mathbb{R}^{N_q \times 3 \times 2}$ and $\Delta p \in \mathbb{R}^{N_q \times N_h \times 3 \times N_p \times 2}$ respectively. The number of sampling offsets and attention head are set as $N_p = 4$ and $N_h = 8$. The separate deformable attention features from 8 attention heads are projected to the q-th element of the overall output deformable attention feature z_o by a linear layer:

$$z_{o_q} = \sum_{h=1}^{N_h} W'_h z'_{qh} \tag{5}$$

where $W'_h \in \mathbb{R}^{c \times c_v}$ and vector $\boldsymbol{z}'_{qh} \in \mathbb{R}^{c_v}$ denote the learnable weight and the q-th separate deformable attention feature $\boldsymbol{z}'_q$ obtained at h-th attention head.

4 Experimental results

In this section, we present an evaluation analysis of the proposed gastric gland classification model as well as the efficiency of multi-scale feature maps for glands' classification. The goal of this experimental evaluation is threefold. Initially, we compared the efficiency of gastric glands' classification, using the IMGL dataset and widely used and state-of-the-art approaches. Secondly, we explored the efficiency of multi-scale deformable attention feature maps extracted from the deformable transformer encoder. Finally, to demonstrate the generality of our model, we applied the proposed method to the pedestrian detection task.

To evaluate the performance of the proposed model, we randomly partitioned the dataset into fivefold training and testing sets and we used precision, recall and F1-score.

4.1 A comparison of state-of-the-art methods: IMGL dataset

In this section, using the IMGL dataset we aim to present a comparison of the proposed methodology against a number of classification approaches. More specifically, in Table 1, we present the evaluation results of the IMGL-VTNet model in comparison to seven classification models. For the comparison, we consider the most widely used models including the state-of-the-art BoTNet-50 [19] architecture that achieves a strong performance on the ImageNet benchmark and has been applied on various tasks.

The results show that the proposed glands' classification approach achieves precision equal to 0.95 and recall equal to 0.94. Moreover, the proposed model achieves F1 score equal to 0.94. The proposed model achieves an F1 score improvement of 0.05 compared to the widely used ResNet-50. Furthermore, the integration of a Multi Head Self-Attention block in ResNet-50 improves the F1 score 0.02. Thus, the proposed model improves the F1 score by 0.03 compared to BotNet-50.

Further experimental results in 39 unannotated WSI (Figure 4) show that the IMGL-VTNet is robust under various cases. It is worth mentioning that normal cases (Figure 4a) include only normal glands, while IM WSI (Figure 4b) include both normal and IM glands. Thus, as it is shown in Figure 4a, only a very small number of glands are misclassified as IM glands. Further analyses of the unannotated normal cases show that less than 3% of the glands have been misclassified.

Table 1. A comparison of glands' classification using different models.

Method	Precision	Recall	F1 score
ResNet-18	0.92±0.04	0.84±0.03	0.88±0.03
ResNet-50	0.91±0.03	0.86±0.03	0.89±0.03
ResNet-101	0.91±0.03	0.82±0.03	0.86±0.03
VGG-19	0.89±0.03	0.89±0.02	0.88±0.02
Inception-V3	0.91±0.04	0.81±0.03	0.86±0.04
Xception	0.82±0.05	0.78±0.04	0.79±0.04
BotNet-50	0.92±0.03	0.90±0.02	0.91±0.02
IMGL-VTNet (proposed)	**0.95±0.03**	**0.94±0.02**	**0.94±0.03**

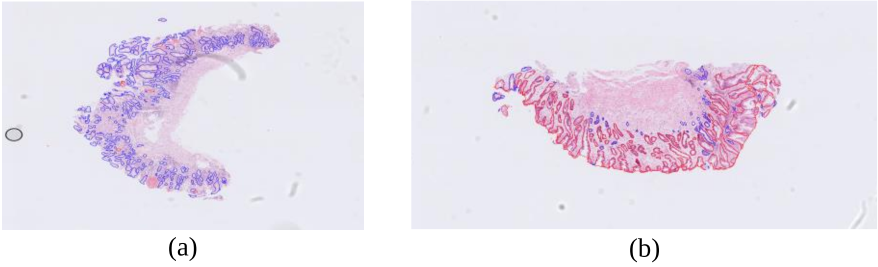
(a) (b)

Fig. 4. Glands' classification results of IMGL-VTNet model on two sample WSI: a) normal case, b) IM case. Blue color denotes the glands that have been detected as normal and red color denotes the glands that have been detected as IM glands.

4.2 Feature map scales analysis

Finally, we internally investigated the efficiency of multi-scale deformable attention feature maps for glands' classification. Thus, we compared the individual use of a single deformable attention feature map instead of the multiple deformable attention feature maps.

Table 2. A comparison of glands' classification efficiency using multi-scale deformable attention feature maps.

Feature map scale	Precision	Recall	F1 score
$W/16 \times H/16$	0.91±0.03	0.96±0.02	0.93±0.02
$W/8 \times H/8$	0.96±0.02	0.92±0.02	0.93±0.01
$W/4 \times H/4$	0.95±0.03	0.93±0.02	0.93±0.02
Multi-scale (IMGL-VTNet)	**0.95±0.03**	**0.94±0.02**	**0.94±0.03**

More specifically, the use of multi-scale feature maps slightly improves the F1-score by 0.01 (Table 2). The results show that higher-level features achieve better precision while lower-level features achieve better recall score.

4.3 Application of the proposed model to pedestrian detection

Finally, to demonstrate the generality of our model, we applied the VTNet to the pedestrian detection task. The average pooling and fully connected layers were replaced by two parallel branches predicting the confidence score and the corresponding bounding boxes respectively. For the evaluation, the Caltech pedestrian dataset was used [25], which contains approximately 2.5 hours of video. The performance was assessed in terms of log-average miss rate over false positives per image denoted as MR^{-2}. Based on the same training and testing protocol, the proposed VTNet outperforms other state-of-the-art pedestrian detectors by reducing the miss rate to 4.1% (Table 3).

Table 3. A comparison of the proposed architecture with five state-of-the-art detectors on the Caltech pedestrian dataset.

Method	MR^{-2} (%)
Faster R-CNN [26]	8.7
ALFNet [27]	8.1
RepLoss [28]	5.0
CSP [29]	4.5
Proposed	**4.1**

5 Conclusion

Multiple risk factors and a multistep process have been associated with gastric carcinogenesis. Among these factors, gastric IM of the mucosa has been recognized as a high-risk precancerous lesion for dysplasia and gastric cancer. However, as the manual assessment of biopsies by histopathologists based on the Sydney System is a laborious and time-consuming task, the accurate detection of IM gastric glands necessitates the adoption of artificial intelligence methods. Thus, in this paper we presented a methodology for the automated classification of gastric glands into normal and IM glands. The proposed IMGL-VTNet model for gastric glands' classification achieves an F1 score equal to 0.94. The results suggest that the proposed methodology obtains promising classification performance on the IMGL dataset. However, limitations of this study include the lack of an end-to-end gland segmentation and classification model that could be adopted on a widespread basis in routine histopathological practice.

Acknowledgements. The EPSRC and CRUK support this work through joint funding in grant number NS/A000069/1.We acknowledge the department of Pathology of UCL Cancer Institute and the UCL Centre for Medical Image Computing for their general support.

References

1. WHO, Cancerm. Available: https://www.who.int/news-room/fact-sheets/detail/cancer, [Last accessed: 24- July- 2022].
2. Waddingham W, Nieuwenburg SA, Carlson S, Rodriguez-Justo M, Spaander M, Kuipers EJ, Jansen M, Graham DG, Banks M.: Recent advances in the detection and management of early gastric cancer and its precursors. Frontline Gastroenterology. 2021 Jul 1;12(4):322-31.
3. Jencks DS, Adam JD, Borum ML, Koh JM, Stephen S, Doman DB.: Overview of current concepts in gastric intestinal metaplasia and gastric cancer. Gastroenterology & hepatology. 2018 Feb;14(2):92.
4. Busuttil RA, Boussioutas A.: Intestinal metaplasia: a premalignant lesion involved in gastric carcinogenesis. Journal of gastroenterology and hepatology. 2009 Feb;24(2):193-201.
5. Pellegrino C, Michele R, Chiara M, Alberto B, Florenzo M, Antonio N, Gioacchino L, Tiziana M, Gian LD, Francesco DM.: From Sidney to OLGA: an overview of atrophic gastritis. Acta Bio Medica: Atenei Parmensis. 2018;89(Suppl 8):93.
6. Dixon MF, Genta RM, Yardley JH, Correa P.: Classification and grading of gastritis: the updated Sydney system. The American journal of surgical pathology. 1996 Oct 1;20(10):1161-81.
7. Sirinukunwattana K, Pluim JP, Chen H, Qi X, Heng PA, Guo YB, Wang LY, Matuszewski BJ, Bruni E, Sanchez U, Böhm A.: Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis. 2017 Jan 1;35:489-502.
8. Gonçalves WG, Dos Santos MH, Lobato FM, Ribeiro-dos-Santos Â, de Araújo GS.: Deep learning in gastric tissue diseases: a systematic review. BMJ open gastroenterology. 2020 Mar 1;7(1):e000371.
9. Dimitropoulos K, Barmpoutis P, Koletsa T, Kostopoulos I, Grammalidis N.: Automated detection and classification of nuclei in pax5 and H&E-stained tissue sections of follicular lymphoma. Signal, Image and Video Processing. 2017 Jan 1;11(1):145-53.
10. Korkmaz SA, Binol H.: Classification of molecular structure images by using ANN, RF, LBP, HOG, and size reduction methods for early stomach cancer detection. Journal of Molecular Structure. 2018 Mar 15;1156:255-63.
11. Barmpoutis P, Kayhanian H, Waddingham W, Alexander DC, Jansen M.: Three-dimensional tumour microenvironment reconstruction and tumour-immune interactions' analysis. In: Proceedings of the IEEE DICTA 2021 (pp. 01-06).
12. England JR, Cheng PM.: Artificial intelligence for medical image analysis: a guide for authors and reviewers. American journal of roentgenology. 2019 Mar;212(3):513-9.
13. Barmpoutis P, Di Capite M, Kayhanian H, Waddingham W, Alexander DC, Jansen M, Kwong FN.: Tertiary lymphoid structures (TLS) identification and density assessment on H&E-stained digital slides of lung cancer. Plos one. 2021 Sep 23;16(9):e0256907.

14. Barmpoutis P, Dimitropoulos K, Apostolidis A, Grammalidis N.: Multi-lead ECG signal analysis for myocardial infarction detection and localization through the mapping of Grassmannian and Euclidean features into a common Hilbert space. Biomedical Signal Processing and Control. 2019 Jul 1;52:111-9.
15. Dimitropoulos K, Barmpoutis P, Zioga C, Kamas A, Patsiaoura K, Grammalidis N.: Grading of invasive breast carcinoma through Grassmannian VLAD encoding. PloS one. 2017 Sep 21;12(9):e0185110.
16. Devlin J, Chang MW, Lee K, Toutanova KB.: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. 2018 Oct 11.
17. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth HR, Xu D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF WACV 2022 (pp. 574-584).
18. Dai Y, Gao Y, Liu F.: Transmed: Transformers advance multi-modal medical image classification. Diagnostics. 2021 Aug;11(8):1384.
19. Srinivas A, Lin TY, Parmar N, Shlens J, Abbeel P, Vaswani A.: Bottleneck transformers for visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2021 (pp. 16519-16529).
20. Gao Z, Hong B, Zhang X, Li Y, Jia C, Wu J, Wang C, Meng D, Li C.: Instance-based vision transformer for subtyping of papillary renal cell carcinoma in histopathological image. In: Proceedings of the MICCAI 2021 (pp. 299-308).
21. Zhu X, Su W, Lu L, Li B, Wang X, Dai J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv:2010.04159. 2020 Oct 8.
22. Li H, Yang F, Zhao Y, Xing X, Zhang J, Gao M, Huang J, Wang L, Yao J.: DT-MIL: Deformable Transformer for Multi-instance Learning on Histopathological Image. In: Proceedings of the MICCAI 2021 (pp. 206-216).
23. Tarvainen A, Valpola H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems. 2017;30.
24. Lin TY, Goyal P, Girshick R, He K, Dollár P.: Focal loss for dense object detection. In: Proceedings of the IEEE ICCV 2017 (pp. 2980-2988).
25. Dollar P, Wojek C, Schiele B, Perona P.: Pedestrian detection: An evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence. 2011 Aug 4;34(4):743-61.
26. Ren S, He K, Girshick R, Sun J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems. 2015;28.
27. Liu W, Liao S, Hu W, Liang X, Chen X.: Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: Proceedings of the ECCV 2018 (pp. 618-634).
28. Wang X, Xiao T, Jiang Y, Shao S, Sun J, Shen C.: Repulsion loss: Detecting pedestrians in a crowd. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018 (pp. 7774-7783).
29. Liu W, Liao S, Ren W, Hu W, Yu Y.: High-level semantic feature detection: A new perspective for pedestrian detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 5187-5196).