DACTransNet: A Hybrid CNN-Transformer Network for Histopathological Image Classification of Pancreatic Cancer

Yongqing Kou[1], Cong Xia[2], Yiping Jiao[3(✉)], Daoqiang Zhang[1], and Rongjun Ge[4(✉)]

[1] Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
[2] Jiangsu Key Laboratory of Molecular and Functional Imaging, Department of Radiology, Zhongda Hospital, Medical School of Southeast University, Nanjing, China
[3] Nanjing University of Information Science and Technology, Nanjing, China
ping@nuist.edu.cn
[4] School of Instrument Science and Engineering, Southeast University, Nanjing, China
rongjun_ge@126.com

Abstract. Automated and accurate classification of histopathological images of pancreatic cancer can lead to higher survival rates for more pancreatic cancer patients in the clinic. However, there are very scarce existing studies for pancreatic cancer, and the diagnosis of pancreatic cancer remains a challenge for pathologists, especially for well-differentiated pancreatic cancer with a clinical histological pattern similar to that of chronic pancreatitis. We propose a hybrid CNN-Transformer model incorporating deformable atrous spatial pyramids (DACTransNet) to perform automated and accurate classification of histopathological images of pancreatic cancer. We elegantly integrate the powerful local feature extraction capability of CNN for spatial features and the global modeling capability of transformer for abstract patterns. Moreover, we imitate pathologists in the clinic by better integrating deformable convolution and multiscale methods to review histopathology slides in pyramidal format. In addition, a migration learning approach was used to improve the classification accuracy of pancreatic cancer histopathology images. The experimental results show that the proposed method not only has a high classification accuracy (up to 96%), but also its good robustness and generalizability as validated by real clinical datasets from multiple centers. Consequently, we provide an effective tool for the clinical diagnosis of pancreatic cancer.

Keywords: Pancreatic cancer · Histopathological image · Transformer

Y. Kou and C. Xia—Contribute equally to this work.

L. Fang et al. (Eds.): CICAI 2023, LNAI 14474, pp. 422–434, 2024.
https://doi.org/10.1007/978-981-99-9119-8_38

1 Introduction

Pancreatic cancer is a highly fatal malignant tumor, known as the "king of cancers", with a low 5-year survival rate of only 10% [1]. Clinically, many patients with pancreatic cancer are mistaken for pancreatitis when early symptoms appear and miss the most optimal time for treatment, resulting in a terrible prognosis. That is why pancreatic cancer has the highest mortality rate among all malignant tumors [2]. Thus, the accurate classification of pancreatic cancer plays an important role in the diagnosis and treatment process.

The gold standard for clinical medical diagnosis is the histopathological image evaluation by pathologists. Currently, there are fewer studies on automated analysis of histopathological images of pancreatic cancer. One reason may be due to the scarcity of resources and lack of high-quality annotation because of the low rate of early diagnosis. Another important reason may be that classification of pancreatic cancer is challenging because early stage pancreatic cancer is clinically very similar to pancreatitis. Most of the existing models used in studies of histopathological images are fine-tuned models [3–6] pre-trained on large natural image datasets (e.g., ImageNet datasets). However, due to the distinctiveness of histopathological images, such as the differences in data structure between histopathological images and natural images, as well as the heterogeneity of tumor cells, these models often result in suboptimal performance.

Recently, there has been significant progress in the accuracy of medical image analysis facilitated by methods based on Vision Transformers (ViT). Advanced approaches [7,8] for medical image analysis tasks rely on the ViT framework, leveraging its remarkable achievements in computer vision tasks. However, compared to methods based on convolutional neural networks, ViT-based models have certain limitations on their performance. Firstly, the serialization operation of ViT results in the loss of spatial information modeling. Secondly, ViT exhibits a higher dependency on large-scale datasets.

To address the aforementioned issues, we propose a novel and efficient hybrid network architecture called the CNN-Transformer hybrid model for Deformable Atrous Spatial Pyramids (DACTransNet). Which combines of local features of convolutional neural networks (CNN) and global features of ViT-based model. This model incorporates a lightweight transformer block at each layer of the CNN, allowing for the extraction of local features while considering global contextual information. Additionally, we employ a novel Deformable Atrous Spatial Pyramids (DC-ASPP) module to capture information from multi-scale irregular objects. Our method offers three primary contributions compared to existing approaches:

- We propose an integrated model that elegantly combines the local information of convolutional neural networks (CNN) and the long-range characteristics of Transformers, allowing for the simultaneous utilization of their strengths to enhance the model's ability to extract distinctive features from pancreatic cancer histopathological images.
- We incorporate deformable convolution into ASPP to extract multi-scale target information as well as irregular target information via multiple atrous

convolution layers with different scales of dilatation rate and deformable convolution in parallel with the features extracted by the encoder.

- We conducted extensive experiments on multiple central datasets, including training on the publicly available TCGA dataset annotated by multiple pathologists and testing on actual clinical datasets from three different regions. This comprehensive evaluation validated the generalization performance and clinical value of our model.

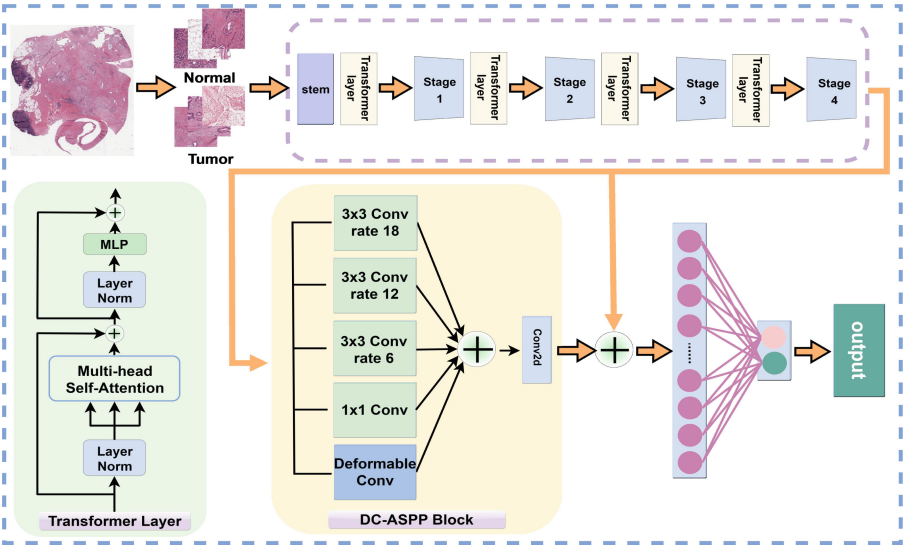

Fig. 1. Illustration of the DACTransNet. DACTransNet contains a CNN-Transformer backbone and a DC-ASPP block incorporating deformable convolution and atrous spatial pyramid pooling (ASPP).

2 Related Works

2.1 CNNs for Histopathological Images Classification

Convolutional Neural Networks (CNNs) have considerably contributed to the development of computational pathology with the development of CNN models due to their robust feature representation capabilities. Most histopathology image classification models [9–15] are derived from the prevalent natural image classification backbone. However, histopathological images are different from other medical images due to their inherent characteristics, such as extremely high resolution images, insufficient labeling, and multi-scale information, trinh et al. [16] developed a multi-scale binary-type coding network to enhance cancer classification by using binary pattern codes to capture and exploit patterns at different scales, which are further converted to decimal numbers. Zhang et al. [17] proposed the concept of a "virtual package" to classify histopathology whole

slide images through a multi-instance learning (MIL) architecture with a two-layer feature distillation. However, since resizing the original image causes information loss when processing high-resolution images, hou et al. [18] proposed a novel spatial hierarchical graph neural network framework to improve the classification accuracy of histopathological images by adding a dynamic structure learning to obtain the spatial topology and hierarchical dependencies of entities.

2.2 Vision Transformers for Histopathological Images Classification

More recently, transformers, originally proposed for natural language processing (NLP), have rapidly become the main architecture in computer vision [7,8,19] and they are considered as alternatives to their CNN counterparts. Several works have been proposed for the processing of medical images [20–22] because the self-attention mechanism of the visual transformer is able to directly capture long-range dependencies. However, due to the limited number of medical images, especially histopathology images, such methods are difficult to optimize and computationally expensive, so most existing studies have focused on creating hybrid CNN-transformer models for feature processing. Zhang et al. [23] proposed a multi-stage hybrid transformer combining the CNN and transformer, achieving high accuracy on the ROSE image dataset. Zheng et al. [24] proposed a graph transformer classifier that fuses graph neural networks and transformers to predict disease grades.

Discussion: Although the mentioned methods (CNNs-based as well as ViT-based methods) have achieved good results, they still ignore some inherent characteristics of histopathological images, such as the heterogeneity and heterogeneity of tumor cells, so they lead to challenging classification tasks of histopathological images. Therefore, the problem is how to elegantly combine the advantages of CNN and transformer, yet reduce the model complexity and classify well for targets with large shape differences and irregularities.

3 Methodology

For the features of histopathology images, the hybrid architecture of DACtransNet is designed to combine the robust local feature extraction capability of CNN for spatial features and the global modeling capability of transformer for abstract patterns. The overall architecture is shown in Fig. 1. DACtransNet consists of two main modules: a hybrid CNN-Transformer network as backbone and an ASPP module that incorporates deformable convolution. And then the ASPP module based on deformable convolution is designed to acquire multi-scale information, and the more robust deformable convolution is used to extract information from irregular targets. Our proposed network DACTransNet is optimized using the standard cross-entropy function as the loss function:

$$L = -[ylog\ \hat{y} + (1-y)log(1-\hat{y})] \tag{1}$$

where y is the true label value (positive class value is 1 and negative class value is 0) and $\hat{y}$ is the predicted label value ($\hat{y} \in (0,1)$).

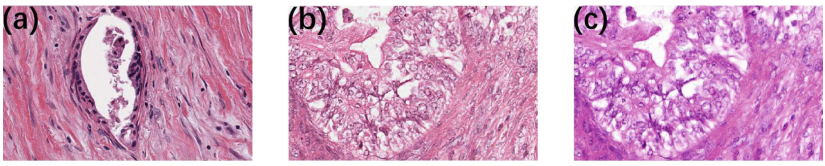

Fig. 2. Color normalization visualization, (a) is the image of the training dataset, (b) is the result of coloring and normalizing the images in the multicenter clinical dataset, and (c) is the original image in the multicenter clinical dataset, after coloring and normalizing the images in the test and training sets are closer in style.

3.1 CNN-Transformer Hybrid Network

We propose an integrated backbone that elegantly combines the local information of a convolutional neural network (CNN) and the long-range properties of transformer. It consists of convolutional and transformer blocks in an alternating superposition.

Convolution Block. To better encode spatial location information, we use convolutional blocks to extract local spatial features. First, fine-tuning of VGG19-Net pre-trained in ImageNet is designed by stacking a stem module and a CNN bottleneck module in four stages, where the blocks downsample the image $I \in R^{3\times H\times W}$ with edge size of H and W into abstractive features. The convolutional blocks in the four stages are the same as $Conv_2$, $Conv_3$, $Conv_4$, $Conv_5$. The modeling process for each convolutional block is shown as follows:

$$F_i^C = ConvBlock_i(F_{i-1}^T), i \in \{1, 2, 3, 4\} \tag{2}$$

where $F_i^C \in R^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i}$ is the local feature obtained from the $i-th$ block.

Transformer Block. The structure of the Transformer layer is shown in the lower left corner of Fig. 1, which contains a Multi-Headed Self-Attention (MHSA) layer to model long-range dependencies, two Layer Normalization layers, and a Multilayer Perceptron (MLP). Traditional ViT-based models use linear position projections for MHSA computation, which results in the loss of spatial information in the transformer, but this is crucial for medical image processing. Existing methods would alleviate this problem by adding positional encoding, however this would add additional computational cost and lead to poor optimization of the model. Therefore, inspired by [25], we replace the position-wise linear projection before each MHSA in the transformer module with a convolutional projection operation that employs $s \times s$ depth-separable convolution on a two-dimensional reshaped token mapping. Such an operation allows the model to further capture local information in the attention mechanism and can remove the original position embedding, simplifying the computational effort. The modeling process of the Transformer block is shown below:

Formally, the two-dimensional feature form of the stage i of the convolutional block output$F_i^C \in R^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i}$ is given as input, we learn a 2D convolution operation of kernel size $s \times s$, stride $s - o$ and p padding as a function $f(\cdot)$. The tokens maps $f(x_i)$ obtained after convolutional mapping. Next, the convolutional projection is implemented using a deeply separable convolutional layer with kernel size s. Finally, the projected tokens are panned to one dimension, i.e., $Query\ Q$, $Key\ K$, $Value\ V$ is used as the input for multi-headed self-attention. The modeling process of this process can be formulated as:

$$x_i^{q/k/v} = Flatten(DepthConv2d(Reshape(x_i), s)) \tag{3}$$

where $x_i^{q/k/v}$ is the token input for $Q/K/V$ matrics at layer i, $DepthConv2d$ is a deep-wise separable convolution, x_i is the token prior to the convolutional projection, and s refers to the convolution kernel size. Hereinafter, applying a MHSA, the output is obtained as follows:

$$SA_i = \sigma(\frac{Q \times K^T}{\sqrt{d}})V \tag{4}$$

where $\sigma$is the softmax function, d is the dimension of the input token. After applying a residual operation and MLP, this process can be expressed as follows:

$$F_i^T = MLP(SA_i + F_i^C) + SA_i \tag{5}$$

We omitted the layer normalization (LN) in the equation for simplicity. Finally, the output of $Transformer\ Block$ is as the input of $Convolutional\ Block$ at stage $i + 1$.

3.2 DC-ASPP Block

To combine information at multiple scales, we introduce an *astrous spatial pyramid pooling* (ASPP) block to detect incoming features by using various filters or pooling operations under multiple perceptual fields and multiple dilation, although using astrous convolution can obtain larger perceptual fields without increasing the computation. Since deformable convolution [26] focuses on adding adaptive 2D spatial offsets to enhance the flexibility of the convolutional sampling locations and to keep the channel dimensions unchanged, we propose a new feature aggregation module that, by introducing deformable convolution instead of normal convolution, the sampling locations of the convolution are no longer limited to fixed sampling locations, making the sampling locations more flexible and capable of producing more accurate localizations. Thus our proposed $DC - ASPP\ Block$ enables the network to focus on both overall features and detailed features and obtain more detailed localization for better feature extraction capability. Our improved $DC - ASPP\ Block$ consists of one 1×1 convolution, three 3×3 convolutions with $rates = (6, 12, 18)$ (all with 256 filters and Batch Normalization) and $Deformable\ Convolution$, and then the features of all branches are then concatenated and pass through another 1×1 convolution.

Table 1. Performance comparison for previous and DACTransNet models on TCGA datasets. Boldfaced results indicate better results.

Method	Accuracy ↑	AUC ↑	F1 Score ↑	Precision ↑		Recall ↑	
				Normal	Tumor	Normal	Tumor
VGGNet	0.8963	0.9765	0.8381	0.9802	0.7332	0.8663	0.9584
ResNet	0.9451	0.9740	0.8919	0.9734	0.9742	0.9921	0.8224
DenseNet	0.9207	0.9785	0.8439	0.9212	0.9216	0.9754	0.7782
ViT	0.8232	0.9451	0.7293	0.9412	0.6293	0.8073	0.8672
SwinTransformer	0.8719	0.9625	0.7836	0.9381	0.7312	0.8823	0.8442
MobileViT	0.8537	0.9387	0.7545	0.9201	0.7061	0.8742	0.8100
DACTransNet (Ours)	**0.9634**	**0.9894**	**0.9791**	**0.9821**	**1.000**	**1.000**	**0.9591**

Table 2. Performance comparison for previous and DACTransNet models on center A datasets. Boldfaced results indicate better results.

Method	Accuracy ↑	AUC ↑	F1 Score ↑	Precision ↑		Recall ↑	
				Normal	Tumor	Normal	Tumor
VGGNet	0.8797	0.9493	0.8692	0.9031	0.8535	0.8764	0.8854
ResNet	0.8822	0.9530	0.8664	0.8822	0.8853	0.9120	0.8482
DenseNet	0.8940	0.9539	0.8758	0.8514	0.8643	0.8833	0.8876
ViT	0.8240	0.9121	0.8166	0.8792	0.7723	0.7893	0.8662
SwinTransformer	0.8541	0.9381	0.8437	0.8852	0.8234	0.8452	0.8651
MobileViT	0.8499	0.9296	0.8372	0.8752	0.8216	0.8482	0.8533
DACTransNet (Ours)	**0.8973**	**0.9714**	**0.8831**	**0.8731**	**0.9344**	**0.9522**	**0.8933**

4 Experiment

4.1 Datasets and Details

Public Datasets (TCGA). In this study, we utilized the TCGA (The Cancer Genome Atlas) dataset [27]. Due to the limited availability of pancreatic cancer histopathological image resources, we opted to use H&E stained tissue slides from 190 pancreatic cancer patients from the TCGA dataset as the sole training dataset for our experiments. The entire image is magnified up to a resolution of 160k×160k pixels at 40× zoom. To facilitate computation and achieve better classification accuracy, we downscaled the images to a 4x zoom level. Moreover, because the whole slide images (WSI) are too large to be loaded into memory, and because the TCGA dataset didn't contain annotations due to the difficulty of annotating pancreatic cancer histopathology images, a group of pathologists collaborated to select annotated portions of the pancreatic cancer WSIs containing both tumor and normal tissue, while maintaining a balanced ratio of negative and positive samples. To facilitate training, we cropped each WSIs into non-overlapping patches of 256×256 pixels. 1336 patches were used as the training dataset, and 164 patches were used as the test dataset.

Table 3. Performance comparison for previous and DACTransNet models on center B datasets. Boldfaced results indicate better results.

Method	Accuracy ↑	AUC ↑	F1 Score ↑	precision ↑		Recall↑	
				Normal	Tumor	Normal	Tumor
VGGNet	0.8426	0.9171	0.8443	0.8441	0.8415	0.8382	0.8472
ResNet	0.8594	0.9362	0.8585	0.8413	0.8830	0.8851	0.8354
DenseNet	0.8605	0.9465	0.8534	0.8222	0.9092	0.9184	0.8041
ViT	0.7945	0.8738	0.7963	0.7923	0.7971	0.7942	0.7956
SwinTransformer	0.8417	0.9250	0.8407	0.8312	0.8524	0.8556	0.8293
MobileViT	0.7814	0.9631	0.9171	0.8592	0.9423	0.9511	0.8933
DACTransNet (Ours)	**0.8714**	**0.9631**	**0.9171**	**0.8592**	**0.9243**	**0.9511**	**0.8933**

Table 4. Performance comparison for previous and DACTransNet models on center C datasets. Boldfaced results indicate better results.

Method	Accuracy ↑	AUC ↑	F1 Score ↑	Precision ↑		Recall↑	
				Normal	Tumor	Normal	Tumor
VGGNet	0.8655	0.9357	0.8660	0.8661	0.8652	0.8644	0.8672
ResNet	0.9087	0.9738	0.8984	0.8834	0.9154	0.9141	0.8821
DenseNet	0.9039	0.9630	0.9013	0.8793	0.9324	0.9365	0.8722
ViT	0.8444	0.9264	0.8447	0.8412	0.8482	0.8483	0.8412
Swin Transformer	0.8544	0.9536	0.8727	0.8563	0.8920	0.8534	0.8543
MobileViT	0.8522	0.9497	0.8538	0.8632	0.8543	0.8432	0.8543
DACTransNet (Ours)	**0.9113**	**0.9801**	**0.9091**	**0.8881**	**0.9374**	**0.9432**	**0.8824**

Multicenter Clinical Dataset for External Validation. In order to evaluate the generalization performance of our proposed model, we applied it to clinical datasets from three different centers, which comprehensively encompassed different types of pancreatic cancer. To ensure patient privacy, the clinical datasets from the three centers were anonymized as Center 1, Center 2, and Center 3, consisting of 30, 35, and 38 H&E stained histopathological slides, respectively. Since the staining of data from different centers varies widely, color normalization becomes an essential step in preprocessing. We adopted the method proposed by Jiao et al. [28], the results after color normalization are shown in Fig2.

Implementation Details. We use Pytorch and the adam optimizer with a learning rate of 1e-4 to run all our experiments. We used the pre-trained weights of VGG19-Net pre-trained on imagenet to train our proposed DACTransNet and ran 300 epochs. To avoid overfitting, our data were enhanced as follows: rotation (90°), horizontal, vertical flip and color disturbance. For the TCGA training dataset, we set the batch size to 4. Appropriate test values, including recall, precision, F1-score, accuracy, and AUC are calculated to quantify and compare the model performance of these four test cohorts.

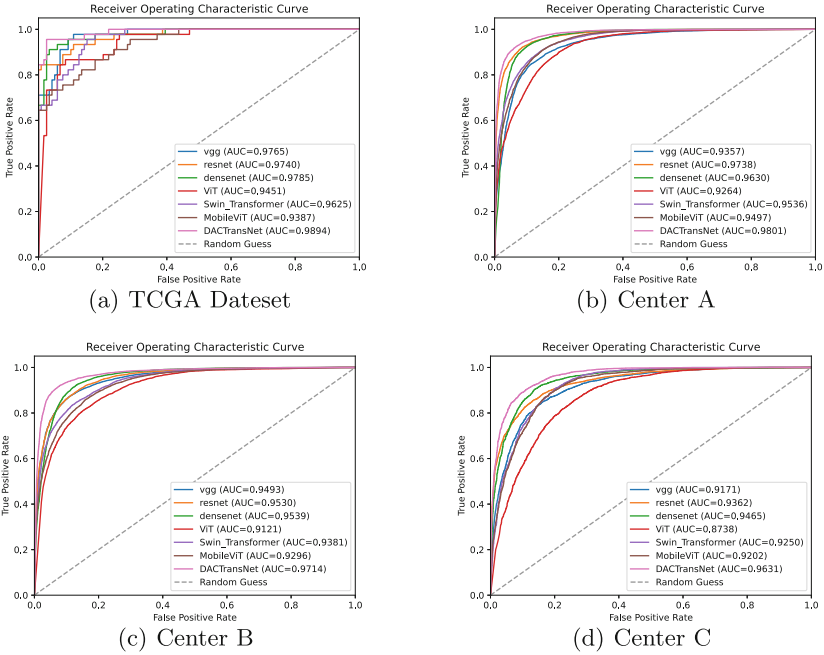

Fig. 3. Acceptance operating characteristic (ROC) curves.

Table 5. Ablation study on TCGA dataset, Trans denotes transformer layer, DC denotes deformable convolution. The baseline model we used was VGG19Net.

Models	Baseline	Trans	ASPP	DC	Accuracy ↑	F1 Score ↑	Precision ↑	Recall ↑
E.1	✓				0.8963	0.8308	0.7332	0.9584
E.2		✓			0.8444	0.8447	0.8482	0.8412
E.3	✓	✓			0.9123	0.9116	0.8829	0.9422
E.4	✓	✓	✓		0.9328	0.9521	0.9533	0.9511
E.5	✓	✓	✓	✓	0.9634	0.9891	1.0000	0.9591

4.2 Comparison with State-of-the-Art Methods

To demonstrate the effectiveness of our proposed method, we compare our approach with state-of-the-art classification methods, including transformer-based models and CNN-based models on the ImageNet dataset.

Results on TCGA Dataset. Compared to these models, our approach largely outperforms both the pre-trained transformer-based model and the CNN-based model. More specifically, our DACTransNet achieves 96.34% accuracy. From the results in Table 1, we found that DACTransNet has a relatively significant advantage in cancer recall compared to the transformer-based and CNN-based

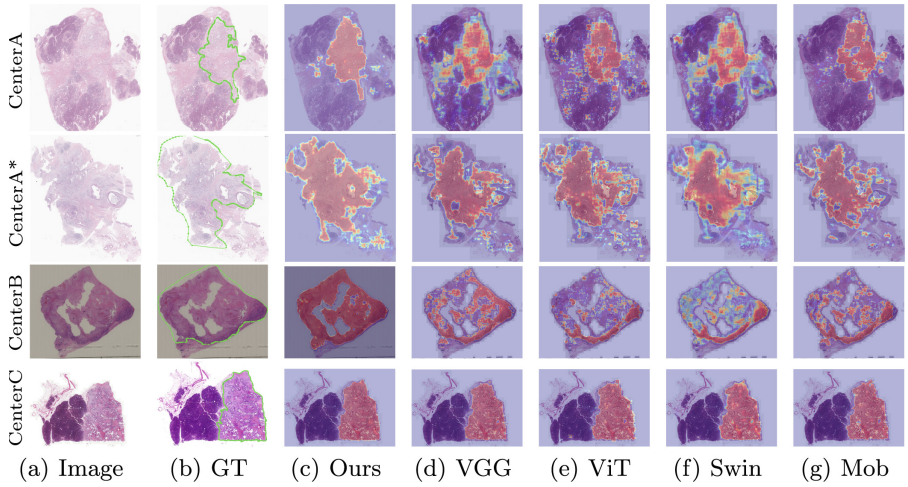

Fig. 4. Visualization results on an multicenter clinical dataset. Where DAT stands for our proposed DACTransNet, Swin stands for Swin Transformer, and Mob stands for MobileViT. The greater the color indicates a higher probability of lesions and conversely a higher probability of normal tissue.

models pre-trained on the ImageNet dataset, a performance that is consistent with the requirements of clinical diagnosis, as pathologists must scrutinize and not overlook any patches that may be cancerous.

Results on Multicenter Clinical Datasets. Then, in order to verify the portability and robustness of the model, we performed the same tests on multicenter clinical datasets. The results are shown in Tables 2, 3 and 4. It can be seen from the results that our model also achieves better results on the multicenter clinical dataset, which is sufficient to prove the good generalization performance of our model.

Visualization Results. In addition, we analyze the performance of our proposed DACTransNet model by means of receiver operating characteristic curve (ROC) curves, as shown in Fig. 3. In the results, our model achieves better performance both on the internal training dataset and the multicenter clinical dataset. Finally, we plotted cancer probability heatmaps on all four datasets, as shown in Fig. 4 and what can be seen is that for the slice shown in CenterA, our model DACTransNet can achieve a classification accuracy of 98.32%, while the classification accuracy of VGG-Net is only 89.45%. For more challenging cases, such as CenterA* and CenterB slices, which are very difficult to classify, because these two slices show a very rare type of cancer in pancreatic cancer, our model can achieve a classification accuracy of 64.42%, while VGG-Net only has a classification accuracy of 56.23%. Finally, for the more common type of pancreatic

cancer histopathology slides in CenterC, VGG-Net has a classification accuracy of about 98.67%, while our model can achieve 99.45% classification accuracy.

4.3 Ablation Studies

We conducted a series of ablation studies on the TCGA dataset to investigate the effectiveness of DACTransNet and to justify the design choices, with the baseline method being VGG19Net with pre-trained weights, as shown in Table 5.

CNN-Transformer Hybrid Network Backbone. Comparing E.3 with E.1 and E.2, we can see that the hybrid CNN-Transformer approach is significantly better than the pure CNN approach and pure ViT-based approach, which shows that CNN and Transformer can indeed make it possible to efficiently process spatial local information and global background information in a unified block.

DC-ASPP Block. Comparing E.3 and E.4, we can see that the ASPP module can bring better performance, which indicates that the multi-scale approach is important for histopathology image classification since pathologists read films are operating at multiple resolutions. And by for E.5 and E.4, we can see that deformable convolution can also bring better performance because of the heterogeneity of tumor cells, and by adding deformable convolution can be a good learning effect for irregular cancer types.

5 Conclusion

In this work, to address some of the challenging classification tasks for histopathological images. We propose a DACTransNet network for pancreatic cancer classification, which elegantly combines the advantages of CNN and transformer to improve the model's ability to model local information and long-distance dependencies, and to classify targets with large differences in shape and irregularities well. It outperforms pure CNN methods pre-trained on ImageNet or pure Transformer methods, and can show better performance on small datasets.

Acknowledgements. This study was supported by the Natural Science Foundation of Jiangsu Province (No. BK20210291), the National Natural Science Foundation (No. 62101249 and No. 62136004), and the China Postdoctoral Science Foundation (No. 2021TQ0149 and No. 2022M721611).

References

1. Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. CA: Cancer J. Clinicians **73**(1), 17–48 (2023)
2. Pereira, S.P., et al.: Early detection of pancreatic cancer. Lancet. Gastroenterol. Hepatol. (2020)

3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
4. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556 (2014)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2015)
6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269 (2017)
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv, abs/2010.11929 (2020)
8. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992–10002 (2021)
9. Chen, H., Dou, Q., Wang, X., Qin, J., Heng, P.-A.: Mitosis detection in breast cancer histology images via deep cascaded networks. In: AAAI Conference on Artificial Intelligence (2016)
10. Tian, Y., et al.: Computer-aided detection of squamous carcinoma of the cervix in whole slide images. arXiv, abs/1905.10959 (2019)
11. Fu, H., et al.: Automatic pancreatic ductal adenocarcinoma detection in whole slide images using deep convolutional neural networks. Front. Oncol. **11** (2021)
12. Yang, H., et al.: Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study. BMC Med. **19** (2021)
13. Coudray, N., et al.: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat. Med. **24**, 1559–1567 (2018)
14. Ianni, J.D., et al.: Tailored for real-world: a whole slide image classification system validated on uncurated multi-site data emulating the prospective pathology workload. Sci. Rep. **10** (2020)
15. Liu, M., Lanlan, H., Tang, Y., Chu Wang, Yu., He, C.Z., et al.: A deep learning method for breast cancer classification in the pathology images. IEEE J. Biomed. Health Inform. **26**, 5025–5032 (2022)
16. Vuong, T.T.L., Song, B., Kim, K., Cho, Y.M., Kwak, J.T.: Multi-scale binary pattern encoding network for cancer classification in pathology images. IEEE J. Biomed. Health Inform. **26**, 1152–1163 (2021)
17. Zhang, H., et al.: DTFD-mil: double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18780–18790 (2022)
18. Hou, W., Huang, H., Peng, Q., Yu, R., Yu, L., Wang, L.: Spatial-hierarchical graph neural network with dynamic structure learning for histological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2022)
19. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., J'egou, H.: Training data-efficient image transformers & distillation through attention. arXiv, abs/2012.12877 (2020)
20. Chen, H., et al.: Gashis-transformer: a multi-scale visual transformer approach for gastric histopathological image detection. Pattern Recogn. **130**, 108827 (2021)

21. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., et al.: TransMil: transformer based correlated multiple instance learning for whole slide image classication. In: Neural Information Processing Systems (2021)
22. Xiong, Y., et al.: Nyströmformer: a nyström-based algorithm for approximating self-attention. In: AAAI Conference on Artificial Intelligence, vol. 35, pp. 16:14138–16:14148 (2021)
23. Zhang, T., Yunlu Feng, Yu., Zhao, G.F., Yang, A., Lyu, S., et al.: MSHT: multi-stage hybrid transformer for the rose image analysis of pancreatic cancer. IEEE J. Biomed. Health Inform. **27**, 1946–1957 (2021)
24. Zheng, Y., Gindra, R., Green, E., Burks, E.J., Betke, M., Beane, J.E., et al.: A graph-transformer for whole slide image classification. IEEE Trans. Med. Imaging **41**, 3003–3015 (2022)
25. Wu, H., et al.: CVT: introducing convolutions to vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 22–31 (2021)
26. Dai, J., et al.: Deformable convolutional networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 764–773 (2017)
27. The cancer genome atlas (TCGA) (2016). http://cancergenome.nih.gov/
28. Jiao, Y., Li, J., Fei, S.M.: Staining condition visualization in digital histopathological whole-slide images. Multimedia Tools Appl. **81**, 17831–17847 (2022)