Hierarchical Pretrained Backbone Vision Transformer for Image Classification in Histopathology

Luca Zedda, Andrea Loddo(✉), and Cecilia Di Ruberto

Department of Mathematics and Computer Science, University of Cagliari, via Ospedale 72, 09124 Cagliari, Italy
{luca.zedda,andrea.loddo,dirubert}@unica.it

Abstract. Histopathology plays a crucial role in clinical diagnosis, treatment planning, and research by enabling the examination of diseases in tissues and organs. However, the manual analysis of histopathological images is time-consuming and labor-intensive, requiring expert pathologists. To address this issue, this work proposes a novel architecture called Hierarchical Pretrained Backbone Vision Transformer for automated histopathological image classification, a critical tool in clinical diagnosis, treatment planning, and research. Current deep learning-based methods for image classification require a large amount of labeled data and significant computational resources to be trained effectively. By leveraging pretrained Visual Transformer backbones, our approach can classify histopathology images, achieve state-of-the-art performance, and take advantage of the pretrained backbones' weights. We evaluated it on the Chaoyang histopathology dataset, comparing it with other state-of-the-art Visual Transformers. The experimental results demonstrate that the proposed architecture outperforms the others, indicating its potential to be an effective tool for histopathology image classification.

Keywords: Histopathology · Deep Learning · Hierarchical ViT

1 Introduction

Histopathology is a critical tool in clinical diagnosis, treatment planning, and research, enabling the study of disease in tissues and organs. Histopathological images provide a wealth of information about the morphology and cellular structure of tissues and organs and can reveal important insights into the underlying pathophysiology of diseases. However, analyzing histopathological images is time-consuming and labor-intensive, requiring expert pathologists to examine and classify the images based on their visual features.

Image classification is a computer-aided technique that can automate the analysis of histopathological images by automatically identifying regions of interest and classifying them into different categories. It helps pathologists by reducing their workload, improving accuracy, and enabling large-scale disease studies.

G. L. Foresti et al. (Eds.): ICIAP 2023, LNCS 14234, pp. 223–234, 2023.
https://doi.org/10.1007/978-3-031-43153-1_19

In this context, deep learning (DL)-based methods have achieved remarkable performance also because they can automatically learn and extract features from raw data, even identifying features that are not easily discernible to humans. Although their advantages, the black-box nature of DL models causes pathologists to hesitate when adopting them in high-stakes environments. In order to comply with regulations and facilitate a feedback loop that integrates model diagnosis and refinement in the development process, there is an increasing need for explainable deep learning [18].

This work proposes a novel DL architecture called Hierarchical Pretrained Backbone Vision Transformer (HPB-ViT) for image classification. The HPB-ViT architecture is based on the Vision Transformer (ViT) model [3], which has shown state-of-the-art performance in computer vision (CV) tasks [3,12]. Nevertheless, they require large amounts of labeled data to train effectively. This aspect is even more pronounced in medical imaging because obtaining labeled data can be challenging due to the need for expert annotation and the potential for variability in annotations between different experts.

Pretraining has become a popular technique in DL to address this challenge. Models are first trained on large-scale datasets, i.e., ImageNet [2] or COCO [7] before being fine-tuned on the target dataset. Pretraining improves the performance of DL models by enabling them to learn generalizable features that can be then transferred to different tasks and datasets [6].

Although pretraining ViT on large-scale datasets is effective, it can be time-consuming and computationally expensive, which hinders their practical use in real-world applications. To tackle this challenge, the HPB-ViT architecture integrates various pretrained off-the-shelf ViT backbones, resulting in faster training with better performance. By leveraging the capability of pretrained ViT backbones, HPB-ViT can effectively learn to classify histopathology images with smaller amounts of labeled data while still achieving state-of-the-art performance.

We evaluated the performance of HPB-ViT on the Chaoyang histopathology dataset [21] and compared it with other state-of-the-art models. The experimental results showed that HPB-ViT achieved superior performance, demonstrating the effectiveness of our proposed architecture for the task at hand.

The rest of this work is organized as follows. Firstly, we review related works in Sect. 2, specifically addressing histopathology and its current issues. Next, we present our proposed approach in Sect. 3. We then report and discuss the experimental results in Sect. 4. Finally, the conclusions are drawn in Sect. 5.

2 Background Concepts and Related Work

When analyzing whole-slide digital pathology images, several challenges need to be addressed. These images are extremely large and are measured in terms of gigapixels, which makes it necessary to break them down into smaller tiles to be processed effectively. Additionally, different magnifications are required for specific tasks and to combine information from multiple scales. Predicting survival

can be challenging since there may only be weak slide-level labels available, and the most crucial areas of the image may not always be obvious. Annotations can also be complex due to the variability in disease subtypes, which requires the expertise of highly trained pathologists. Cell-based methods involve detecting and characterizing thousands of objects, which can pose a challenge. DL architectures have become increasingly adapted to this task to tackle these issues, and new approaches specifically designed for digital pathology are emerging [11]. These approaches have replaced traditional handcrafted methods [10].

Deep Learning in Histopathology. CNNs have been the reference approach since the release of AlexNet. However, several promising architectures, such as T2T-ViT [17], Swin [8,9], DeepViT [19], CvT [15] have emerged after the introduction of Transformer architecture [14] and the advent of ViT [3].

One of the earliest studies that employed ViT for histopathology was conducted by Zhou et al. [20]. They proposed a hybrid model that combined convolutional residual networks [5] and ViT mechanisms to create a network with inductive solid bias capabilities and scale and rotation invariance, obtaining state-of-the-art performance in a brain biopsies dataset.

Building upon this, Chen et al. [1] also proposed a hybrid convolutional and vision transformer network similar to inception networks [13]. They achieved state-of-the-art results on the HE-GHI-DS dataset, further highlighting the effectiveness of the hybrid approach in histopathology classification.

Issues in Deep Learning for Histopathology. Although the promising results indicate that ViTs have the potential to revolutionize the field of histopathology, some limitations and challenges must be considered. First, the typical extensive size of these images can lead to high computational requirements. Some studies addressed this problem by dividing the original high-resolution images into smaller patches [11]. However, selecting the appropriate magnification level requires a profound understanding of the analyzed task [4].

Also, there is a shortage of adequately labeled images for training. To tackle this problem, Xu et al. [16] developed a generative approach that employs ViTs to generate synthetic data for training. The results were promising, particularly for the underrepresented minority classes, with ≈5% performance gain.

3 The Proposed Architecture

In this section, we present the proposed HPB-ViT architecture and its modules. Section 3.1 describes the concept of attention in Transformers [14]. In Sect. 3.2, we describe the approach to standardize the backbone's output. Section 3.3 discusses the module which learns patch pooling. Finally, in Sect. 3.4, we provide a detailed description of the overall architecture.

3.1 Attention Mechanism

Attention mechanisms have become increasingly popular due to the advancements in transformer architecture and multi-head self-attention (MHSA) [14].

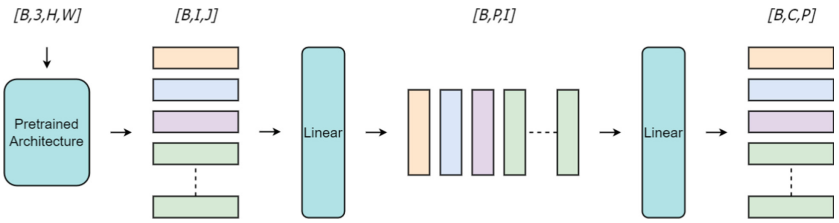


Fig. 1. From left to right: example feature map derived from the backbone pretrained Architecture, Liner transformation, second linear transformation after feature transpose.

The main goal of attention mechanisms is to allow models to focus on important parts of the input while disregarding irrelevant information.

In the MHSA formulation, the attention mechanism is defined as:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \tag{1}$$

where Q, K, and V represent the query, key, and value matrices. The W_i^Q, W_i^K, and W_i^V matrices represent the learned projection matrices for the i^{th} head, and W^O represents the learned output projection matrix. h is the number of heads, and d_k is the dimensionality of the key vectors.

The attention mechanism calculates a weighted sum of their values to prioritize important input elements and ignore irrelevant ones. The weights are determined by comparing the queries and keys, which creates a soft alignment between the query and key vectors. This then allows the corresponding value vectors to be weighted accordingly. Our proposed architecture employs the attention mechanism in the *Transformer layer* shown in Fig. 3.

3.2 Backbone Encapsulator

The Backbone Encapsulator (BE) module receives an input batch of images of size $[B, 3, H, W]$, where B is the batch size, 3 is the number of channels that forms the images, while H and W are height and width, respectively. Then, BE passes the images through a selected pre-trained backbone, obtaining features of size $[B, I, J]$, where I is the number of channels of the output feature map provided by the backbone, and J is the backbone's feature map size. Then, BE applies two different linear transformations to the backbone output features. The first linear transformation expands the last dimension of the features to the predefined size. Then, the features are transposed and passed through a second linear layer to expand the number of patches to the predefined size. This transformation serves the purpose of reorganizing the backbone output to match the expected input size of the HPB model, i.e., $[B, C, P]$, where C is the final number of channels and P is the final feature map size.

In summary, the Backbone Encapsulator module is used to make the output of the pre-trained backbone consistent with the HPB model. This condition is

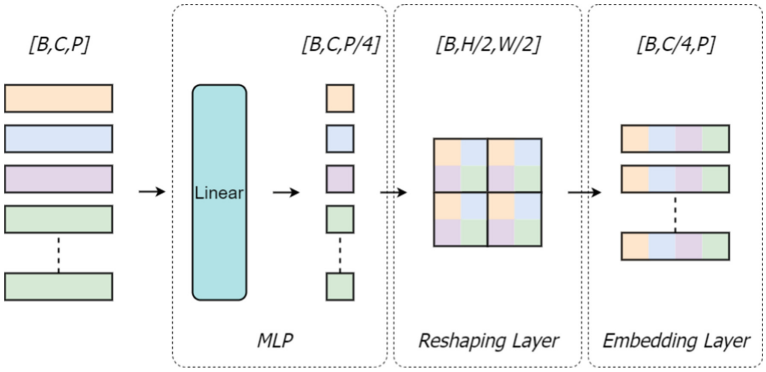


Fig. 2. Representation of the PPHR module. From left to right: example feature map derived from the image, patch representation of the image (MLP block), linear transformation (MLP block), patch representation of the image after linear transformation (Reshaping layer), patch reorganization (Embedding layer).

achieved by modifying the backbone output to match the expected input size of the HPB model. A schematic illustration is shown in Fig. 1.

3.3 Patch Pooling and Hierarchical Reconstruction Module

The Swin Transformer [9] is a state-of-the-art transformer-based architecture for image classification. It features a hierarchical structure that efficiently handles large images by gradually reducing the feature map's resolution while increasing the transformer layers' receptive field.

The Patch Pooling and Hierarchical Reconstruction (PPHR) module works by merging nearby patches into larger ones, reducing the number of patches, and increasing the receptive field of the following transformer layers.

Some distinctions exist between the Swin Transformer's patch merging layer and our PPHR module. The PPHR module comprises three primary components: a *multi-layer perceptron* (MLP), a *reshaping layer*, and an *embedding layer*. A schematic illustration is shown in Fig. 2.

The MLP takes a series of patches as input, with a $[B, C, P]$ size. Here, B represents the batch size, C is the number of patches, and P denotes the size of each patch squared. These patches are then projected into a space of size $[B, C, E]$, where E represents the expansion embedding dimension. Finally, the patches are further projected into a lower-dimensional space of size $[B, C, P/4]$. These patches are normalized using a *Layer Normalization* layer before passing through the linear layer.

The output of the MLP is then passed through the Reshaping layer, which reshapes it into a 2D grid of patches employing the *rearrange function*. Finally, the grid is rearranged into shape $[B, C/4, P]$.

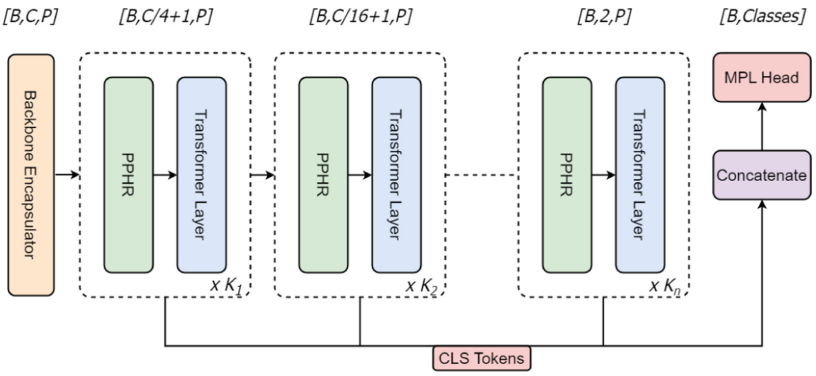


Fig. 3. The architecture of the proposed Hierarchical Pretrained Backbone Vision Transformer over different stages.

After the reshaping layer, the embedding layer applies a linear and another Layer Normalization layer. The final result is an output tensor of size $[B, C/4, K]$, where K is the same as P in our approach.

3.4 Overall Architecture

Our proposal's overall architecture involves several stages that lead to a single vector of image features. It is illustrated in Fig. 3.

Each stage is composed of a PPHR module (see Sect. 3.3) and a set of Transformer Layers [14], which are repeated K_n times. The value of K_n is chosen, and n is the current stage's index.

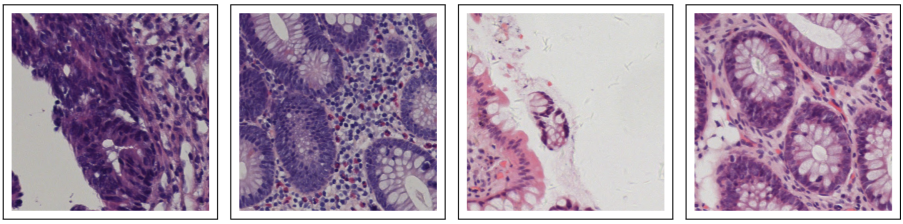
Our approach involves using learnable class tokens, similar to traditional ViTs. However, we use them separately for each stage instead of simultaneously like conventional ViTs.

At the beginning of each stage, a learnable positional encoding is added, and class tokens with size $[B, 1, P]$ are concatenated. After the final stage, the class tokens from each stage are concatenated and used as input to an MLP head, which produces the scores for each possible class. The concatenation of the class tokens for each stage produces a hierarchical dense representation of the image, resulting in higher accuracy for the current task.

Ultimately, the proposed architecture combines the strengths of the PPHR module and Transformer Layers to learn a hierarchical representation of the input image, resulting in improved performance on image classification tasks.

4 Experiments

The goal of the experiments is to determine if the proposed method can enhance the accuracy of histopathological image classification concerning off-the-shelf Transformer-based architectures. We will begin by introducing the dataset



(a) Adenocarcinoma. (b) Adenoma. (c) Normal. (d) Serrated.

Fig. 4. Visual representation of the four classes included in the Chaoyang dataset.

(Sect. 4.1), outlining the experimental setup (Sect. 4.2), and then analyzing the results in Sect. 4.3.

4.1 Chaoyang Dataset

The Chaoyang dataset is a histopathology image collection containing 512×512 sample patches of four different types of colon-rectum tissue conditions: normal, serrated, adenocarcinoma, and adenoma (see Fig. 4). The dataset was constructed in a realistic and practical context, meaning that it may contain noise and other imperfections that can make the classification task more challenging. The authors provide the dataset in two sets: a training set with 4,022 samples and a test set with 2,039 samples.

The class distribution in the dataset is shown in Fig. 5. As can be observed, the dataset is imbalanced, with a more significant number of adenocarcinoma samples compared to the other classes. Overall, the Chaoyang dataset provides a challenging and realistic benchmark for evaluating the performance of image classification models on histopathology images of the colon rectum. The dataset's imbalance and potential imperfections make it a more realistic representation of real-world applications.

4.2 Experimental Setting

We validated our approach using the Chaoyang dataset [21]. To provide a fair comparison between our proposed architecture and several off-the-shelf Vision Transformers, i.e., ViT [3] and SwinV2 [8], we reported accuracy, F1-score, precision, and recall as performance metrics, calculated using the macro average. Additionally, we provided information on the backbone used and the number of parameters for each employed configuration.

Training Details. To help avoid overfitting, we reserved 10% of the training set with the actual class distribution to build a validation set.

Then, we implemented an oversampling strategy on the training set, comprising two steps to address the class imbalance. Firstly, we repeated the underrepresented samples until we had equal samples for each class. Secondly, we augmented

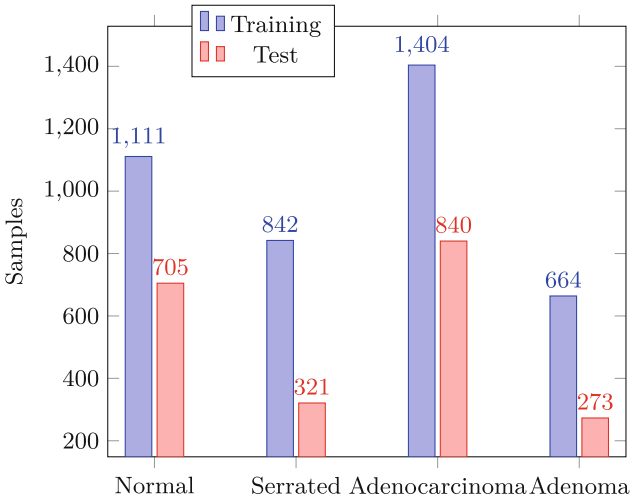


Fig. 5. Class distribution in the Chaoyang dataset. It is imbalanced, with a meager presence of serrated and adenoma classes.

Table 1. Image augmentation parameters adopted for models training.

Augmentation	Parameters	Probability
HorizontalFlip	–	0.5
VerticalFlip	–	0.5
RandomRotate90	[90, 180, 270] degrees	0.5
RandomResizedCrop	[0.5, 1.0] of original size	0.5

the repeated samples online to enhance the diversity of the images' representations with various geometric transformations while preserving important visual features. The augmentations are indicated in Table 1. No color augmentations were considered to preserve the staining procedure and obtain a first baseline.

Our augmentation pipeline effectively increased the diversity of our training set and improved our model's ability to handle variations in input data.

Implementation Details. The experiments were conducted on a workstation with an Intel(R) Core(TM) i5-9400 @ 4.10 GHz CPU, 32 GB RAM, and an NVIDIA RTX 3060 GPU with 12 GB memory. Every method was trained with the following hyperparameters: AdamW was set as the optimizer with a weight decay of 1×10^{-2} and momentum of 0.9. The initial learning rate was 1×10^{-4} across a total of 100 epochs. Dropout was set to 0.2.

HPB Configuration. As for HPB, SwinV2 Tiny [8] and ViT base [3] was chosen as the backbones to experiment with the architecture with two different

kinds of image representations: hierarchical provided by SwinV2 and standard by ViT base. Each stage had a different number of repeats, with 2, 2, 4, and 2, respectively. We incorporated 8 attention heads with a size of 64 for each transformer layer. For the embedding dimension, we opted for a small value of 384. Lastly, we set the expansion factor for the feed-forward layer inner dimension to 3.

4.3 Experimental Results

Table 2 presents the results obtained by our proposed methods in its two versions: HPB-ViT using ViT Base [3] as the backbone, from now on referred to as **HPB-Base**, and **HPB-Tiny**, HPB-ViT with the SwinV2 Tiny [8] backbone, and some state-of-the-art methods on the Chaoyang dataset. For fairness, the table is divided into three sections: the first is, to the best of our knowledge, the current baseline on this dataset [21]; the second represents ViT-based architectures, and the last includes SwinV2-based architectures. We point out that we applied the same oversampling and augmentation strategy for every experimented approach, except for the baseline, which was reported from the work of Zhu et al. [21].

In particular, the method of Zhu et al. [21], referred to in our table as *NSHE+CNN*, uses ResNet-34 CNN [5] as the classification model. In addition, the authors proposed the *noise suppressing and hard enhancing* (NSHE) technique to make the classification model resistant to possible noise interference in the images, as the Chaoyang dataset was collected from real-world settings and the main goal was to create a method robust to noise. According to the authors, their approach achieved an accuracy of 0.83 and an F1-score of 0.77. In addition, we also used ViT [3] and SwinV2 [8] for comparison purposes. ViT comes in two versions: *Base* and *Large*, while SwinV2 has *Tiny* and *Small* versions.

Both proposed versions result in significant performance improvements across all reported metrics, with some differences. The HPB-Base version outperforms the baseline work and ViT architectures in every metric. Despite having slightly more parameters than ViT Base, HPB-Base is a suitable compromise between ViT Base and ViT Large in terms of performance gains.

On the other hand, HPB-Tiny performs better than both SwinV2 models and is the top performer among all the tested models. It is based on the SwinV2 Tiny backbone and has fewer parameters than HPB-Base but slightly more than SwinV2 Small. However, it still has double the parameters of SwinV2 Tiny. Overall, HPB-Tiny strikes a good balance between performance and complexity.

Overall, the results demonstrate that the proposed HPB models outperformed the baseline on the Chaoyang dataset [21], highlighting the effectiveness of the proposed approach, even compared to off-the-shelf Transformers architectures. Additionally, we observed that our proposed HPB architecture outperformed the work of Zhu et al. [21], specifically designed to take into account the noise interference commonly found in real-world acquired images. This aspect highlights the robustness and effectiveness of our approach and underscores its potential for handling noise interference images with improved accuracy and reliability.

Table 2. Experimental results obtained on the Chaoyang dataset [21]. The reported performance metrics, calculated using the macro average, include accuracy, F1-score, precision, and recall. Additionally, we provided information on the backbone used and the number of parameters for each model tested. The best results are highlighted in bold, while the second best are underlined.

Model	Backbone	Image size	Acc↑	F1↑	Pre↑	Rec↑	Params↓
NSHE+CNN [21]	ResNet-34	224	0.83	0.77	0.78	0.75	–
ViT Base [3]	–	224	0.84	0.79	0.78	0.79	85.8M
ViT Large [3]	–	224	0.84	0.79	0.79	0.78	303.3M
HPB-Base (Our)	ViT Base [3]	224	0.85	0.80	0.81	0.80	112.3M
SwinV2 Tiny [8]	–	256	0.84	0.79	0.80	0.79	27.6M
SwinV2 Small [8]	–	256	0.84	0.79	0.80	0.78	51.3M
HPB-Tiny (Our)	SwinV2 Tiny [8]	256	0.86	0.81	0.82	0.80	54M

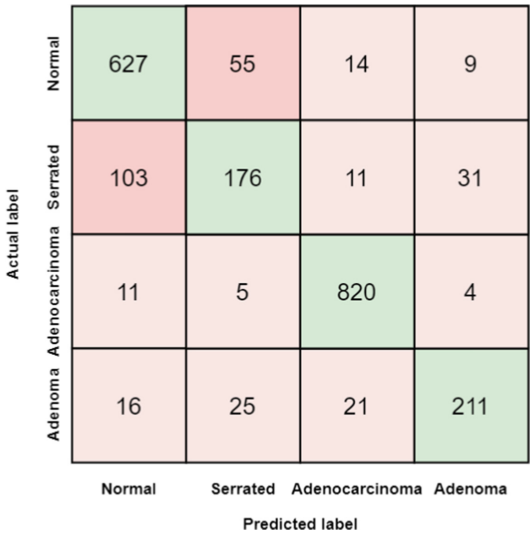


Fig. 6. Confusion matrix obtained with HBP-Tiny. It shows the misclassification issues between normal and serrated classes.

Limitations. Despite its positive outcomes, the proposed HPB architecture has some limitations. One is the accurate differentiation between normal and serrated classes. Even seasoned pathologists struggle with this classification, which could result in incorrect labeling. Therefore, this issue must be addressed appropriately to ensure the architecture's practical application in certain circumstances.

We use the confusion matrix in Fig. 6 to illustrate this aspect. It shows misclassifications for the serrated and normal classes in the best-proposed architecture (HPB-Tiny). Further refinements are necessary to accurately classify these classes.

5 Conclusions

The HPB-ViT architecture has shown great potential in automating the classification of histopathological images. This result is achieved by utilizing the pretrained SwinV2 Tiny backbone, which allows the HPB architecture to learn how to classify these images with less labeled data, yet still perform at a state-of-the-art level. The effectiveness of this proposed architecture is demonstrated through its evaluation of the Chaoyang histopathology dataset for image classification.

There are several ways to improve the performance, practicality, and classification accuracy of the proposed HPB-ViT architecture. One possible direction is to include new pretrained ViT backbones like T2T-ViT, DeepViT, and CvT. Additionally, further integrating CNNs as backbones in the HPB architecture could improve its performance. Testing the approach on different datasets may help generalize the proposed architecture's effectiveness in various fields and show its potential against noise interference. Finally, incorporating multiple backbones into the HPB architecture could improve classification accuracy.

Acknowledgments. We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No.3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR)"

References

1. Chen, H., et al.: Gashis-transformer: a multi-scale visual transformer approach for gastric histopathological image detection. Pattern Recogn. **130**, 108827 (2022)
2. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Li, F.-F.: A large-scale hierarchical image database. In: Imagenet (2009)
3. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 3–7 May 2021 (2021)
4. Glotsos, D., et al.: Improving accuracy in astrocytomas grading by integrating a robust least squares mapping driven support vector machine classifier into a two level grade classification scheme. Comput. Methods Progr. Biomed. **90**(3), 251–261 (2008)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. Hendrycks, D., Lee, K., Mazeika, M.: Using pre-training can improve model robustness and uncertainty. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA, 9–15 June 2019, vol. 97 of Proceedings of Machine Learning Research, pp. 2712–2721. PMLR (2019)

7. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
8. Liu, Z., et al.: Swin transformer V2: scaling up capacity and resolution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022, pp. 11999–12009. IEEE (2022)
9. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021, pp. 9992–10002. IEEE (2021)
10. Putzu, L., Fumera, G.: An empirical evaluation of nuclei segmentation from h&e images in a real application scenario. Appl. Sci. **10**(22), 7982 (2020)
11. Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: a survey. Medical Image Anal. **67**, 101813 (2021)
12. Steiner, A.P., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L.: How to train your vit? data, augmentation, and regularization in vision transformers. In: Transactions on Machine Learning Research (2022)
13. Szegedy, C., et al.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015, pp. 1–9. IEEE Computer Society (2015)
14. Vaswani, A., et al. Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
15. Wu, H., et al.: Introducing convolutions to vision transformers. In: Cvt (2021)
16. Xu, X., Kapse, S., Gupta, R., Prasanna, P.: Vit-dae: transformer-driven diffusion autoencoder for histopathology image analysis. CoRR, abs/2304.01053 (2023)
17. Li, Y., et al.: Training vision transformers from scratch on imagenet. In: Tokens-to-Token Vit (2021)
18. Zhang, X., Chan, F.T.S., Mahadevan, S.: Explainable machine learning in image classification models: an uncertainty quantification perspective. Knowl. Based Syst **243**, 108418 (2022)
19. Zhou, D., et al.: Towards deeper vision transformer. In: Deepvit (2021)
20. Zhou, X., Tang, C., Huang, P., Tian, S., Mercaldo, F., Santone, A.: Asi-dbnet: an adaptive sparse interactive resnet-vision transformer dual-branch network for the grading of brain cancer histopathological images. Interdisc. Sci. Comput. Life Sci. **15**(1), 15–31 (2023)
21. Zhu, C., Chen, W., Peng, T., Wang, Y., Jin, M.: Hard sample aware noise robust learning for histopathology image classification. IEEE Trans. Med. Imaging **41**, 881–894 (2021)