TransPath: Transformer-Based Self-supervised Learning for Histopathological Image Classification

Xiyue Wang[1], Sen Yang[2], Jun Zhang[2], Minghui Wang[1], Jing Zhang[3(✉)], Junzhou Huang[2], Wei Yang[2], and Xiao Han[2]

[1] College of Computer Science, Sichuan University, Chengdu, China
haroldhan@tencent.com
[2] Tencent AI Lab, Shenzhen, China
[3] College of Biomedical Engineering, Sichuan University, Chengdu, China
jing_zhang@scu.edu.cn

Abstract. A large-scale labeled dataset is a key factor for the success of supervised deep learning in histopathological image analysis. However, exhaustive annotation requires a careful visual inspection by pathologists, which is extremely time-consuming and labor-intensive. Self-supervised learning (SSL) can alleviate this issue by pre-training models under the supervision of data itself, which generalizes well to various downstream tasks with limited annotations. In this work, we propose a hybrid model (*TransPath*) which is pre-trained in an SSL manner on massively unlabeled histopathological images to discover the inherent image property and capture domain-specific feature embedding. The *TransPath* can serve as a collaborative local-global feature extractor, which is designed by combining a convolutional neural network (CNN) and a modified transformer architecture. We propose a token-aggregating and excitation (TAE) module which is placed behind the self-attention of the transformer encoder for capturing more global information. We evaluate the performance of pre-trained *TransPath* by fine-tuning it on three downstream histopathological image classification tasks. Our experimental results indicate that *TransPath* outperforms state-of-the-art vision transformer networks, and the visual representations generated by SSL on domain-relevant histopathological images are more transferable than the supervised baseline on ImageNet. Our code and pre-trained models will be available at https://github.com/Xiyue-Wang/TransPath.

Keywords: Self-supervised learning · Transformer · Histopathological image

1 Introduction

Benefiting from a massive amount of labeled data, deep learning has obtained success in medical image analysis. However, such careful annotations are very scarce in histopathological whole-slide images (WSIs). Gigapixel size of the image

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12908, pp. 186–195, 2021.
https://doi.org/10.1007/978-3-030-87237-3_18

creates a large search space for labeling, and the wide variations, even between the same class, further increase the annotation challenge. Extracting effective features from unlabeled histopathological images can promote the development of digital pathology and aid pathologists for faster and more precise diagnoses.

To address these issues, transfer learning from large labeled ImageNet [14] are proven to be an effective training strategy [12]. However, a large domain shift exists between natural images and histopathological images. The desired approach to tackle this domain shift is pre-training or training from scratch on the domain-relevant data, which is limited by the annotation-lacking problem. To address this, self-supervised pre-training is a possible alternative, which learns the image representation using the supervision signal produced from the data itself. Self-supervised learning (SSL) has achieved superior performance in the field of natural images for image classification, semantic segmentation, and object detection tasks [2,6].

For the histopathological image analysis, there have been several works that apply SSL techniques (e.g. CPC, SimCLR, and MoCo) to improve the classification performance [3,8,9,11,15]. However, there are still two aspects that could be improved. First, these works lack a large domain-specific histopathological image dataset for self-supervised pre-training, and their pre-trained image patches were cropped from a small number of WSIs (up to 400 WSIs). The small dataset results in a lack of sample diversity and prevents robust and generic representation learning. Second, only CNN structures are applied. CNN has a good capacity to learn low-level texture content features (local features) which is a crucial determinant in the classification task. The learning of global context features is often limited by the receptive field. The cropped histopathological image patches are usually enough to capture both the cell-level structure (e.g., cellular microenvironment) and tissue-level context (e.g., tumor microenvironment). Thus, both the local and global features are required for digital pathology. To learn global features, the transformer-based architecture may be a viable alternative, which originates from the natural language processing (NLP) field and has shown great potential in the computer vision field [4].

In this work, we collected the current largest histopathological image dataset for self-supervised pre-training, which comprises approximately 2.7 million images with the size of 2048×2048 cropped from a total of 32,529 WSIs from the cancer genome atlas (TCGA[1]) and pathology AI platform (PAIP[2]) datasets. The utilized datasets guarantee sample diversity and cover multi-sites (over 25 anatomic sites) and multi-cancer subtypes (over 32 classes). We propose a hybrid CNN-transformer framework to enhance the local structure and global context features extraction, where the transformer is modified by adding a customized TAE module into the self-attention block of the transformer. CNN extracts local features by convolutional computation. Transformer captures global dependencies through the information interaction among the tiled patches (tokens) from

[1] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/.
[2] http://www.wisepaip.org/paip/.

the generated features by CNN. The customized TAE is designed to further enhance global weight attention by taking all tokens into account. The combination of CNN and transformer networks can provide a robust and generalized feature extractor to capture both local fine structure and global context for the histopathological image analysis.

Our main contributions can be summarized as follows:

(1) To the best of our knowledge, this is the first self-supervised pre-training work carried out on the current largest public histopathological image dataset, which guarantees the sample diversity and helps the network capture sufficient domain-specific features.
(2) A novel hybrid architecture, which combines CNN and transformer to simultaneously capture the local structure and global context, is proposed to perform the histopathological image classification.
(3) We modified the self-attention layer of the vanilla transformer encoder by inserting our customized TAE module to strengthen the global feature learning ability further.
(4) Benefiting from the above design, our model outperforms state-of-the-art vision transformer networks on three public histopathological datasets. The proposed model can serve as a feature extractor for the downstream pathology image analysis. Our code and pre-trained model have been released online to facilitate reproductive research.

2 Methods

The overview of the proposed framework is provided in Fig. 1. There are two key points in our algorithm, namely construction of backbone and self-supervised pre-training, which will be introduced in the following.

2.1 Proposed Hybrid Network Backbone (*TransPath*)

The proposed hybrid network backbone *TransPath* aims to utilize both the local feature mining ability of CNN and the global interaction ability of the transformer. As shown in Fig. 1, the input image x is augmented as x_1 and x_2, which are then input to our designed backbone. In the backbone, CNN is first used to extract features ($\frac{H}{32} \times \frac{W}{32} \times 1024$). The produced features are the grid outputs $\frac{H}{32} \times \frac{W}{32} \times D$, where each grid element has a feature representation of length D. We denote each grid element as a token (i.e. token or word as in NLP), and the two-dimensional elements are flattened to a sequence of tokens $N \times D$, where $N = \frac{H}{32} \times \frac{W}{32}$. Besides the image features, we also provide the corresponding learnable position embeddings (*pos*). Then, similar to the ViT, an extra learnable classifier embedding (*cls*) as a single token is added to the transformer input to form a new input X with $N + 1$ patches. The final state of *cls* is acted as the final features generated by the transformer.

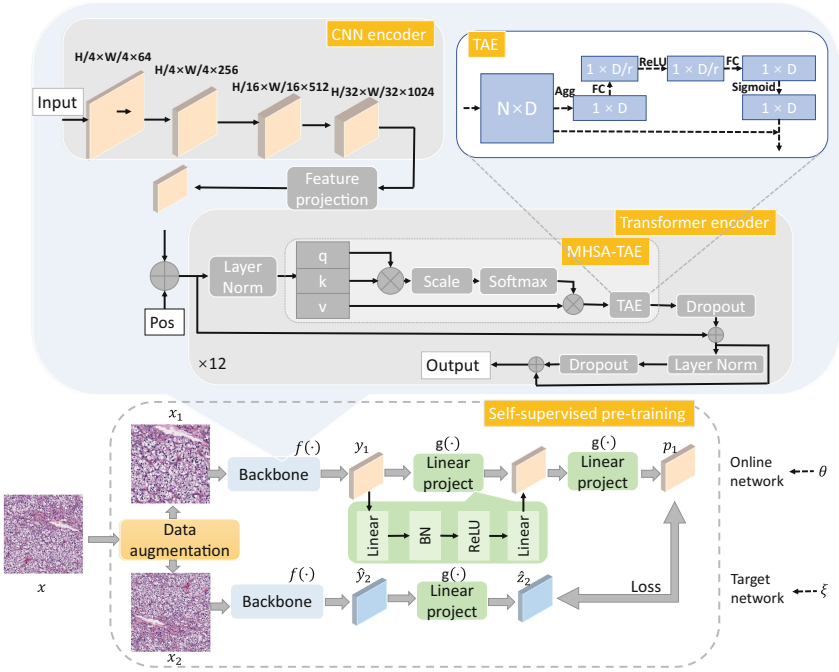

Fig. 1. Overview of the proposed CNN-transformer-based self-supervised pre-training procedure. BYOL architecture is used for self-supervised learning due to its negative sample independence [5]. The CNN block utilized in this work is ResNet50, which can be replaced by any CNN architecture. The transformer encoder is modified by inserting our customized TAE module for more sufficient global feature learning. BN: batch normalization.

Different from previous studies, the multi-head self-attention (MHSA) is modified in this work by adding our customized token-aggregating and excitation (TAE) module to produce a new MHSA block called MHSA-TAE, which is computed as follows.

$$\begin{aligned} \text{Linear projection}: \boldsymbol{Q} \leftarrow W_q X, \boldsymbol{K} \leftarrow W_k X, \boldsymbol{V} \leftarrow W_v X \\ \text{Self-attention}: F_{MHSA} = \text{Softmax}\left(\boldsymbol{Q}\boldsymbol{K}^\top / \sqrt{d}\right)\boldsymbol{V} \\ \text{Attention aggregation}: F_{agg} = \frac{1}{N+1}\sum_{i=1}^{N+1} F_{MHSA}(i,j) \\ \text{Output}: F_{MHSA-TAE} = \sigma(W' \times \sigma(W \times F_{agg})) \times F_{MHSA} \end{aligned} \tag{1}$$

where the input X is linearly projected into 3 subspaces with weights W_k, W_q, and W_v to get $\boldsymbol{K}$, $\boldsymbol{Q}$, and $\boldsymbol{V}$, which consider all tokens, thereby helping generate global attention vectors. In the self-attention computation process, the

interaction between $\boldsymbol{K}$ and $\boldsymbol{Q}$ is computed by the dot product. And then, the weight is scaled with factor b and projected into $\boldsymbol{V}$ space to obtain a new feature embedding F_{MHSA}. To further enhance the global feature extraction, the TAE module averages the input tokens to obtain aggregated features F_{agg} with the size of $1 \times D$, where the i and j denote the row and column in F_{MHSA}, respectively. The final step excites F_{agg} and re-projects it to F_{MHSA}, where W and W' represent the weights in the fully connected layers, σ denotes the ReLU activation function. The final produced feature weight $F_{MHSA-TAE}$ considers more sufficient global information since each element in $F_{MHSA-TAE}$ is the aggregated result across all tokens. After that, the attention weight $F_{MHSA-TAE}$ is imposed on the input feature embedding X by residual connection to form the output features y of our *TransPath*.

2.2 Self-supervised Pre-training

Self-supervised pre-training aims to learn a visual representation of raw data without manual supervision. We adopt BYOL architecture [5], which avoids the definition of negative samples in contrastive learning. As illustrated in Fig. 1, there are two parallel paths that share a similar structure but different network weights. The backbone adopts our proposed *TransPath*. We train the online network with parameter θ and to-be-updated parameter ξ of target network by $\xi \leftarrow \tau\xi + (1-\tau)\theta$.

Given a random histopathological image x, its two augmentations (x_1 and x_2) are then alternately fed to the two parallel paths. When x_1 and x_2 respectively pass through our *TransPath* (formulated as $f(\cdot)$) in the online and target networks, which generates visual representations $y_1 = f^{\theta}(x_1)$, $\hat{y_2} = f^{\xi}(x_2)$. Then, linear projection head $g(\cdot)$ is adopted to transform the representations to other latent spaces, e.g. $p_1 = g^{\theta}(g^{\theta}(y_1))$ in online network, $\hat{z_2} = g^{\xi}(\hat{y_2})$ in target network. Symmetrically, the swapped prediction separately feeds x_1 and x_2 to the target and online networks, obtaining $y_2 = f^{\theta}(x_2)$, $\hat{y_1} = f^{\xi}(x_1)$, $p_2 = g^{\theta}(g^{\theta}(y_2))$, $\hat{z_1} = g^{\xi}(\hat{y_1})$. The objective function is optimized by minimizing the distance error of ℓ_2-norm between online and target networks:

$$L(p, z) = -\frac{p}{\|p\|_2} \cdot \frac{z}{\|z\|_2} \tag{2}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm. The symmetric objective function L_{loss} is calculated as:

$$L_{loss} = \frac{1}{2}L(p_1, \hat{z}_2) + \frac{1}{2}L(p_2, \hat{z}_1) \tag{3}$$

After self-supervised pre-training, the pre-trained backbone can then be fine-tuned to various downstream tasks.

3 Datasets

The datasets used for pre-training and fine-tuning are collected from different projects. The pre-trained histopathological image dataset is collected from two

WSI-level datasets: TCGA and PAIP. We crop these WSIs as approximately 2.7 million unlabeled images (2048×2048 pixels). After the self-supervised pre-training process, we fine-tune our CNN-transformer network to prove its classification ability on three public histopathological image datasets: NCT-CRC-HE, PatchCamelyon, and MHIST. The following will introduce them in detail.

TCGA. TCGA includes 30,072 WSIs covering over 25 anatomic sites and over 32 cancer subtypes. We crop these WSIs into images with the size of 2048×2048 pixels. After excluding images without tissues, we randomly select M images from each WSI. When the number of images in a WSI is less than M, all images available will be used. It is worth noting that because our image size is large enough, $M = 100$ can almost cover the entire WSI area. Finally, we generate a TCGA pre-training dataset with 2,476,964 unlabeled histopathological images.

PAIP. PAIP releases 2457 WSIs, including 6 cancer types (liver, renal, colorectal, prostatic, pancreatic, and cholangio cancers). Following the same image extraction strategy as the TCGA dataset, we produce a PAIP pre-training dataset with 223,557 unlabeled histopathological images.

NCT-CRC-HE. NCT-CRC-HE is provided by National Center for Tumor Diseases (NCT) to identify 8 colorectal cancer tissues and 1 normal tissue. A total of 100,000 images with 224×224 pixels are extracted from 86 WSIs [7], which is used as the training set in the fine-tuning process. The test set contains 7180 images extracted from 50 patients with colorectal adenocarcinoma.

PatchCamelyon. PatchCamelyon (PCam) contains 327,680 images (96×96) [1] for breast cancer detection, which are extracted from Camelyon16 challenge dataset with 400 WSIs. The data splitting in our fine-tuning experiment follows the PCam, resulting in 245,760, 40,960, and 40,960 images for training, validation, and test, respectively.

MHIST. MHIST is designed to classify colorectal polyps as benign and precancerous [18], which consists of 3,152 images with the size of 224×224 pixels. Following the official data splitting, 2,175 images (1,545 benign and 630 precancerous) are used for training and 977 images (617 benign and 360 precancerous) are used for evaluation.

4 Experimental Results and Discussions

4.1 Experimental Setup

In the pre-training process, we use BYOL architecture to train our *TransPath*. The input images are downsampled as 512×512 pixels and batch size is set as 256. The data augmentation strategies keep the same as SimCLR [2]. The network is optimized by an SGD optimizer [10] with an initial learning rate of 0.03 and its weight decay of 0.0001. Our pre-trained model is implemented in PyTorch [13] framework and with 32 Nvidia V100 GPUs, which takes 100 epochs and around 200 h to converge. In the fine-tuning process, the input image size keeps consistent with the corresponding datasets. SGD is used as the default optimizer with a batch size of 64 and an initial learning rate of 0.0003.

4.2 Ablation Study

We conduct a set of ablation studies to investigate the effects of three key components (CNN encoder, transformer encoder, and TAE module) within our proposed *TransPath*. The results are listed in Table 1. In this experiment, CNN is initialized by ImageNet pre-training. It is seen that CNN can obtain a satisfactory classification result, especially in the NCT-CRC-HE dataset. To alleviate the weak global feature extraction problem of CNN, we then integrate CNN and transformer (ViT), which generates consistent performance gains across three datasets in terms of three metrics (AUC +2%, ACC +3%, and F1-score +3%). To further enhance the global context features in the transformer, we insert a customized TAE module into the MHSA block, which further improves the classification performance by 3% in terms of the F1-score on MHIST. The above-reported results demonstrate that the combination of CNN with local feature capture ability and transformer with global dependence learning ability can produce robust visual representation for the histopathological image analysis.

Table 1. Ablation study

Networks	**Datasets and metrics**							
	MHIST			**NCT-CRC-HE**		**PatchCamelyon**		
	F1-score	ACC	AUC	F1-score	ACC	F1-score	ACC	AUC
CNN	0.7957	0.8095	0.9188	0.9099	0.9081	0.8227	0.844	0.9311
CNN+Trans	0.8277	0.8302	0.9378	0.9313	0.9319	0.8587	0.8605	0.9536
CNN+Trans+TAE (ours)	0.8586	0.8651	0.9476	0.9486	0.9483	0.8661	0.8766	0.9631
CNN+Trans+TAE +SSL (our full method)	0.8993	0.8968	0.9727	0.9582	0.9585	0.8983	0.8991	0.9779

4.3 Comparisons with Vision Transformer Networks

This subsection compares our *TransPath* with current state-of-the-art vision transformer networks, including pure transformer-based architectures (ViT [4] and T2T-ViT-24 [20]) and hybrid CNN-transformer paradigms (VT-ResNet [19] and BoTNet-50 [16]). In this experiment, our model is initialized by ImageNet pre-training and other models keep consistent as their publications. As shown in Fig. 2, the hybrid paradigms consistently achieve better performance compared with the pure-transformer across three datasets. Our method follows the design of CNN-transformer and achieves the best performance in these histopathological image classification tasks.

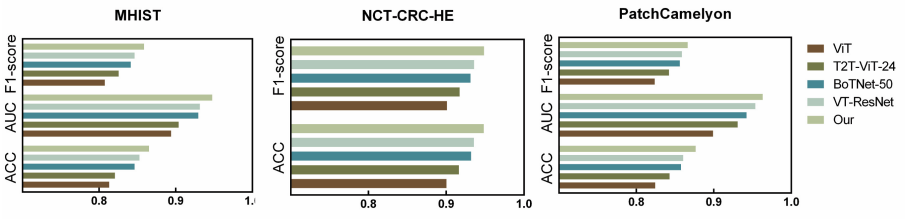

Fig. 2. Comparisons with current state-of-the-art vision transformer networks evaluated on three public datasets.

4.4 Comparisons with Various Pre-training Strategies

This subsection fine-tunes our *TransPath* under different pre-training strategies, including random parameter initialization, ImageNet pre-training, and self-supervised pre-training. As shown in Table 2, when the network is trained from scratch (random parameter initialization), the lowest performance can be seen in all three databases. The pre-trained *TransPath* on domain-relevant data in a self-supervised manner outperforms ImageNet pre-training by 4% of F1-score and 3% of ACC and AUC on MHIST, which demonstrates the potential of SSL on the histopathological image analysis.

Table 2. Comparisons with various pre-training strategies

Networks	Datasets and metrics							
	MHIST			NCT-CRC-HE		PatchCamelyon		
	F1-score	ACC	AUC	F1-score	ACC	F1-score	ACC	AUC
Random	0.8183	0.8206	0.8789	0.9332	0.9322	0.856	0.8567	0.9354
ImageNet	0.8586	0.8651	0.9476	0.9486	0.9483	0.8662	0.8766	0.9631
SSL (ours)	0.8993	0.8968	0.9727	0.9582	0.9585	0.8983	0.8991	0.9779

5 Conclusion

We proposed a customized CNN-transformer architecture for histopathological image classification. Our approach makes use of both local and global receptive fields to extract discriminating and rich features. The proposed TAE module is inserted in the transformer structure to better capture global information. The self-supervised pre-trained model on large domain-specific histopathological images can benefit various downstream classification tasks through transfer learning. Experimental results on three public datasets for three different classification tasks demonstrated the effectiveness of our proposed method and the pre-trained model. Future work will investigate the effects of various SSL methods, other data augmentation methods tailored for histopathological images [17],

and whether it is beneficial to separate the frozen slides from the formalin-fixed paraffin-embedded (FFPE) slides in the TCGA dataset.

Acknowledgements. This research was funded by the National Natural Science Foundation of China (No. 61571314), Science & technology department of Sichuan Province, (No. 2020YFG0081), and the Innovative Youth Projects of Ocean Remote Sensing Engineering Technology Research Center of State Oceanic Administration of China (No. 2015001).

References

1. Bejnordi, B.E., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA **318**(22), 2199–2210 (2017)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
3. Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A., Courtiol, P.: Self-supervision closes the gap between weak and strong supervision in histology. arXiv preprint arXiv:2012.03583 (2020)
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Grill, J.B., et al.: Bootstrap your own latent: a new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 (2020)
6. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
7. Kather, J.N., et al.: Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Med. **16**(1), 1–22 (2019)
8. Koohbanani, N.A., Unnikrishnan, B., Khurram, S.A., Krishnaswamy, P., Rajpoot, N.: Self-path: self-supervision for classification of pathology images with limited annotations. arXiv preprint arXiv:2008.05571 (2020)
9. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. arXiv preprint arXiv:2011.08939 (2020)
10. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
11. Lu, M.Y., Chen, R.J., Wang, J., Dillon, D., Mahmood, F.: Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. arXiv preprint arXiv:1910.10825 (2019)
12. Mormont, R., Geurts, P., Marée, R.: Multi-task pre-training of deep neural networks for digital pathology. IEEE J. Biomed. Health Inform. **25**(2), 412–421 (2020)
13. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. **32**, 8026–8037 (2019)
14. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
15. Srinidhi, C.L., Kim, S.W., Chen, F.D., Martel, A.L.: Self-supervised driven consistency training for annotation efficient histopathology image analysis. arXiv preprint arXiv:2102.03897 (2021)

16. Srinivas, A., Lin, T.Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. arXiv preprint arXiv:2101.11605 (2021)
17. Tellez, D., et al.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med. Image Anal. **58**, 1–9 (2019)
18. Wei, J., et al.: A petri dish for histopathology image analysis. arXiv preprint arXiv:2101.12355 (2021)
19. Wu, B., et al.: Visual transformers: token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677 (2020)
20. Yuan, L., et al.: Tokens-to-Token ViT: training vision transformers from scratch on ImageNet. arXiv preprint arXiv:2101.11986 (2021)