Multi-scale Contrastive Learning with Attention for Histopathology Image Classification

Jing Wei Tan[a], Khoa Tuan Nguyen[b], Kyoungbun Lee[c], and Won-Ki Jeong[a]

[a]Department of Computer Science and Engineering, Korea University, Seoul, South Korea
[b]Department of Environmental Technology, Food Technology and Molecular Biotechnology, Ghent University Global Campus, Incheon, South Korea
[c]Department of Pathology, Seoul National University Hospital, Seoul, South Korea

ABSTRACT

Whole slide images (WSIs) in histopathology naturally provide multi-scale information. Several previous studies have shown that leveraging such multi-scale information in histopathology image analysis is effective to improve performance. Here, we propose making use of recent advances in contrastive learning and self-attention techniques in multi-scale WSIs for cancer subtype classification using weak labels. The proposed method is based on a Siamese architecture to share a common encoder network for images on different scales to reduce the model size and training cost. In addition, we propose a variant of the self-attention module specifically designed for multi-scale WSIs so that the network can focus on important textural features across different image scales. We assess the efficacy of the proposed method via an ablation study on a real intrahepatic cholangiocarcinoma dataset. The result confirms that our method outperforms conventional multi-scale models with fewer model parameters.

Keywords: Multi-scale, Contrastive learning, Attention, Digital pathology

1. INTRODUCTION

Digitized histopathology images (i.e., whole slide images (WSIs)) are composed of several magnification levels that commonly range from 1.25× to 40×.This allows pathologists to visualize and analyze the WSIs using a computer by zooming in and out to find regions of interest without a microscope. Each scale of a WSI provides a different kind of information as the tissues show differences in their structures when visualized at different magnifications, as illustrated in Fig. 1.

Many existing histopathology image analysis studies are based only on a single scale, mostly at the highest magnification.[1–3] However, more recent studies have shown that leveraging multi-scale WSIs is effective to improve the performance.[4–6] A common strategy to leverage multi-scale WSIs in the workflow is using multiple neural networks, where each of them is trained individually using the data from a specific magnification level. Then, the intermediate result from each network is combined at the end to conduct the final downstream task (i.e., classification). The drawback of this approach is the increasing model size and computational burden.

Recently, contrastive learning and attention modules have been widely adopted and proven to be robust in histopathology image analysis .[2,7,8] The general concept of contrastive learning is enforcing similar images to stay closer in the feature space (and vice versa), whereas that of the attention mechanism is focusing more on the relevant context in the data. Although employing such techniques contributed to improving the model performance, there has been no intensive study on leveraging multi-scale data with contrastive learning and attention modules, which is the main motivation of this work.

Further author information: (Send correspondence to Won-Ki Jeong)
Jing Wei Tan: E-mail: jingwei_92@korea.ac.kr
Khoa Tuan Nguyen : E-mail: khoatuan.nguyen@ghent.ac.kr
Kyoungbun Lee : E-mail: azi1003@snu.ac.kr
Won-Ki Jeong : E-mail: wkjeong@korea.ac.kr

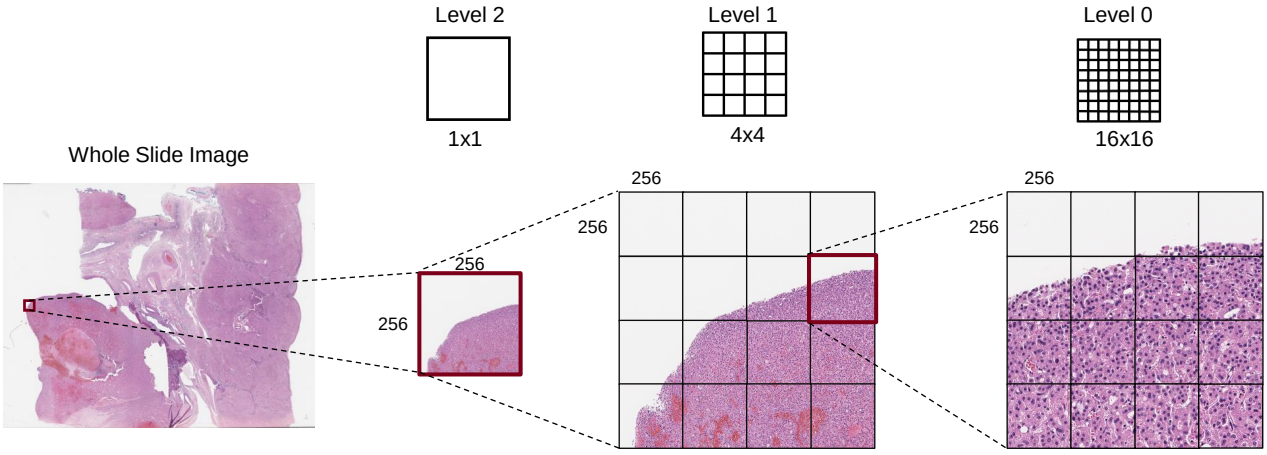

Figure 1. From a specific location in a WSI, we can extract one patch from Level 2, 16 patches from Level 1, and 256 patches from Level 0. These patches illustrate different morphological features at different levels (scales).

In this paper, we propose a novel multi-scale contrastive learning model and attention module that take advantage of the multiple magnification levels of WSIs specifically targeting the intrahepatic cholangiocarcinomas (IHCCs) subtype classification using weak labels (i.e., per-slide labels). Inspired by the supervised contrastive learning concept, we build a unified model that learns scale-invariant features from images at various magnification levels using a single Siamese encoder network. We also propose multi-scale attention (MSA), a variant of self-attention for multiple inputs, for multi-instance learning classification to make the model focus more on the important patch features in different scales. We demonstrate the efficacy of the proposed method via an ablation study on a real IHCC dataset.

2. RELATED WORK

2.1 WSI classification

Some previous studies employed only the highest magnification scale in their experiments. For example, Gao *et al.*[9] extracted patches from 40× magnification WSIs to classify the melanoma skin cancer WSIs into eight tissue types and grades of invasive ductal carcinomas of breast tumor WSIs. Some more recent work employed multi-scale learning; for example, in MS-DA-MIL,[4] the authors carried out single-scale learning, which trained the feature extractors for each scale separately in Stage 1, followed by the feature extractors from multiple scales plugged into the final model for bag classification. Another study, DSMIL,[5] proposed building a multi-scale model by training each scale with a self-supervised contrastive learning approach, SimCLR .[10] After that, the authors extracted the features from each scale and then concatenated them. MS-DA-MIL used only two scales (10× and 20×), whereas DSMIL used up to three different scales (1.25×, 5×, and 20×); however the model with two scales outperformed the model with three scales. Therefore, instead of the conventional methods using multiple single-scale models (Fig. 2(a)), we propose building a single unified model trained by images from various scales (Fig. 2(b)).

2.2 Contrastive learning

SimTriplet[7] proposed taking a multi-view of the WSIs as the inputs of the Siamese contrastive learning model instead of using the self-augmentation method. This model maximized the similarities of the intra-sample and inter-sample by using only positive pairs, without any negative pairs. Three augmented views from each adjacent pair were generated as the inputs of the SimTriplet. Furthermore, DSMIL[5] trained the feature extractors using SimCLR[10] first and used these pretrained feature extractors for their downstream tasks. The authors trained SimCLR to maximize the similarities between the patches from the same WSI and applied random

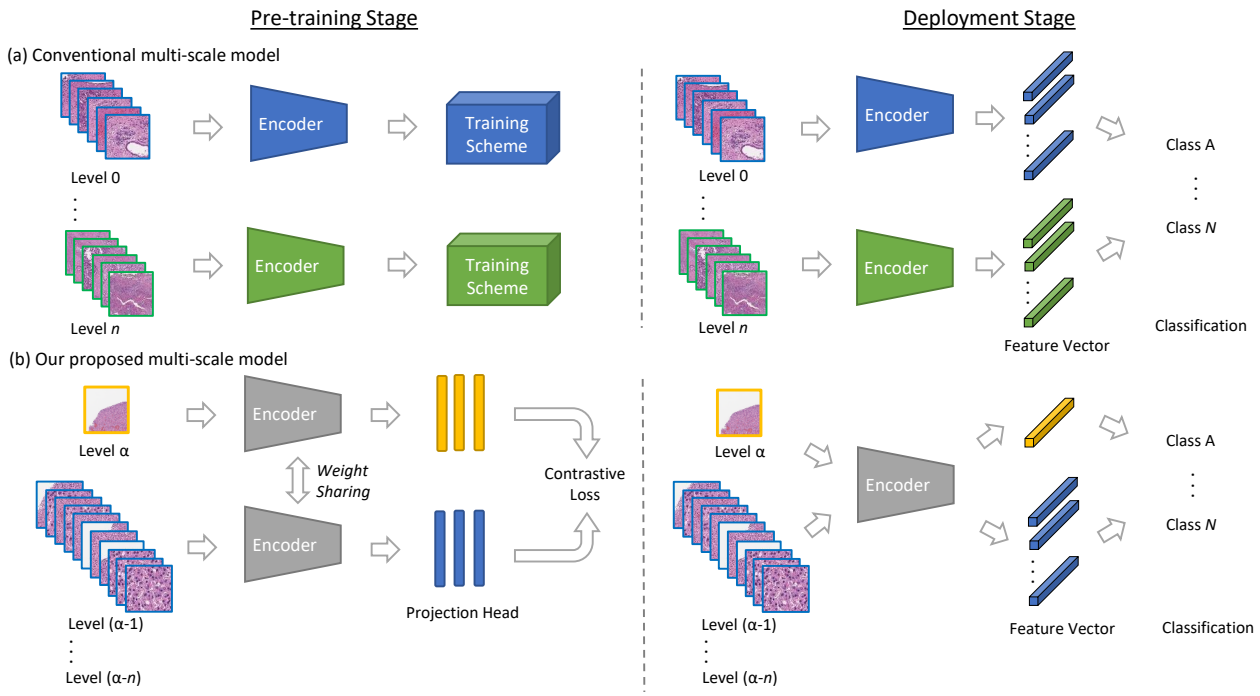

Figure 2. Comparison of (a) conventional multi-scale model, which consists of multiple single-scale models, and (b) our proposed model, which consists of a single unified model to train images from various scales.

image augmentation to the model inputs. Our proposed method takes advantage of the WSIs that provide multiple magnification levels instead of applying the scale augmentation method to obtain the synthetic positive pairs.

2.3 Attention

Given the set of feature vectors extracted from a well-trained encoder, the transformer-based method[11] has been widely applied to the classification stage.[8,12,13] With the combination of the local-patch feature vectors and the corresponding position embedding, Huang *et al.*[12] directly fed them to their transformer encoder so that it could aggregate the local information. Meanwhile, TransMIL[8] proposed that categorizing the feature vectors into positive, negative, and uncertain instances can yield more useful information for their transformer module. On the other hand, Kalra et al.[13] proposed the FocAtt-MIL technique, which aggregates the prediction by the learned focal factor. However, instead of aggregating the local features[8,12] or using the group of local features as the global information,[13] we use both extracted features from different multi-scale levels, which means the higher scale features can aggregate the information for the corresponding lower scale feature. It should also be noted that our idea is different from the "MIL aggregator" in DSMIL[5] since they used the highest score feature to mask the others in a bag of mixed features, whereas our attention mechanism distinguishes important higher scale features related to the inquired lower scale features. Furthermore, the methods of using patch features have been used in recent transformer-based papers[14,15] but still in a self-attention fashion, which means aggregating from the same scale level but not joint aggregating different scale levels at the same time, as in our model.

3. METHOD

Our proposed framework consists of two stages: (1) train a unified multi-scale model by using contrastive learning and (2) classify the WSIs with the pretrained encoder from the first stage.

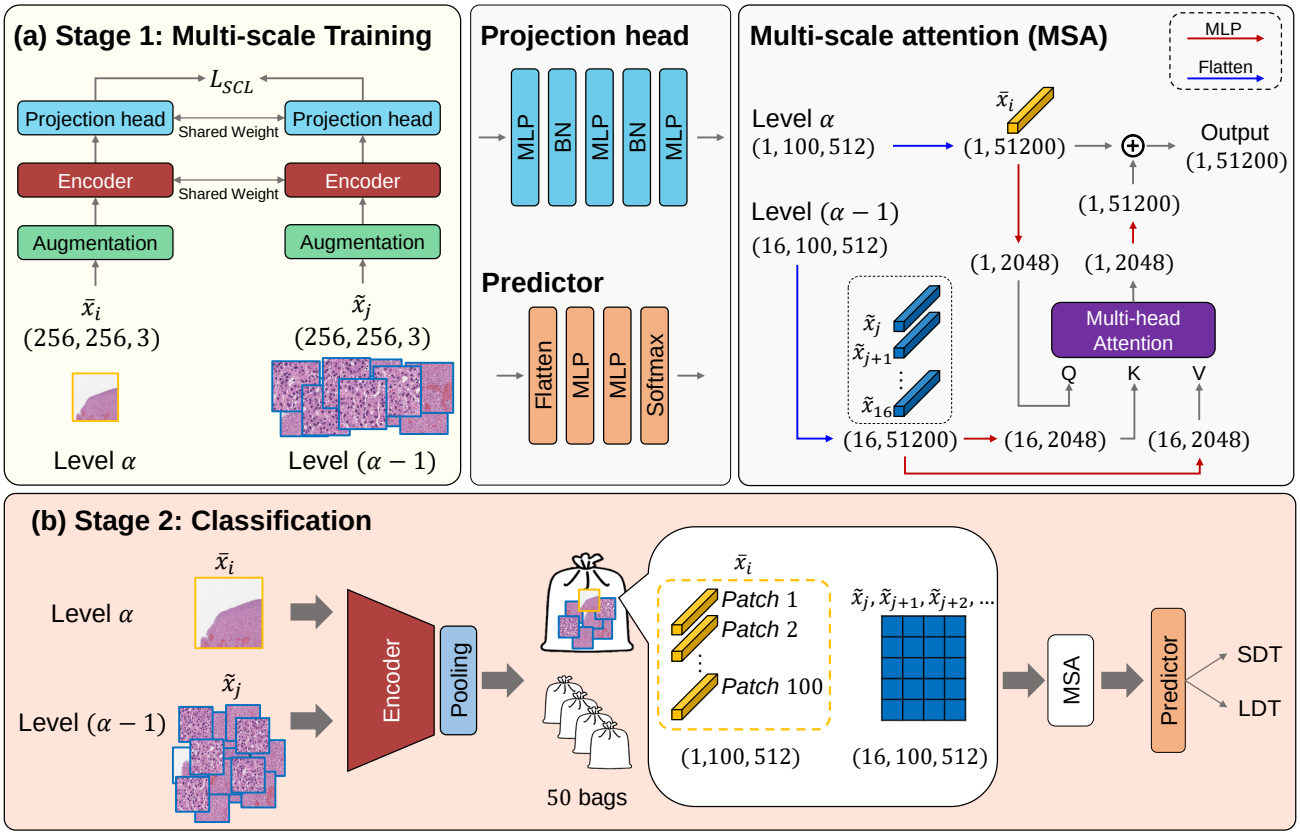

Figure 3. The overall framework of our proposed method. (a) Multi-scale training network. (b) Classification stage.

3.1 Data Pre-processing

We extracted the patches in $256 \times 256 \times 3$ dimension from each WSI from three different scales: Level 0, with a magnification scale of $20\times$; Level 1, with a magnification scale of $10\times$; and Level 2, with a magnification scale of $5\times$. Patches with less than 50% of tissue coverage were discarded.

3.2 Stage 1: Multi-scale Training using Contrastive Learning

As shown in Fig. 3(a), we adopted a Siamese network as a baseline architecture for our multi-scale encoder network, in which both the sister networks were made from the same architecture sharing the same parameters (the weights of both networks are updated at the same time). Data augmentations such as random flip, rotation, and contrast are first applied to the input images before feeding into the backbone network. Because the WSIs are composed of multiple magnification scales, crop-and-scale augmentation is not used. Each sister network consists of a backbone network (VGG16[16]) and a projection head, which is made up of two sets of a dense layer with 1024 neurons and a batch normalization layer, followed by a dense layer with 256 neurons. The outputs from both sister networks are the feature vectors in 256-dimension, used for the contrastive loss computation.

To make use of the multi-scale nature of WSIs, the model takes two different inputs from the same spatial location. For example, if we take an image patch from Level α (i.e., the main patch), then its corresponding patches in the higher magnification image (Level $(\alpha - 1)$) are taken. Then, positive pairs are made using patches from the same location but at different scales or patches at the same scale but from different locations in the same WSI. Negative pairs are made from patches from different classes but at the same scale as the main patch. The purpose of creating the input to the encoder in this way is to make the model maximize not only the similarities of the patches at a different scale from the same location but also the similarities of patches from the same WSI and to minimize the similarities of the patches from different classes.

For encoder training, we adopted supervised contrastive learning (SCL)[17] in our proposed framework. The main difference between SCL and the general self-supervised contrastive learning is that it leverages additional class label information to ensure that the feature distance between the elements from the same class becomes small.

The SCL loss is formulated as follows:

$$L_{SCL} = \sum_{i=1}^{M} -\frac{1}{M_{y_i} - 1} \sum_{j=1}^{M} \mathbb{1}_{\bar{y}_i = \bar{y}_j} \cdot log \frac{exp(h_i \cdot h_j / \tau)}{\sum_{k=1}^{M} exp(h_i \cdot h_k / \tau)} \quad ()$$

where M indicates the number of samples in a mini batch, h_i is the feature vector (256-dimensional) of the main patch from Level α within a mini batch, h_j is the feature vector of the corresponding patches from Level $(\alpha - 1)$ or the patches from the same WSI as the main patch, and h_k is the feature vector from other classes based on the label $\bar{y}$. The label $\bar{y}$ is the pair label, where 0 is assigned to negative pairs and 1 is assigned to positive pairs. Furthermore, τ is a scalar temperature parameter. $(\cdot)$ is the inner dot product between the L2 normalized projected feature of two feature vectors.

The initial learning rate is set to 1e-4 along with the exponential decay learning schedule with a decay rate of 1e-6. Stochastic gradient descent (SGD) is employed as the optimizer. We train the models with 50 epochs, and the models with the lowest validation loss are then saved to use in the downstream tasks.

3.3 Stage 2: Classification using Multi-scale Attention

In stage two, the encoder of the pretrained multi-scale model is then used in the downstream task. Our dataset is a weak label dataset, in which only image-level labels are provided. Therefore, multiple instance learning (MIL) is adopted to classify our dataset in the way of treating WSIs as bags that consist of a small number of patches (instances). MIL assumes that a positive bag should include at least one positive instance, whereas a negative bag should be made up of negative instances only. We randomly generate 50 bags for each WSI, and each bag consists of the latent feature vectors (i.e., the output of the pretrained multi-scale encoder) of 100 patches (see Fig. 3(b)). We generate the bags with the extracted features of the Level α patches along with their corresponding patches from Level $(\alpha - 1)$. After the bag formation, we classify the bags via a predictor which composes of two multi-layer perceptrons (MLPs), each with 512 neurons and has ReLU activation in between followed by a softmax layer before the output. For the final results, because 50 different bags are generated from each WSI, the majority voting is used to obtain the final predicted class for each WSI. Binary cross-entropy loss along with SGD as the optimizer is used in training the classification model.

3.3.1 Multi-scale attention (MSA)

To leverage the benefit of the multi-scale nature of WSI, we introduce a multi-scale attention (MSA) module before the predictor module to exploit the relationship between multi-scale features. MSA module is a bottleneck residual that has a multi-head attention (MHA) transformer layer[11] in the middle (see Fig. 3 upper right). Unlike the original self-attention module, which generates query, key, and value from the same input, our MSA module accepts two inputs from different scales (Level α and $(\alpha - 1)$). Specifically, after flattening the two last axes of the two inputs, we have a token from the main patch at Level α and 16 tokens from 16 corresponding patches at Level $(\alpha - 1)$. In the MHA layer, we assign the Level $(\alpha - 1)$ feature as the key and value, whereas the Level α feature is assigned as the query. By using this scheme, with the coarse view in Level α, we can inquire the meaningful information from the finer view in the level $(\alpha - 1)$ instead of combining them all and including the self-attention mechanism. To reduce the computation cost in the MHA layer, we use the MLP layers to reduce the size of the token's feature to 2048, which means, in the MHA layer, there are 4 heads with $d_{model} = 2048$ and $d_k = 128$. See the MHA design by Vaswani et al.[11] for more details. The output from the MHA layer is enlarged and aggregated with the flattened Level α feature to form a residual connection.

Table 1. The classification performance on single scale dataset.

Area	Level 0	Level 1	Level 2
WSI	0.6364	0.5879	0.5669
Tumor	0.7250	0.7441	0.7086

Table 2. The ablation study of classification performance with the proposed multi-scale pretrained model as the feature extractor. (MS: Multi-scale, CL: Contrastive Learning, MSA: Multi-scale Attention. Note that the conventional (baseline) multi-scale model has two encoders, whereas our multi-scale model has only one encoder to leverage two different scales.)

Method	Area	Level (1+0)	Level (2+1)	Encoder Size
Conventional MS	WSI	0.6081	0.5915	98M
	Tumor	0.7398	0.7200	
MS + CL (Ours)	WSI	0.5882	0.6250	49M
	Tumor	0.7331	0.7026	
MS + CL + MSA (Ours)	WSI	0.6415	0.5993	
	Tumor	0.7715	0.7166	

4. RESULT

4.1 Dataset

We used an IHCC dataset consisting of 332 WSIs in total, collected from 168 patients in Seoul National University hospital. IHCC is categorized into two subtypes: (1) small duct type (SDT) and (2) large duct type (LDT). The classification of the duct type is highly related to the KRAS and IDH mutated genes. Based on the gene mutation information, we select the WSIs with the wild type in the KRAS gene and mutated type in the IDH gene, and the mutated type in the KRAS gene and wild type in the IDH gene, for the training set of the SDT and LDT, respectively. For inference on the test set, we collected patches using two different strategies: one is random sampling on the entire WSI and the other is random sampling only within the tumor area (region annotations are provided).

4.2 Experimental Result

In this section, we compare the performance (accuracy measured over 10 iterations) of different models we tested. We first conducted the single-scale experiments on three different scales (i.e., each model is trained using single-scale data) and two different areas by using a pretrained VGG16 model to extract features for bag generation. These bags were then classified with a predictor with same parameter setting as mentioned in Section 3.3. We can see that the bags from the tumor area always outperformed those from the entire WSI by about 9% to 14% (see Table 1).

For the WSI area, expectedly, the classification accuracy decreases as the magnification level (scale) lowers. This is because the patches from the lower magnification level carry less information than those from higher magnification level. On the other hand, the bags from the tumor area performed the best at the magnification Level 1 rather than Level 0. Therefore, we can say that the highest level is not always the best for classification even if it could provide the most detailed information. Among the different combinations of scale and area, the best combination is the bags from the tumor area from Level 1, with an accuracy of 0.74.

Next, we conducted an ablation study of comparing multi-scale methods by adding the proposed components one by one, starting from a baseline multi-scale method without contrastive learning (see Table 2). The baseline conventional multi-scale model consists of multiple single scale models each of which is trained independently with the SCL loss. Then, the features were extracted from multiple single-scale models and concatenated to form the bags. The model size of the conventional multi-scale encoder is composed of 98 million parameters, which is twice the size of our proposed multi-scale encoder, with only 48 million parameters. Unlike the encoders in the conventional multi-scale model trained separately for each specific scale, our encoder is trained in a single Siamese model. Note that our proposed MS + CL model performed comparably to the conventional multi-scale models with only a half-size encoder due to the proposed multi-scale contrastive learning model. Each single patch from level α comprises 16 patches from level (α-1), but not all 16 patches would carry the representative features; hence, we proposed the MSA to further improve the classification of duct types. As shown in Table 2,

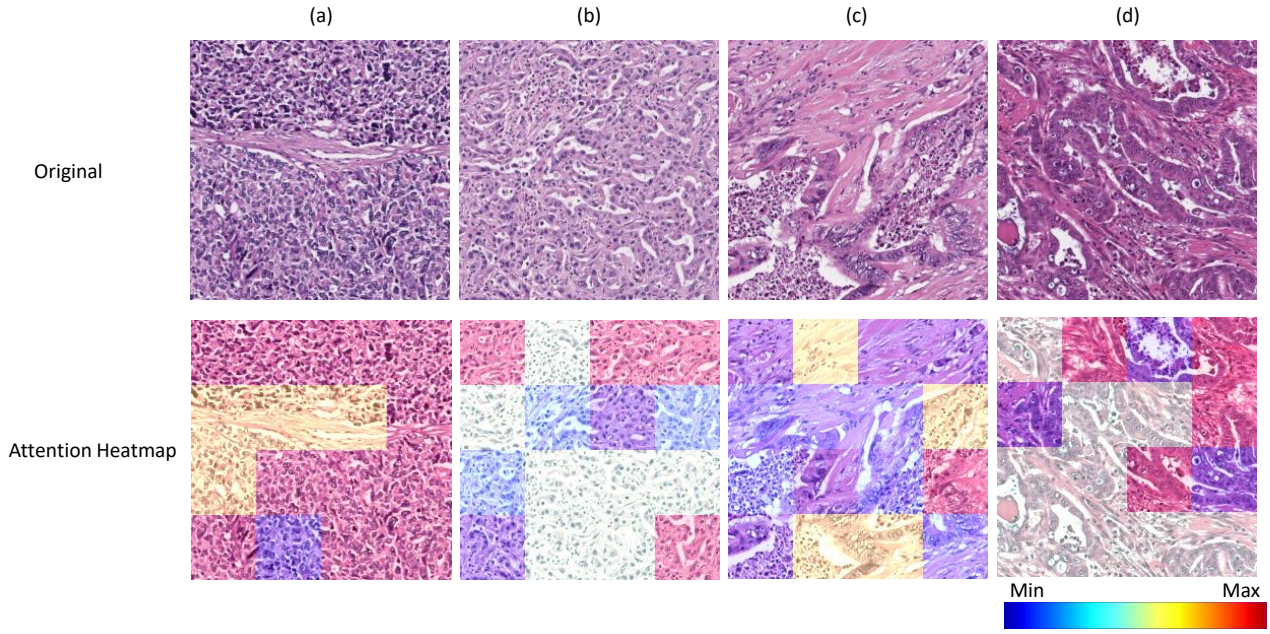

Figure 4. Each patch from Level $(\alpha - 1)$ showed different weight from MSA model. (a-b) are the patches from Small duct type and (c-d) are the patches Large duct type.

the MSA models outperformed both conventional multi-scale models and our proposed models without MSA except for the Level $(2 + 1)$ case; this is in line with what we observed in the single scale experiments that Level 0 and 1 show better performance than Level 2 (see Table 1). The models of Level $(1 + 0)$ showed about 3% to 6% performance improvement as compared with the conventional multi-scale model and MS + CL. Finally, we observed that multi-scale models always performed better than single-scale models as expected. The Level 1 models improved 5% and 3% with the help of the MS + CL + MSA model in WSI and tumor areas, respectively. For the Level 2 models, it increased 3% in the WSI area and 1% in the tumor area with the multi-scale model. The weights of 16 patches from Level (α-1) have been extracted from the MSA model and illustrated as the heatmap as showed in Fig. 4. We can see that the different patch from level (α-1) showed different weight in the heatmap even they are from the same patch in level α.

5. CONCLUSION

In this study, we introduced a unified multi-scale model, in which the multi-scale WSIs are trained with a single model instead of training multiple single-scale models. The proposed multi-scale model is about half of the conventional multi-scale approach. Our results showed that our proposed multi-scale model yielded comparable results to the conventional multi-scale model and helped in reducing the computational cost. Furthermore, we propose the MSA module to attend to the multi-scale information to further improve the performance of the duct type classification in IHCCs. In future work, we plan to extend our MSA to handle more than two scales. Conducting in-depth analysis of constrastive learning and attention in the context of histopathology image analysis is another future research direction.

ACKNOWLEDGMENTS

This study was approved by the institutional review board of Seoul National University Hospital (IRB NO.H-1011-046-339). This work was partially supported by the National Research Foundation of Korea (NRF-2019M3E5D2A01063819, NRF-2021R1A6A1A13044830), the Institute for Information & Communications Technology Planning & Evaluation (IITP-2023-2020-0-01819), the Korea Health Industry Development Institute (HI18C0316), and a Korea University Grant.

REFERENCES

[1] Gao, Z., Shi, J., Zhang, X., Li, Y., Zhang, H., Wu, J., Wang, C., Meng, D., and Li, C., "Nuclei grading of clear cell renal cell carcinoma in histopathological image by composite high-resolution network," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 132–142, Springer (2021).

[2] Li, J., Lin, T., and Xu, Y., "Sslp: Spatial guided self-supervised learning on pathological images," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 3–12, Springer (2021).

[3] Zhang, C., Song, Y., Zhang, D., Liu, S., Chen, M., and Cai, W., "Whole slide image classification via iterative patch labelling," in [*2018 25th IEEE International Conference on Image Processing (ICIP)*], 1408–1412, IEEE (2018).

[4] Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., and Takeuchi, I., "Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 3852–3861 (2020).

[5] Li, B., Li, Y., and Eliceiri, K. W., "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 14318–14328 (2021).

[6] Marini, N., Otálora, S., Podareanu, D., van Rijthoven, M., van der Laak, J., Ciompi, F., Müller, H., and Atzori, M., "Multi_scale_tools: a python library to exploit multi-scale whole slide images," *Frontiers in Computer Science* **68** (2021).

[7] Liu, Q., Louis, P. C., Lu, Y., Jha, A., Zhao, M., Deng, R., Yao, T., Roland, J. T., Yang, H., Zhao, S., et al., "Simtriplet: Simple triplet representation learning with a single gpu," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 102–112, Springer (2021).

[8] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al., "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in Neural Information Processing Systems* **34** (2021).

[9] Gao, Z., Shi, J., and Wang, J., "Gq-gcn: Group quadratic graph convolutional network for classification of histopathological images," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 121–131, Springer (2021).

[10] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G., "A simple framework for contrastive learning of visual representations," in [*International conference on machine learning*], 1597–1607, PMLR (2020).

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., "Attention is all you need," *Advances in neural information processing systems* **30** (2017).

[12] Huang, Z., Chai, H., Wang, R., Wang, H., Yang, Y., and Wu, H., "Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 561–570, Springer (2021).

[13] Kalra, S., Adnan, M., Hemati, S., Dehkharghanian, T., Rahnamayan, S., and Tizhoosh, H. R., "Pay attention with focus: A novel learning scheme for classification of whole slide images," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 350–359, Springer (2021).

[14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N., "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR* (2021).

[15] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows," *International Conference on Computer Vision (ICCV)* (2021).

[16] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).

[17] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D., "Supervised contrastive learning," *Advances in Neural Information Processing Systems* **33**, 18661–18673 (2020).