AutoEncoder-Based Feature Transformation with Multiple Foundation Models in Computational Pathology

Woojin Chung[1], Yujun Park[2], and Yonnho Nam[1]

[1] Deparment of Biomedical Engineering, Hankuk University of Foreign Studies, Yongin-si 17035, Gyeonggi-do, Korea.
{goglxych97,yoonhonam}@hufs.ac.kr
[2] Department of Pathology, CHA Bundang Medical Center, CHA University, Seongnam-si 13496, Gyeonggi-do, Korea.
isutar_star@naver.com

Abstract. The performance of deep learning models is highly dataset-dependent. Pretrained models on large-scale datasets have significant advantages in understanding general patterns by leveraging large volumes of data. Some of these models, which are adaptable to a wide range of downstream tasks, are referred to as foundation models. Recently, several foundation models have been published in the field of computational pathology, recognized for their potential to advance deep learning applications in several downstream tasks. In order to effectively utilize multiple foundation models, each with its own advantages, it is crucial to effectively summarize or ensemble their advantages. In this paper, we propose a feature transformation method for the effective utilization of features from multiple foundation models using an autoencoder-based architecture. This method facilitates the extraction of integrated features from multiple foundation models, enabling more generalized training. We demonstrated that the proposed approach resulted in more robust representations for out-of-distribution datasets in our patch-level classification tasks.

Keywords: Computational Pathology · Foundation Model · Feature Extraction · Deep Learning · AutoEncoder

1 Introduction

The performance of deep learning models highly depends on the training datasets. Models pretrained on large-scale, high-quality datasets not only provide a strong starting point but also help in understanding general patterns, aiding in finding optimal convergence points for various tasks. These models, depending on the data spectrum they were trained on and their intended use, are called foundation models [17]. In medical imaging, where data access is limited due to patient privacy concerns and high data collection costs, foundation models can be effectively utilized. They leverage pretrained information, minimizing the need

for direct data access, reducing data acquisition costs, and safeguarding patient privacy. This highlights the growing importance of foundation models in the medical domain [6, 17].

Whole Slide Images (WSIs), which are often gigapixel in size, are difficult to handle as input for image processing models. Additionally, generating annotations for these images is expensive and labor-intensive work. Foundation models are considered to address these conventional challenges, and several foundation models have recently been published in the field of computational pathology [18]. The performance of these models have demonstrated on multiple benchmark datasets, proving their robustness and generalizability [7, 8]. With advancements in technology and resources, the prevalence of these models is expected to continue increasing. Therefore, the strategic use of foundation models should be investigated to advance computational pathology.

Foundation models have been trained on different datasets using various methodologies, resulting in unique strengths for each. To leverage these strengths, it is essential to have a method that can effectively summarize or ensemble these advantages. This approach is particularly useful in fields like computational pathology, where capturing the vast array of universal patterns is challenging. By using multiple foundation models information, it is possible to achieve more robust and general representations. This can lead to more accurate and reliable support diagnostic tools.

However, obtaining integrated information from multiple foundation models in computational pathology involves several considerations. First, handling the massive size of WSIs for both training and inference requires significant time. As foundation models grow in size, using multiple models in parallel without proper tuning poses problems in terms of computing resources and practical application. Additionally, since these models were trained by different institutions, the resulting features have varying properties, making it difficult to determine their correlations. This diversity complicates the task of achieving optimal performance, necessitating effective integration methods. Finally, while there may be an optimal combination of models for specific datasets and tasks, manually selecting these combinations is impractical due to time and resource constraints. Therefore, developing a robust and efficient method for integrating the knowledge from multiple foundation models is necessary.

There have been various model fusion strategies [10] in the field of deep learning to effectively utilize and integrate information from multiple models, such as multi-teacher student knowledge distillation [11, 19] and model ensembling [13]. Multi-teacher student knowledge distillation involves transferring knowledge from multiple teacher models to a single student model, thereby capturing common representations and improving the robustness of the student model. Model ensembling aggregates models using various methods to achieve better performance compared to a single model. These methods aim to enhance learning ability by improving generalization and robustness through the use of multiple pretrained models.

In this paper, we propose a feature transformation method that extracts integrated representations from multiple foundation models using an autoencoder-based architecture [9]. This approach extracts transformed features highly correlated to the original features from each of these multiple models. Using this transformed information, we conducted patch-level classification tasks in pathology and compared the performance of the proposed method with those from individual foundation models, particularly on out-of-distribution data.

2 Method

2.1 Used Pretrained Models

There are several pretrained models available in computational pathology. For our method, we utilized four models: three self-supervised pretrained models, each with a different architecture [1,3,16] and one task-specific pretrained model [15]. All of these models are trained at the patch-level. Due to being trained on vast datasets using self-supervised methods, the three self-supervised models are also classified as foundation models. Table 1 provides details of the models used.

Table 1. Details of the pretrained models used for the proposed method.

Model Name	Model Architecture	Trained Method	Trained Dataset
MONAI pathology tumor detection [15]	ResNet18 [5]	Task-specific Learning	Camelyon16
Ciga et al. [3]	ResNet18 [5]	SimCLR [2]	TCGA, TUPAC16, CPTAC, SLN-Breast Camelyon16,17
CTransPath [16]	Swin-T [12]	SimCLR [2]	TCGA, PAIP
UNI [1]	ViT-L [4]	DINOv2 [14]	Private Dataset

2.2 Proposed Method for Feature Transformation

In computational pathology, foundation models pretrained at the patch-level receive patches extracted from WSIs as input and output features. Each foundation model produces unique representations in its features due to differences in training methods and model architectures. These original features, after passing through the models, differ in both the information contained in their dimensions and their shapes. To harmonize this information, we utilized the capability of autoencoders to re-arrange dimensional information in the internal latent space during the training process. By using autoencoders, features was compressed and reconstructed, effectively filtering out noise and preserving essential information.

The overview of the proposed method is illustrated in Fig. 1. Our proposed method consists of the following two steps:

In the first step, we used an autoencoder-based architecture, as shown in Fig. 1(a), to train the reconstruction of features that had passed through several foundation models. The encoder of the autoencoder effectively compressed the input data into its latent space. While training autoencoders to reconstruct the original representations from each foundation model, we imposed a constraint using the cosine similarity loss function to increase the correlation at the bottlenecks formed after passing through each encoder. This training process involved applying both the reconstruction loss for the features and the cosine similarity loss in the latent space. As the cosine similarity loss converged, these encoders explored the common representations shared by all the foundation models and constructed integrated latent spaces in the bottlenecks. We expect these highly correlated spaces to have learned a more robust representation from multiple pretrained models and to perform better on out-of-distribution data than a single foundation model.

In the second step, we used the transformed features from the trained encoder to conduct downstream classification tasks. The transformed features, compared to the original features, incorporated information from other foundation models. To demonstrate the effectiveness of the proposed method, we compared the results of training on the original features provided by the foundation model with those from training on the transformed features.

2.3 Evaluation Tasks

We evaluated our method on two tasks in computational pathology: (1) predicting lymph node metastasis in early gastric cancer and (2) cancer region segmentation in WSIs. All training steps were performed at the patch-level using 224×224 pixel size patches at 10× magnification. Since each pretrained model requires different input normalization, the patches were transformed accordingly for each model. To confirm whether our method produces more robust representations, we included both internal and external cases for each task. All results were compared across three scenarios: scratch learning using ResNet18 [5], transfer learning with the promising foundation model named UNI, trained with the DINOv2 framework [1,14], and our method, which involves using the transformed features from the original features of this foundation model. Transfer learning was performed using linear probing with initialized layers.

For predicting lymph node metastasis in early gastric cancer, we conducted classification on patch-level metastasis using the labels from the slide-level annotation. Afterward, we observed the probability for each patch in the slides and then calculated the Area Under the ROC Curve (AUC) value for the average probability at the slide-level. Quantitative evaluations were performed on both internal and external datasets.

Gastric cancer region segmentation was trained at the patch-level. Patches were extracted from WSIs annotated at the slide-level, and a classification task was performed to determine if each patch contained a cancer region. Quantitative

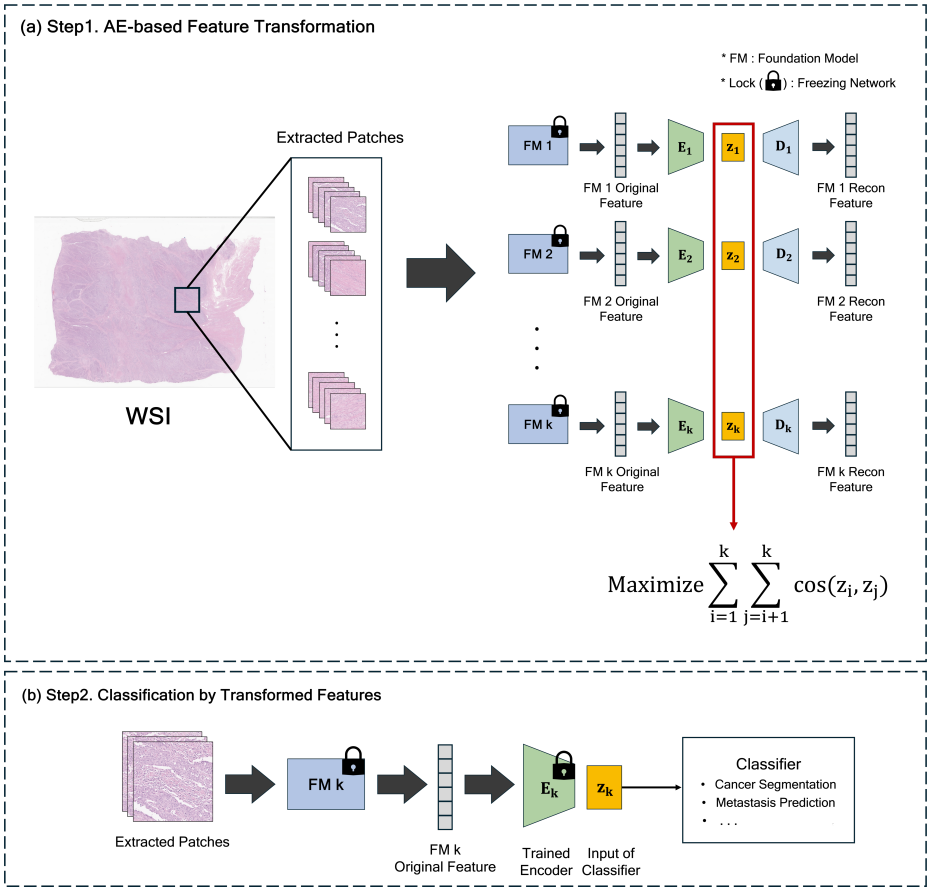

Fig. 1. Overview of the proposed method. This method consists of the following two steps. (a) Step 1: patches extracted from WSIs were used to train the reconstruction of the original features from each foundation model using autoencoders. During this process, cosine similarity loss function was employed to increase the correlation of the latent spaces at the bottleneck. (b) Step 2: after the autoencoders converged, we selected one foundation model and its corresponding encoder to conduct the classification task using the transformed features.

evaluation was performed only on the internal dataset, as gold standard labels for the external dataset were not available.

3 Result

3.1 Datasets

Lymph Node Metastasis in Early Gastric Cancer We used 40× magnification hematoxylin and eosin stained WSIs from our institution for our study. The training dataset consisted of 200 WSIs, with 100 slides from patients with lymph node metastasis (LNM) and 100 slides from patients without LNM. These slides were sourced from patients with early gastric cancer who underwent curative surgical resection with lymph node dissection. The internal validation dataset included an additional 60 surgical cases of early gastric cancer from our institution, comprising 30 cases with LNM and 30 cases without LNM. The external validation dataset included 46 endoscopic resection cases with additional lymph node dissection from external sources, comprising 23 cases with LNM and 23 cases without LNM.

Cancer Region Segmentation An expert pathologist annotated the cancer regions at the slide-level for 80 gastric cancer WSIs from our institution. From this dataset, 64 slides were used for training and 16 slides for internal testing. Additionally, we obtained 20 gastric cancer slides from other institution as an external dataset and performed cancer region segmentation to observe the trends in the results.

3.2 Lymph Node Metastasis Prediction in Early Gastric Cancer

We evaluated the performance of lymph node metastasis prediction in early gastric cancer for each method on both internal and external datasets. The results are shown in Fig. 2(a). For scratch learning, the AUC values for the internal and external datasets were 0.46 and 0.36, respectively, indicating overfitting due to the insufficient number of datasets, resulting in poor prediction performance. Our proposed method and the foundation model using transfer learning achieved AUC values of 0.79 and 0.77 on the internal dataset, respectively. However, on the external dataset, there was a significant difference, with AUC values of 0.71 for our method and 0.63 for the transfer learning.

3.3 Cancer Region Segmentation

The results of cancer region segmentation for the internal dataset are shown in Fig. 2(b). Both the transfer learning with foundation model and our proposed method outperformed scratch learning, with accuracy scores of 0.984 and 0.984 compared to 0.966, respectively. Additionally, our proposed method demonstrated more robust outcomes with less variance.

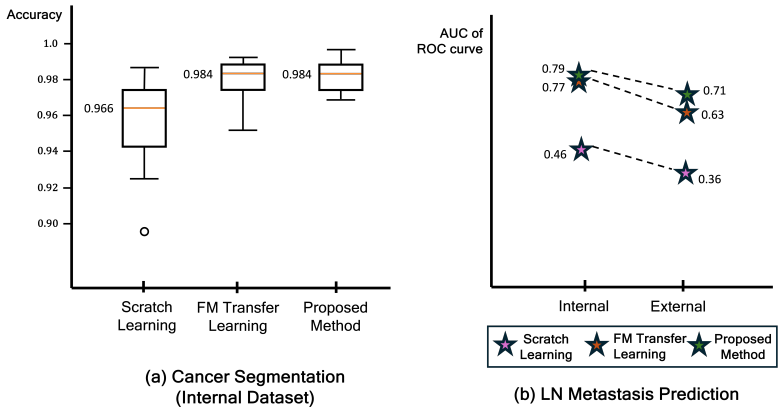

Fig. 2. (a) AUC values for lymph node metastasis prediction in early gastric cancer using internal and external datasets. (b) Box plots of accuracy for cancer segmentation in the internal dataset.

We examined the segmentation results for each patch in both the internal and external datasets, with examples shown in Fig. 3. The results from the internal dataset showed no significant difference between our proposed method and the foundation model with transfer learning. However, compared to both methods, scratch learning exhibited several false positive regions. In the external dataset, the trained scratch learning model failed to predict most of the regions. When comparing the results with and without feature transformation, the predictions made using the original features showed some false positive regions in cancer detection. As shown in Fig. 3(b), the results of transfer learning without our feature transformation method predicted muscle tissue and lymphocyte patches as cancer. Additionally, we observed that our method improved the detection of patterns in the boundary areas of WSIs and ulcer patches in our out-of-distribution segmentation.

3.4 Ablation Study on Proposed Method

An ablation study was conducted to determine whether the improved generality of the proposed method is merely attributable to the large model size. In the experiment for lymph node metastasis prediction in early gastric cancer, the same deep layers used for feature transformation and the downstream task were applied to the original features, and results were observed for each pretrained model. Evaluated based on the mean probability at the slide level, the AUC values showed that the method produced more robust representations on the external dataset compared to when the feature transformation was excluded. The AUC values increased across all foundational models, suggesting that the improved generalization is due not only to the model size but also to the ef-

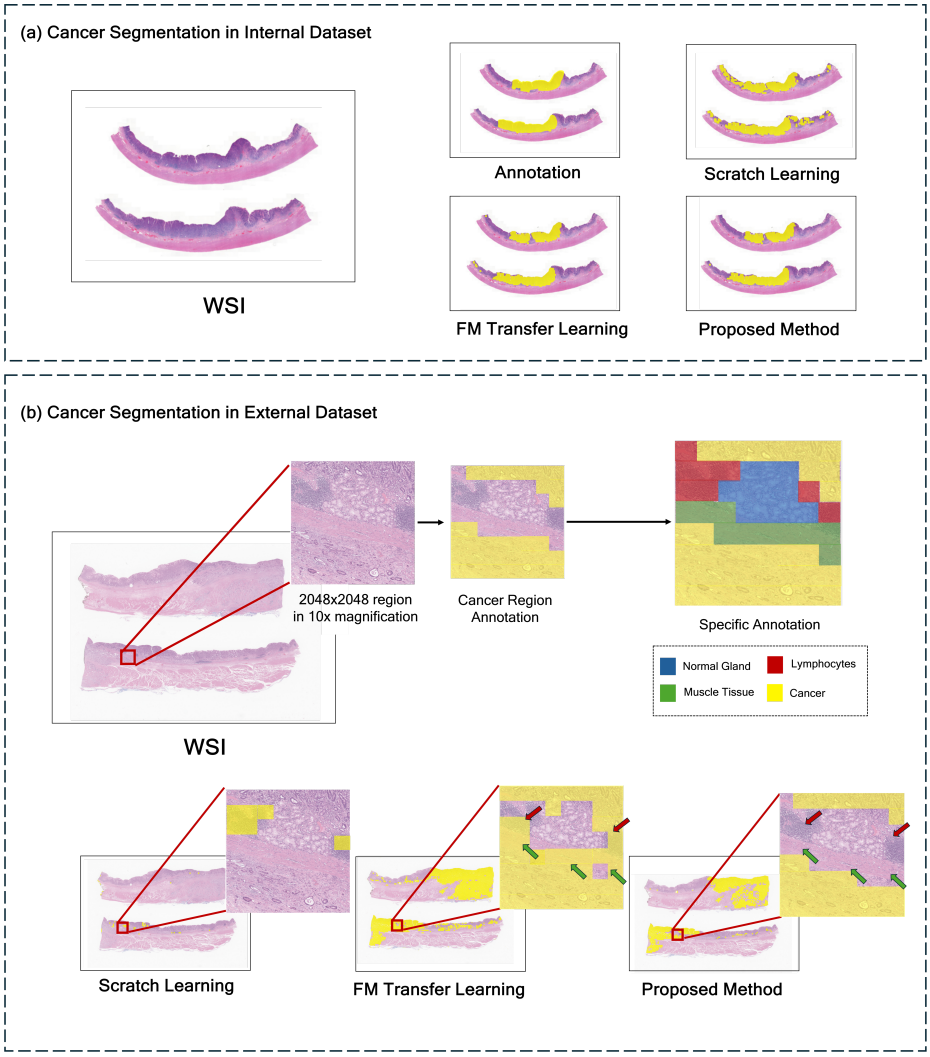

Fig. 3. Representative examples of cancer segmentation at the slide-level for scratch learning, the foundation model with transfer learning, and our proposed method. (a) Examples from the internal dataset. (b) Examples from the external dataset. The red arrow indicates the lymphocyte regions, and the green arrow indicates the muscle tissue regions. Our proposed method showed a reduction in the false positive areas around these patches.

fectiveness of the feature transformation approach. The results are shown in Table 2.

Table 2. The ablation study results for Lymph Node Metastasis Prediction in Early Gastric Cancer.

Model Name	**AUC w/o Proposed Method**	**AUC w/ Proposed Method**
MONAI pathology tumor detection [15]	0.63	0.64
Ciga et al. [3]	0.56	0.61
CTransPath [16]	0.61	0.67
UNI [1]	0.63	0.71

4 Discussion and Conclusion

In this paper, we propose a method to capture the common features of multiple foundation models. By leveraging information from several foundation models, our method can help find a more integrated representation, providing more robust results on both internal and external datasets. Processing giga-pixel sized WSIs, our method did not significantly increase the inference time compared to using a single model, as it only used one foundation model and its corresponding trained encoder (as shown in Fig. 1(b)). Additionally, our method can be flexibly applied not only to the foundation models used in this study but also to other pretrained networks suitable for various downstream tasks. This approach can be helpful when working with limited data, as is often the case in the medical domain.

There are some considerations in our study. First, our approach must be rigorously validated, as it was tested within the limited context of a few tasks in computational pathology. It should be evaluated across various domains and benchmark datasets. Additionally, training the autoencoders for feature transformation requires sufficient data; using a small dataset may lead to overfitting and reduced performance. Furthermore, applying our method to a large number of foundation models imposes more constraints, potentially requiring significant time and computing resources to converge.

References

1. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)

2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (02 2020)
3. Ciga, O., Martel, A., Xu, T.: Self supervised contrastive learning for digital histopathology (11 2020)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv **abs/2010.11929** (2020), https://api.semanticscholar.org/CorpusID:225039882
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2015), https://api.semanticscholar.org/CorpusID:206594692
6. He, Y., Huang, F., Jiang, X., Nie, Y., Wang, M., Wang, J., Chen, H.: Foundation model for advancing healthcare: Challenges, opportunities, and future directions (2024)
7. kaiko.ai, Gatopoulos, I., Känzig, N., Moser, R., Otálora, S.: eva: Evaluation framework for pathology foundation models. In: Medical Imaging with Deep Learning (2024), https://openreview.net/forum?id=FNBQOPj18N
8. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3344–3354 (June 2023)
9. Kramer, M.A.: Autoassociative neural networks. Computers & Chemical Engineering **16**, 313–328 (1992), https://api.semanticscholar.org/CorpusID:62207837
10. Li, W., Peng, Y., Zhang, M., Ding, L., Hu, H., Shen, L.: Deep model fusion: A survey. ArXiv **abs/2309.15698** (2023), https://api.semanticscholar.org/CorpusID:262942062
11. Liu, Y., Zhang, W., Wang, J.: Adaptive multi-teacher multi-level knowledge distillation (03 2021)
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9992–10002 (2021), https://api.semanticscholar.org/CorpusID:232352874
13. Mohammed, A., Kora, R.: A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University - Computer and Information Sciences **35** (02 2023). https://doi.org/10.1016/j.jksuci.2023.01.014
14. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.Q., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y.B., Li, S.W., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. ArXiv **abs/2304.07193** (2023), https://api.semanticscholar.org/CorpusID:258170077
15. Team, M.: Pathology tumor detection (2022), https://monai.io/model-zoo.html
16. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis **81**, 102559 (07 2022). https://doi.org/10.1016/j.media.2022.102559
17. Zhang, S., Metaxas, D.: On the challenges and perspectives of foundation models for medical image analysis. Medical Image Analysis **91**, 102996 (10 2023). https://doi.org/10.1016/j.media.2023.102996

18. Zhang, Y., Gao, J., Tan, Z., Zhou, L., Ding, K., Zhou, M., Zhang, S., Wang, D.: Data-centric foundation models in computational healthcare: A survey. arXiv preprint arXiv:2401.02458 (2024)
19. Zuchniak, K.: Multi-teacher knowledge distillation as an effective method for compressing ensembles of neural networks (02 2023)