Multi-modal Pathological Pre-training via Masked Autoencoders for Breast Cancer Diagnosis

Mengkang Lu, Tianyi Wang, and Yong Xia(✉)

National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
yxia@nwpu.edu.cn

Abstract. Breast cancer (BC) is one of the most common cancers identified globally among women, which has become the leading cause of death. Multi-modal pathological images contain different information for BC diagnosis. Hematoxylin and eosin (H&E) staining images could reveal a considerable amount of microscopic anatomy. Immunohistochemical (IHC) staining images provide the evaluation of the expression of various biomarkers, such as the human epidermal growth factor receptor (HER2) hybridization. In this paper, we propose a multi-modal pre-training model via pathological images for BC diagnosis. The proposed pre-training model contains three modules: (1) the modal-fusion encoder, (2) the mixed attention, and (3) the modal-specific decoders. The pre-trained model could be performed on multiple relevant tasks (IHC Reconstruction and IHC classification). The experiments on two datasets (HEROHE Challenge and BCI Challenge) show state-of-the-art results.

Keywords: Breast cancer · Hematoxylin and eosin staining · Immunohistochemical staining · Multi-modal pre-training

1 Introduction

Breast cancer (BC) is one of the most common malignant tumors in women worldwide and it causes nearly 0.7 million deaths in 2020 [26]. The pathological process is usually the golden standard approach for BC diagnosis, which relies on leveraging diverse complementary information from multi-modal data. In addition to obtaining the histological characteristics of tumors from hematoxylin and eosin (H&E) staining images, immunohistochemical (IHC) staining images are also widely used for pathological diagnoses, such as the human epidermal growth factor receptor 2 (HER2), the estrogen receptor (ER), and the progesterone receptor (PR) [22]. With the development of deep learning, there are a lot of multi-modal fusion methods for cancer diagnosis [6,7,20,21].

Recently, with the development of Transformer, multi-modal pre-training has achieved great success in the fields of computer vision (CV) and natural language

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 457–466, 2023.
https://doi.org/10.1007/978-3-031-43987-2_44

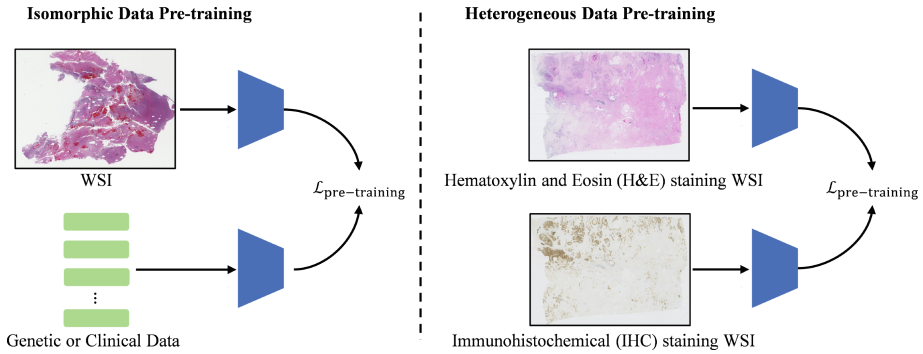

Fig. 1. Illustration of different multi-modal pre-training methods. The WSIs, genetic and clinical data from a patient could be used for isomorphic data pre-training. Pairs of H&E and IHC staining WSIs are used for heterogeneous data pre-training in our method.

processing (NLP). According to the data format, there are two main multi-modal pre-training approaches, as shown in Fig. 1. One is based on isomorphic data, such as vision-language pre-training [5] and vision-speech-text pre-training [3]. The other is based on heterogeneous data. Bachmann *et al.* [2] proposed Multi-MAE to pre-train models with intensity images, depth images, and segmentation maps. In the field of medical image analysis, it is widely recognized that using multi-modal data can produce more accurate diagnoses than using single-modal data. However, the development of multi-modal pre-training methods has been limited due to the scarcity of paired multi-modal data. Most methods focus on chest X-ray vision-language pre-training [8,11]. To our best knowledge, there is no work for multi-modal pre-training based on pathological heterogeneous data.

In this paper, we propose a multi-modal pre-training method based on masked autoencoders for BC downstream tasks. Our model consists of three parts, i.e., the modal-fusion encoder, the mixed attention, and the modal-specific decoder. We choose paired H&E and IHC (only HER2) staining images, which are cropped into non-overlapped patches as the input of our model. We randomly mask some patches by a ratio and feed the remaining patches into the modal-fusion encoder to get corresponding tokens. Then the mixed attention module is used to take the intra-modal and inter-modal correlation into account. Finally, we use modal-specific decoders to reconstruct the original H&E and IHC staining images respectively. Our contributions are summarized as follows:

- We propose a Multi-Modal Pre-training via Masked AutoEncoders MMP-MAE for BC diagnosis. To our best knowledge, this is the first pre-training work based on multi-modal pathological data.
- We evaluate the proposed method on two public datasets as HEROHE Challenge and BCI challenge, which shows that our method achieves state-of-the-art performance.

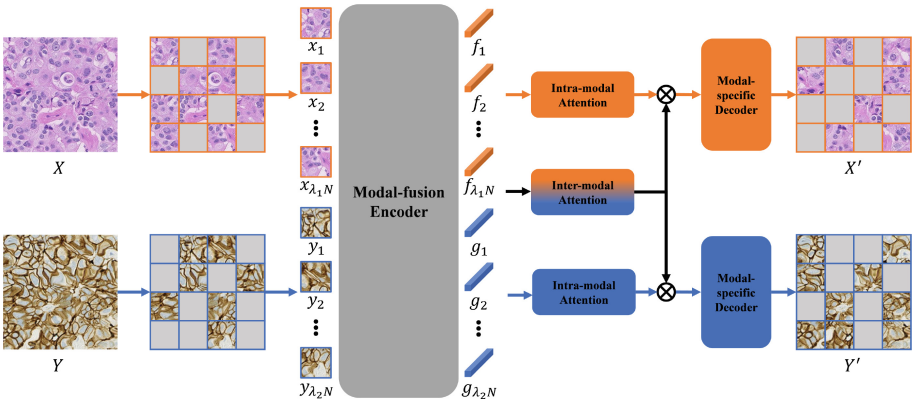

Fig. 2. Framework of our proposed MMP-MAE. A pair of images X and Y (H&E and IHC) are cropped into N non-overlapped patches, which are randomly masked by ratio λ_1 and λ_2. We feed the remaining patches $\{x_i\}_{i=1}^{\lambda_1 N}$ and $\{y_i\}_{i=1}^{\lambda_2 N}$ into the modal-fusion encoder to extract the patch tokens $\{f_i\}_{i=1}^{\lambda_1 N}$ and $\{g_i\}_{i=1}^{\lambda_2 N}$. Then we use intra-modal attention and inter-modal attention to take patch correlation into account. X' and Y' are reconstructed by modal-specific decoders respectively.

2 Method

2.1 Overview

The proposed MMP-MAE consists of three modules, i.e., the modal-fusion encoder, the mixed attention, and the modal-specific decoder, as shown in Fig. 2. A pair of H&E and HER2 images are cropped into regular non-overlapping patches. We mask some of the patches of two modalities with a ratio. The remained patches are fed into the modal-fusion encoder to get the corresponding tokens. Then we use the mixed attention module to extract intra-modal and inter-modal complementary information. Finally, the modal-specific tokens are fed into the modal-specific decoders to reconstruct the original H&E and HER2 images. The pre-trained modal-fusion encoder could be used for downstream tasks (e.g., HER2 status prediction and HER2 image generation based on H&E images).

2.2 MMP-MAE

Modal-Fusion Encoder. We use ViT-base [12] as the backbone of the modal-fusion encoder, which contains a linear projection, 12 transformer blocks, and a Multi-Layer Perceptron (MLP) head. We remove the MLP head and use the remained part to extract patch tokens. An image is cropped into several non-overlapping patches, and these patches are mapped to D dimension tokens with the linear projection and added position embeddings to retain positional information. Each transformer block consists of a multi-head self-attention layer (MHSA)

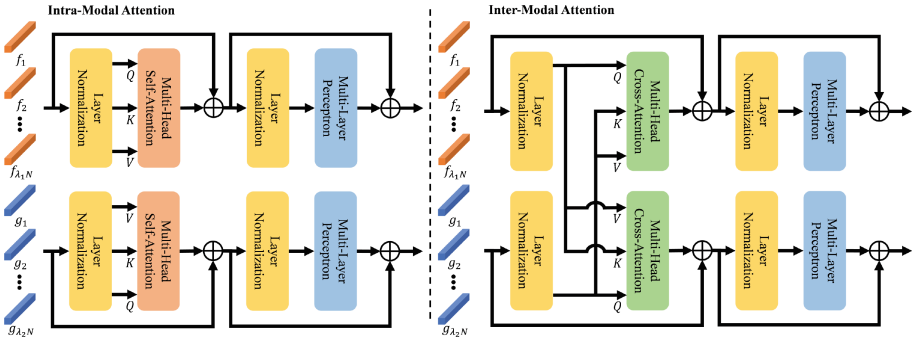

Fig. 3. Diagram of Intra-modal attention and inter-domain attention. The input of both attention modules is single-modal patch tokens. The intra-modal attention is the original transformer block, and there is no interaction between two modalities. We replace the MHSA with MHCA in the inter-modal attention to learn complementary information.

and a feedforward network (FFN). Layer normalization (LN) is applied before each layer, and the residual connection is used after each layer. The processing flow of ViT is shown in Alg. 1.

Mixed Attention. The mixed attention module contains intra-modal attention and inter-modal attention, as shown in Fig. 3. The intra-modal attention is the original transformer block, which consists of MHSA, MLP, LNs, and residual connections. It is defined as

$$A_x(F) = \text{softmax}(\frac{Q_x K_x^\top}{\sqrt{d}})V_x, \;\; A_y(F) = \text{softmax}(\frac{Q_y K_y^\top}{\sqrt{d}})V_y, \tag{1}$$

Algorithm 1. Transformer processing flow.

Input: A set of patches from one image $X = \{x_i\}_{i=1}^{N}$, $X \in \mathbb{R}^{N\times(R\times R\times C)}$

Transfer patches into linear embeddings
for $i = 1$ to N **do**
  $f_i \leftarrow \text{LP}(x_i)$, where $F = \{f_i\}_{i=1}^{N}$, $F \in \mathbb{R}^{N\times D}$
end for
Position encoding concatenation
$F_0 \leftarrow \text{Concat}\,(F_p, F)$, where $F_p \in \mathbb{R}^{1\times D}$
for $l = 1$ to L **do**
  $F_l^{'} \leftarrow \text{MHSA}(\text{LN}(F_{l-1})) + F_{l-1}$
  $F_l \leftarrow \text{FFN}(\text{LN}(F_l^{'})) + F_l^{'}$
end for

Output: Class token and patch tokens $F_l \in \mathbb{R}^{(N+1)\times D}$

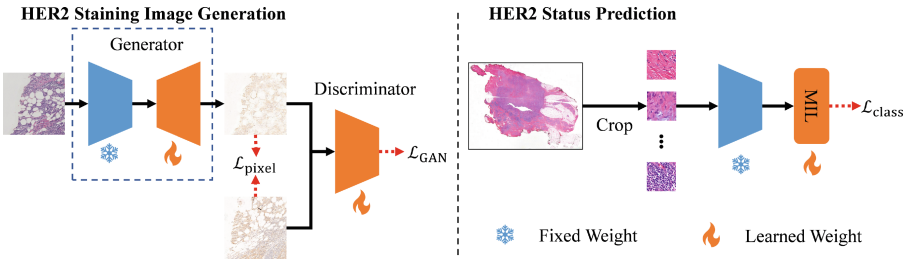

Fig. 4. Workflow of two downstream tasks. In the HER2 staining image generation task, we remain the structure of GAN and replace the generator with our pre-trained model. In the HER2 status prediction task, we replace the feature extractor with our pre-trained model to obtain representations with HER2 semantics.

In inter-modal attention, we replace MHSA with the multi-head cross-attention (MHCA) module. We use MHCA to leverage diverse complementary information between two modalities.

$$A_x(F) = \text{softmax}(\frac{Q_x K_y^\top}{\sqrt{d}})V_y,\ \ A_y(F) = \text{softmax}(\frac{Q_y K_x^\top}{\sqrt{d}})V_x, \tag{2}$$

Modal-Specific Decoder. Each modal-specific decoder is a shallow block with two transformer layers. Different from the transformer encoder, the target of the transformer decoder is used to reconstruct the original image.

Reconstruction Loss. Given a pair of H&E image X and HER2 image Y, which is cut into 16×16 non-overlapping patches $\{x_i\}_{i=1}^{N}$ and $\{y_i\}_{i=1}^{N}$. We mask some of the patches randomly with the ratio λ_1 and λ_2 ($\lambda_1 + \lambda_2 = 1$). The remained patches are fed into the modal-fusion encoder and the output is corresponding patch tokens $\{f_i\}_{i=1}^{\lambda_1 N}$ and $\{g_i\}_{i=1}^{\lambda_2 N}$. We randomly generate masked patch tokens $\{e_j^x\}_{j=1}^{(1-\lambda_1)N}$ and $\{e_j^y\}_{j=1}^{(1-\lambda_2)N}$, which are learnable vectors for masked patch prediction. The input of the mixed attention module is the full set of tokens $\{f_i, e_j^x\}_{i=1,j=1}^{i=\lambda_1 N, j=(1-\lambda_1)N}$ and $\{g_i, e_j^y\}_{i=1}^{i=\lambda_2 N, j=(1-\lambda_2)N}$, which include both the remaining patch tokens and the masked patch tokens. After the process of the mixed attention module, H&E and HER2 patch tokens are fed into the modal-specific decoders respectively to reconstruct the original H&E image X' and HER2 image Y'. The reconstruction loss is computed by the mean squared error between the original images X, Y and the generative images X', Y', which is computed as

$$\mathcal{L}_{\text{H\&E}} = \frac{1}{T_1}\sum_{i=1}^{T_1} \mid p_i - p_i' \mid^2,\ \ \mathcal{L}_{\text{HER2}} = \frac{1}{T_2}\sum_{i=1}^{T_2} \mid q_i - q_i' \mid^2 . \tag{3}$$

We use an adjustable hyperparameter θ to balance the losses of two modalities. The final loss $\mathcal{L}$ is defined as

$$\mathcal{L} = \theta \mathcal{L}_{\text{H\&E}} + (1 - \theta)\mathcal{L}_{\text{HER2}} \tag{4}$$

2.3 Downstream Tasks

The pre-trained encoder could be used for downstream tasks, as shown in Fig. 4. We choose two relevant tasks: HER2 image generation based on H&E images and HER2 status prediction. In the HER2 generation task, we replace the generator of Pyramid Pix2pix, a generative adversarial network (GAN) in [16], with our pre-trained encoder and a light-weight decoder. The weights of the pre-trained encoder are fixed, and the light-weight decoder in the generator and the discriminator are learnable. We use pairs of H&E and IHC images for GAN training. In the HER2 status prediction task, we replace the universal extractor ResNet-50 [14] with our pre-trained encoder. We use CLAM-MIL [19] as the aggregator in our training process.

3 Experimental Results

3.1 Datasets

ACROBAT Challenge. The AutomatiC Registration Of Breast cAncer Tissue (ACROBAT) Challenge [27] provides H&E WSIs and matched IHC WSIs (ER, PR, HER2, and KI67), which consists of 750 training cases, 100 validation cases, and 300 testing cases. We choose paired H&E and HER2 WSIs for pre-training. We extract the key points and descriptors from paired WSIs using SIFT [18] and SuperPoint [10]. Then the extracted key points and descriptors are matched using RANSAC [13] and SuperGlue [25]. We repeat this procedure several times on the rotated, downsampled, or transformed moving WSI to fetch the best transformation based on mean squared error (MSE) loss between source and target WSIs' descriptors. After that, the selected transformation is optimized across different levels of WSIs by gradient descent with local normalized cross-correlation (NCC) as its cost function. In the final phase of nonrigid registration, we use the optimized transformation to get the initial displacement field, which is optimized across different levels of WSIs by gradient update. The loss function of which is the weighted sum of NCC and diffusive regularization. We resize the displacement field and apply it to the original moving WSI. After all the WSI pairs are well registered, we convert the padded H&E image to grayscale and apply median blur to it. Next, the Otsu threshold is applied to extract the foreground area, which is cropped into non-overlapping 256×256 images. Finally, all the chosen images (around 0.35 million) from WSI in the same pair are saved for MMP-MAE pre-training.

BCI Challenge. Breast Cancer Immunohistochemical Image Generation Challenge [16] consists of 3896 pairs of images for training and 977 pairs for testing, which are used to generate HER2 images based on H&E images.

Table 1. Performance comparison on BCI Challenge.

Method	PSNR(dB)	SSIM
cycleGAN [28]	16.20	0.373
Pix2pix [15]	18.65	0.419
Pyramid Pix2pix [16]	21.16	0.477
Proposed	**22.76**	**0.484**

HEROHE Challenge. HER2 On H&E (HEROHE) Challenge [9] is developed to predict the HER2 status in invasive BC cases via the analysis of HE slides. It contains 359 training samples and 150 test samples for WSI classification.

3.2 Experimental Setup

Experiments are implemented in PyTorch [24] and with 4 NVIDIA A100 Tensor Core GPUs. We pre-train our MMP-MAE on the ACROBAT dataset with AdamW [17] and the learning rate of $1e^{-4}$. The batch size of pre-training is 1024 and it takes about 30 h for 100 epochs. We use warmup for the first 10 epochs and the learning rate is set to $1e^{-6}$.

In the HER2 staining image generation task, we use 2 GPUs with a batch size of 4. The learning rate is $2e^{-4}$ and the optimizer is Adam. We use the learning rate decay strategy for stable training. Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) are used as the evaluation indicators for the quality of the HER2 generated images.

In the HER2 status prediction task, we use 1 GPU with a batch size of 1 (WSI level). The learning rate is $1e^{-4}$ and the Adam optimizer is used. Four standard metrics are used to measure the HER2 status prediction results, including the area under the receiver operator characteristic curve (AUC), Precision, Recall, and F1-score.

3.3 Method Comparison

HER2 Staining Image Generation. Three methods on BCI datasets are compared in our experiments, as shown in Table 1. CycleGAN is a representative unsupervised method, which doesn't need paired images for training. So

Table 2. Performance comparison on HEROHE Challenge.

Method/Team	AUC	Precision	Recall	F1-Score
Macaroon	0.71	0.57	**0.83**	0.68
MITEL	0.74	0.58	0.78	0.67
Piaz	**0.84**	**0.77**	0.55	0.64
Dratur	0.75	0.57	0.70	0.63
IRISAI	0.67	0.58	0.67	0.62
Proposed	**0.84**	0.72	0.82	**0.74**

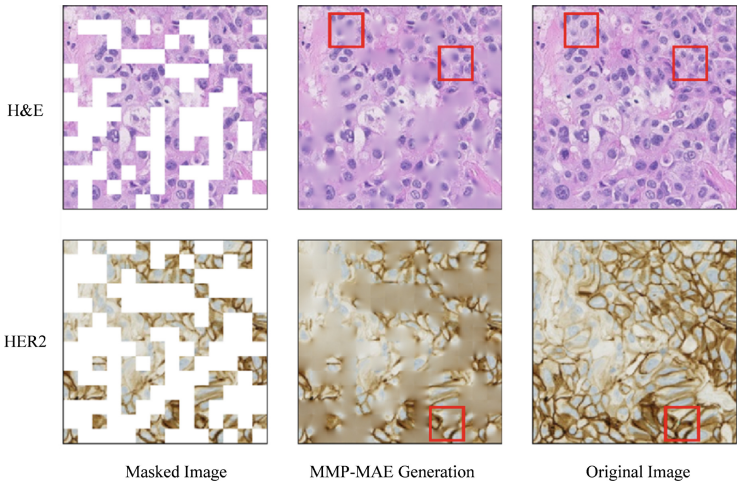

Fig. 5. Visualization of MMP-MAE generation results on the ACROBAT dataset. The region in the red box shows our MMP-MAE could learn the semantic information from the adjacent area. (Color figure online)

cycleGAN focuses more on style transformation, and it is difficult to match the cell-level information in detail. Pix2pix and Pyramid pix2pix use paired data, which obtain better results than cycleGAN. Pyramid pix2pix uses the multi-scale constraint, which performs better than pix2pix. Our method is based on the framework of Pyramid pix2pix and we replace the generator with our pre-trained encoder and a lightweight decoder. Our MMP-MAE further improves the performance, which achieves higher PSNR by 1.60, and SSIM by 0.007. The visualization on the ACROBAT dataset also shows our model could learn the modality-related information, as shown in Fig. 5.

HER2 Status Prediction. We compare our method with the top five methods reported in HEROHE challenge review [9]. Most of these methods use the multi-network ensemble strategy and extra datasets. Team Macaroon uses the CAMELYON dataset [4] for tumor classification. Team MITEL uses BACH dataset [1] for tumor classification. Team Piaz and Dratur both use a multi-network ensemble strategy to improve their performances. Team IRISAI first segment the tumor area and then predict the HER2 status. MMP-MAE still achieves competitive results by using a single pre-trained model, which is shown in Table 2. Our model improves and F1-Score by 6%. The results show our model pre-training has the ability to predict status from one modality.

4 Conclusion

In this paper, we propose a novel multi-modal pre-training framework, MMP-MAE for BC diagnosis. MMP-MAE use paired H&E and HER2 staining images

for pre-training, which could be used for several downstream tasks such as HER2 staining image generation and HER2 status prediction only by H&E modality. Both the experiment results on BCI and HEROHE datasets show our pre-trained MMP-MAE demonstrates strong transfer ability. Our future work will expand our work to more modalities.

Acknowledgment. This work was supported in part by the Key Research and Development Program of Shaanxi Province, China, under Grant 2022GY-084, in part by the National Natural Science Foundation of China under Grant 62171377, and in part by the Key Technologies Research and Development Program under Grant 2022YFC2009903/2022YFC2009900.

References

1. Aresta, G., et al.: Bach: grand challenge on breast cancer histology images. Med. Image Anal. **56**, 122–139 (2019)
2. Bachmann, R., Mizrahi, D., Atanov, A., Zamir, A.: MultiMAE: multi-modal multi-task masked autoencoders. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol. 13697, pp. 348–367. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19836-6_20
3. Baevski, A., Babu, A., Hsu, W.N., Auli, M.: Efficient self-supervised learning with contextualized target representations for vision, speech and language. arXiv preprint arXiv:2212.07525 (2022)
4. 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)
5. Chen, F.L., et al.: VLP: a survey on vision-language pre-training. Mach. Intell. Res. **20**(1), 38–56 (2023)
6. Chen, R.J., et al.: Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Trans. Med. Imaging **41**(4), 757–770 (2020)
7. Chen, R.J., et al.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4025 (2021)
8. Chen, Z., et al.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science, vol. 13435, pp. 679–689. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_65
9. Conde-Sousa, E., et al.: HEROHE challenge: predicting HER2 status in breast cancer from hematoxylin-eosin whole-slide imaging. J. Imaging **8**(8), 213 (2022)
10. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: self-supervised interest point detection and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 224–236 (2018)
11. Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A.: Multiple meta-model quantifying for medical visual question answering. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 64–74. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_7

12. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
14. 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)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
16. Liu, S., Zhu, C., Xu, F., Jia, X., Shi, Z., Jin, M.: BCI: breast cancer immunohistochemical image generation through pyramid pix2pix. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1815–1824 (2022)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
18. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157. IEEE (1999)
19. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomed. Eng. **5**(6), 555–570 (2021)
20. Mobadersany, P., et al.: Predicting cancer outcomes from histology and genomics using convolutional networks. Proc. Natl. Acad. Sci. **115**(13), E2970–E2979 (2018)
21. Nakhli, R., et al.: Amigo: sparse multi-modal graph transformer with shared-context processing for representation learning of giga-pixel images. arXiv preprint arXiv:2303.00865 (2023)
22. Onitilo, A.A., Engel, J.M., Greenlee, R.T., Mukesh, B.N.: Breast cancer subtypes based on ER/PR and HER2 expression: comparison of clinicopathologic features and survival. Clin. Med. Res. **7**(1–2), 4–13 (2009)
23. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
24. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., et al.: PyTorch: an imperative style, high-performance deep learning library. Adv. Neural. Inf. Process. Syst. **32**, 8026–8037 (2019)
25. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4938–4947 (2020)
26. Sung, H., et al.: Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **71**(3), 209–249 (2021)
27. Weitz, P., Valkonen, M., Solorzano, L., Hartman, J., Ruusuvuori, P., Rantalainen, M.: ACROBAT-automatic registration of breast cancer tissue. In: 10th Internatioal Workshop on Biomedical Image Registration (2022)
28. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)