Generating IHC-stained slices from HE slices using the BCI Dataset

Anakha Ganesh[1]
anakhag
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 02139
anakhag@mit.edu

Emily Zhou[1]
emilyz26
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 02139
emilyz26@mit.edu

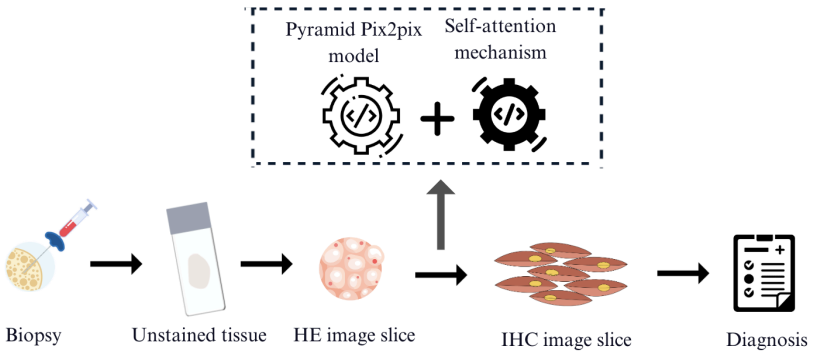

Figure 1. The flow of breast cancer diagnosis stems from creating IHC image slices from HE images obtained through tumor extraction. We automated the process of generating the IHC slices.

Abstract

This study explores the synthesis of immunohistochemical (IHC)-stained images from hematoxylin and eosin (HE)-stained slices using the breast cancer immunohistochemical (BCI) dataset, leveraging computer vision techniques to address the high costs and resource demands of traditional histopathological methods. Two studies were conducted to observe the performance of IHC image generation: a modified PyramidPix2Pix model with self-attention mechanisms and a U-NET model. We aim to improve the accuracy of translating HE images to IHC representations. The PyramidPix2Pix approach utilizes a conditional generative adversarial network (GAN) framework enhanced with a multi-scale pyramid architecture and self-attention layers, enabling the model to capture intricate details and global dependencies critical for medical image processing. Experimental results demonstrate that our enhanced model not only achieves higher fidelity in image quality, as evidenced by improved Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) metrics, but also facilitates a more detailed understanding of tumor heterogeneity. The U-NET framework utilizes a convolutional neural network for image segmentation. The results demonstrated that this framework is not well suited to generating IHC images. This research on modified PyramidPix2Pix holds potential for reducing diagnostic turnaround times and refining treatment planning, thereby advancing breast cancer management and care.

1. Introduction

Breast cancer is the leading cause of cancer-related deaths in women worldwide [1]. Histopathological checks (an examination of changes in a patient's tissue due to disease) facilitate early detection of breast cancer, and computer vision (CV) techniques are being used with increasing frequency to help quicken and specify the treatment process.

The process of detecting breast cancer involves extracting tumor material and making them into hematoxylin and eosin (HE) stained slices. These slides are then examined by pathologists who utilize either a microscope or the digital images of the entire tissue slice (these images are called WSI). For patients who have already been diagnosed with breast cancer, doctors then examine protein expression – in particular, patients with high expression of human epidermal growth factor receptor 2 (HER2), a protein involved in cell growth, are at risk of more aggressive cancer progression [2].

Immunohistochemical (IHC) techniques are necessary to check HER2 expression. Doctors prepare an additional IHC-stained slice to determine how widespread HER2 is in regard to the tumor [3]. In clinical practice, only one IHC-stained slice is prepared, even with multiple HE-stained slices

available, hindering the full heterogeneous expression of the tumor. Current CV models aim to generate IHC-stained slices without additional patient interventions.

Given the necessity of HER2 expression level checks, it's valuable to perform these series of operations for every patient. However, creating and evaluating the IHC-stained images are expensive, not only in terms of money but also resources [2]. Our project will seek to improve on the current methods of synthesizing an IHC-stained image solely based on the HE-stained digital images, in order to help decrease the cost of analyzing a patient's cancer as well as accelerate a patient's treatment.

2. Related Work

Some of the most popular models within the field of medical imaging include cycleGAN, Pix2pix, MGGAN, and U-NET. These models, in particular, have been used for IHC image generation in previous models.

CycleGAN, a type of generative adversarial network (GAN), consists of two sets of GANs, each with a generator and a discriminator, designed for unpaired image-to-image translation tasks [4]. CycleGANs ensure an image translated from one domain to the other can be translated back to the original domain, preserving key attributes [5]. Certain attributes of CycleGAN are applied to Pix2Pix, like its GAN framework, which is used as a conditional GAN framework in Pix2Pix. MGGAN, or multi-generator generative adversarial network, uses multiple generators to focus on a subset of features [6]. Adding generators to Pix2pix was considered as a possible extension of the existing model to focus on specific features with each generator.

Pix2Pix leverages a conditional GAN (cGAN) framework, which is essential for its functionality to translate input images to corresponding output images based on a dataset of aligned pairs [7]. This model is publicly accessible and has been widely adopted for various image-to-image translation tasks. In this framework, the generator attempts to produce output that can't be distinguished from "real" (or target) images in the dataset, while the discriminator evaluates the authenticity of the generated images against the actual images. Pix2Pix enhances the basic GAN loss with an L1 loss component. This L1 regularization term penalizes the absolute differences between the generated images and their real counterparts in the training set, encouraging the network to produce results that are both realistic and closely aligned with the ground truth.

In particular, a 2022 improvement of Pix2Pix, which was named PyramidPix2pix, focuses specifically on generating IHC images for breast cancer diagnosis. The PyramidPix2pix model enhances the traditional Pix2Pix framework by incorporating a multi-scale pyramid architecture, building upon the basic Pix2Pix's use of a conditional GAN structure [8]. PyramidPix2pix introduces a hierarchical, scale-wise refinement mechanism where the image generation process is performed at multiple resolutions. Each level of the pyramid applies Gaussian convolutions and downsampling to gradually refine the image from coarse to fine details. This multi-scale approach allows PyramidPix2pix to produce more precise and higher-quality translations than the original Pix2Pix model, making it especially effective for complex image translation tasks like converting between different types of medical stains.

The U-NET CNN architecture has been used for fast and precise segmentation of images, including for biomedical image segmentation. However, U-NET has not been documented for the translation of HE images to IHC images for breast cancer. Nevertheless, U-NET has been used for image segmentation for breast cancer identification [9]. IHC staining for tissue slices typically allows pathologists to identify malignant growth among the cells present, so image segmentation could be useful for finding relevant outlines of malignant cells in HE slices to generate IHC-stained images. Additionally, a deep neural network using convolution and a GAN framework has been used in the past to generate IHC-stained images, so the success of convolution by itself was explored with the U-NET experiment [10]. This experiment serves to use a simple U-NET framework for the task of generating IHC images, not only segmenting them.

3. Methods/Algorithms and Data

3.1 Data Setup and Preprocessing

BCI (breast cancer immunohistochemical) is a publicly available dataset that focuses on translating HE-stained slices to an IHC result. There are a total of 4870 pairs of image patches coming from the WSIs of 51 different patients, covering different cancer severities [5]. We obtained an academic license for the BCI dataset to download and utilize it. The BCI dataset consists of two folders, one for the HE dataset and the other for the IHC dataset. Each folder has a test and train subgroup. The HE dataset functions as the input and the IHC dataset functions as the answer.

In order to pre-process the dataset, we combined the HE and IHC images such that we were left with pairs of images {HE, IHC}. Each pair of images represents roughly the same foundational medical situation, allowing us to translate the HE images to IHC images.

3.2 Architecture

3.2.1 Proposed PyramidPix2pix Improvement

To improve the PyramidPix2pix architecture, we decided to incorporate self-attention, a core component of transformer models that has progressed the way neural networks handle sequences and spatial relationships. The self-attention mechanism can be mathematically described as follows [11]:

- *Query, Key, Value Vectors:* For each element *i* in the input sequence, the model computes vectors q_i, k_i, and v_i (queries, keys, and values) through linear transformations of the inputs.
- *Attention weights:* The model then calculates the attention weights α_{ij} for each element *j* concerning *i* by taking the dot product of q_i and k_j, followed by a softmax operation to normalize the weights [Equation 1].

$$\alpha_{ij} = \frac{exp(q_i \cdot k_j)}{\sum_{k=1}^{n} exp(q_i \cdot k_j)} \quad (1)$$

- *Weighted sum:* The output for each position *i* is then a weighted sum of the value vectors, weighted by the computed attention weights [Equation 2]. This operation allows each output element to dynamically focus on the most relevant parts of the input data.

$$y^{(i)} = \sum_{j=1}^{n} \alpha_{ij} v_j \quad (2)$$

To integrate self-attention into the PyramidPix2pix framework, we inserted self-attention layers within both the encoding and decoding stages of the network.

Essentially, after each downsampling step in the encoder, we applied a self-attention layer. This layer helps the model capture global dependencies between different regions of the image, which is crucial for understanding complex patterns in medical images or other detailed textures. Similarly, we inserted self-attention layers after each upsampling step in the decoder. This placement ensures that as the model reconstructs the image, it retains the ability to focus on the most salient features identified during encoding.

In more detail, the self-attention module after each downsampling layer in the encoder consists of three linear transformations that map the input feature map into queries *Q*, keys *K*, and values *V* [Equation 3].

$$\begin{aligned} Q &= W_Q \bullet F \\ K &= W_K \bullet F \\ V &= W_V \bullet F \end{aligned} \quad (3)$$

F represents the output feature map from the previous layer, and W_Q, W_K, W_V are the trainable weight matrices for the queries, keys, and values respectively.

The self-attention scores are then calculated by taking the dot product of queries and keys, followed by a softmax to ensure the scores are normalized, as shown in Equation 4.

$$\alpha = softmax(\frac{QK^T}{\sqrt{d_k}}) \quad (4)$$

d_k represents the dimensionality of the keys, ensuring the proper scaling. The output feature map at each position is then updated as a weighted sum of the value vectors, where the weights are given by the attention scores.

For training, we ran the model for 10 epochs with the default learning rate of 0.0002. Additionally, we downsized the images from (1024, 1024) to (512, 512) in order to account for these limitations.

3.2.2 Proposed U-Net Architecture

The U-NET model used for this task is novel, since U-NET has not been used for the task at hand in the past. U-NET utilizes an encoder and decoder structure, as shown in Figure 1, to be able to reconstruct generated images as their input image. It also utilizes skip connections to maintain local and global information to segment the image [12]. The model was constructed in a similar method to a keras based U-NET model, but it does not use dropout, and it uses less convolutional blocks to simplify the task [13]. The model ran for 10 epochs with a learning rate of 0.01. Each image was also resized from (1024, 1024) to (256, 256) to account for local memory constraints.

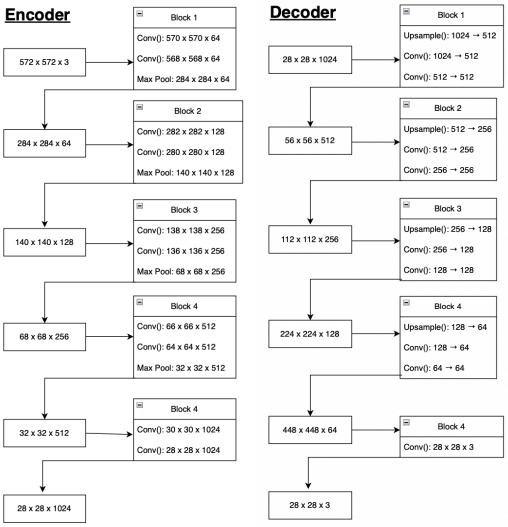

Figure 1. U-NET architecture, separated into encoder and decoder

3.3 Experiments

The experiments we ran involved comparing the generated images from our U-NET model and our improved PyramidPix2pix model to the original images from the BCI dataset. Then, using pre-defined metrics, we compared the metric values associated with the images we generated from the U-Net model and the improved Pyramid Pix2pix model to those obtained from running the original PyramidPix2pix model. To accomplish this, not only did we run our U-NET/improved model, but we also re-trained and tested the original Pyramid Pix2pix model locally to replicate results.

3.4 Evaluation

We evaluated the model by splitting the BCI dataset into training, testing, and validation datasets. We then applied two metrics, Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM), to quantitatively assess the quality of our generated images. Additionally, we performed a qualitative analysis of the generated images by looking at visual differences between the dataset's ground truth images and our generated images.

PSNR assesses the ratio of the maximum possible signal (power of the image) to the power of corrupting noise that affects its fidelity. Higher PSNR values typically indicate better image quality. PSNR values typically range between 30 and 50 dB for image compression and video compression, so a lower PSNR value is expected by virtue of the different staining techniques used for HE and IHC images. We will use PSNR to evaluate the fidelity of the generated IHC images in terms of noise and error levels compared to the originals.

SSIM evaluates image quality based on luminance, contrast, and structure comparisons between the generated image and the reference image. It ranges from -1 to 1, where 1 indicates perfect similarity. We will use this metric to measure how well the structural integrity and textural details of the generated IHC images match the original.

4. Results

The following results compare the PSNR and SSIM metrics between the three models we ran.

Model	PSNR	SSIM
Original PyramidPix2pix	21.16	0.477
Improved PyramidPix2pix	21.31	0.493
U-NET	3.464	0.0283

Table 1. Comparison of PSNR and SSIM values between the original Pyramid Pix2pix model, the improved Pyramid Pix2pix model, and the U-Net model.

The IHC images generated by our improved PyramidPix2pix model are shown in Figure 2.

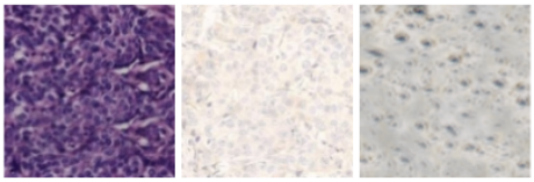

Figure 2. HE image (left), ground truth IHC image (center), our generated IHC image from the improved Pyramid Pix2pix model (right)

We also generated IHC images from our U-NET model, as shown in Figure 3.

Figure 3. HE image (left), ground truth IHC image (center), our generated IHC image from the U-NET model (right)

5. Discussion/Significance

From the PSNR and SSIM metric values for the original Pyramid Pix2pix model and our improved model, we can see a slight increase in both values; this indicates that our model was able to effectively preserve and reconstruct the original image quality as well as the structural integrity.

Visually, based on the IHC image generated by our model compared to the ground truth IHC image, the general pattern of dark spots and other nuances are represented in our image. The main difference is a slight discoloration of the general image, but because we were limited on compute resources and thus had to downsample our training images to effectively train our model, our generated image essentially captures the essence of the HE image.

U-NET produced images with little to no similarity to the ground truth image. The reconstructed image shown appears entirely black; however, upon further inspection of the pixel values in the tensors, it was observed that the values were small-valued decimals, with the first significant figures in the hundredths or

thousandths place. We saw this issue with numerous randomly-inspected images. The convolution technique without a GAN seems to make it more difficult for the model to generate a new image. The nature of the task of generating a new IHC image entails that the model creates a new image. It is possible that deeper CNNs would be able to generate images closer to ground truth with enough training. However, the design of U-NET is meant to be used specifically for precise image segmentation, and it lacks depth in the network due to the need to be able to reconstruct the original input image using the decoder [9]. This lower capability to generate new images likely fed into the failure of the U-NET model. Further investigation of the number of epochs trained, learning rate, and other hyperparameters would confirm that the type of training was not the issue in U-NET's failure, it was the task U-NET was applied to.

Overall, our proposed model with self-attention in PyramidPix2pix has the potential to contribute towards more effective treatment for diagnosed breast cancer. By generating IHC-stained slices for each HE slice, physicians can gain more information about heterogeneous tumors. Our enabling of the PyramidPix2pix model to focus on relevant parts of the image throughout the processing pipeline with self-attention enhances the model's ability to handle tasks that require a high level of detail and accuracy, such as medical image translation. With an accurate model that can do the proposed image generation, we can formulate a complete treatment plan for women with diagnosed breast cancer to prevent recurrence and administer targeted protocols.

6. Conclusion

Our research has demonstrated that integrating self-attention mechanisms into the PyramidPix2pix architecture enhances its capability to generate IHC-stained images from HE-stained slices with greater fidelity and structural accuracy. The improvements observed in both PSNR and SSIM metrics signify that the modified model not only preserves the essential visual features of the original images but also improves the overall image quality.

There are numerous future steps we could take to further refine and improve our model's accuracy. The first is incorporating other types of improvements in our model to compare the effectiveness of these new adjustments. For example, the foundation of stable diffusion models seems promising, as general stable diffusion techniques can help with learning generalized features to mitigate issues with overreliance on one feature [14]. Furthermore, we could incorporate another accuracy metric of requesting real-world pathologists to diagnose the level of breast cancer using our generated IHC images; we could then compare their final diagnosis with our ground truth.

With regard to the U-NET model, we could run future experiments with different hyperparameters and ensure that the resultant images were not due to a lack of time to converge. Another experiment would be to reconstruct the model that the usage of U-NET was inspired based off of. A deep learning framework using a GAN and deep convolutional networks to generate IHC stained images inspired the usage of U-NET, a smaller convolutional network, to generate IHC images [10]. Rather than isolating U-NET, we would be interested to see the results of adding a GAN to generate images and having fewer layers of convolution with a U-NET model as opposed to the multiple layers of convolution in the original deep learning framework [10].

Ultimately, our implementation of this computer vision technology with PyramidPix2pix could lead to increased accessibility of high-quality diagnostic services in resource-limited settings, enhancing diagnostic accuracy and treatment efficacy in oncology. The advancement of computer vision techniques in medical imaging holds significant promise for contributing to the global effort to reduce disparities in healthcare outcomes.

7. Individual Contributions

Together, we wrote all our reports, such as our progress report and this final project report, collaboratively. Furthermore, we brainstormed our ideas and next steps together, relying on each other when any of us encountered any obstacles.

Individually, Emily worked on the Pyramid Pix2pix model and results write-up for Pyramid Pix2pix for this project, whereas Anakha focused on the U-Net model and results write-up. Emily specifically worked on re-running the original Pyramid Pix2pix model as well as creating/running the improved version on the BCI dataset. Anakha focused on reading existing literature on U-Net architecture and adapting it to fit our task at hand (IHC image generation from HE images). She then trained, tested, and validated the new U-Net model she made.