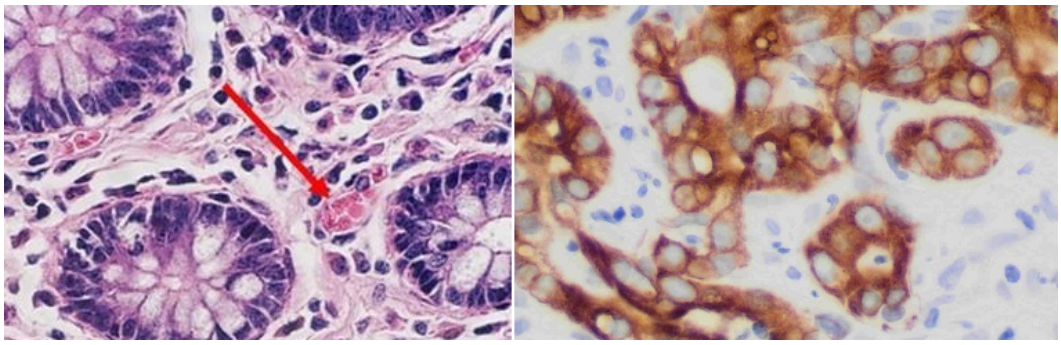
Deep Learning Models for Tumour Ratio Estimation in Histological Images

Introduction and Background: Hematoxylin and Eosin (H&E) staining is one of the most widely used techniques for imaging tissue structures under a microscope. It provides a contrast between cellular components, making it useful in medical diagnostics and research. In this technique, hematoxylin stains cell nuclei blue or purple by binding to nucleic acids, identifying cell nuclei. Eosin stains the cytoplasm and other structures pink or red. This allows the differentiation of tissue components, including the distinction between normal and pathological areas, such as cancerous tissues.

H&E staining is widely used because it is simple to apply and cost-effective in detecting abnormalities, including metastases. However, H&E results lack consistency compared to other techniques like immunohistochemistry (IHC). It cannot label tumour cells directly and requires an experienced pathologist to determine whether a cluster of cells represents metastasis based on the context of cellular morphology and nuclear arrangement. This process can introduce subjectivity and, therefore, variability between different observers.



Left: H&E result, red arrow indicates tumour cells. Right: IHC result, tumour cells are brown, normal cells are blue.

The development of machine learning for analyzing H&E-stained whole slide images (WSI) has gained significant clinical success. Various research groups have utilized deep learning models to detect tumour metastases in sentinel lymph nodes, as demonstrated in the Camelyon challenges (16 and 17), grade cancer stages, and categorize diagnostic outcomes with high accuracy. However, these models cannot directly apply to cancer biology research due to their different training targets and objectives.

In clinical settings, metastatic cancer cells shed from advanced-stage primary tumours and form metastases (Mets), with staining results typically showing low numbers of metastatic clusters. Machine learning models trained for clinical use often operate with a binary outcome (presence or absence of Mets in lymph node images). In cancer staging or diagnostic categorizing problems, the outcomes are limited to a few discrete categories. In contrast, cancer biology research involves introducing cancer cells into immunodeficient mice to generate metastases and quantify the number of metastases in different tissue samples. This allows researchers to assess whether a treatment can inhibit metastasis by reducing the number of metastatic clusters, which can be high in number. Therefore, more precise segmentation is required in research applications than in clinical models.

Problem Statement: The project addresses the challenge of accurately estimating tumour ratios in histological images, particularly for cases lacking pixel-perfect paired datasets. This solution aims to benefit researchers and pathologists working on cancer detection and staging.

Summary: For this project, we plan to use immunohistochemistry (IHC) staining results as ground truth, leveraging data from publicly available online databases.[1] A key challenge in using IHC as a ground truth lies in its imperfect alignment with hematoxylin and eosin (H&E) staining since these stains are obtained from consecutive, yet distinct, tissue slides. To mitigate this issue, we will focus on the number of tumour cells rather than pixel-wise precision for training, as we base our approach on the assumption from prior research that the ratio of tumour to normal tissue area remains consistent at corresponding locations across the two slides.

Generative Adversarial Networks (GANs) provide an adversarial learning framework where a generator creates images, while a discriminator evaluates them, making GANs highly suitable for biomedical image processing. In this project, we will explore two different GAN-based methods to generate Ki67 IHC-stained images from H&E slides.

First, we will employ pix2pix [2,3], a conditional GAN model that requires paired input images for training. Here, we will use a modified version called Pyramid Pix2Pix, and use a pre-trained model to generate IHC-like images from H&E. Second, we will use CycleGAN[4], which does not require paired image inputs, thus providing a flexible option for unpaired training sets. By implementing these two approaches, we aim to improve the accuracy of our results and offer a novel way to evaluate potential treatment effectiveness.

Petríková, D., Cimrák, I., Tobiášová, K., & Plank, L. (2024). Ki67 expression classification from HE images with semi-automated computer-generated annotations. In Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 1: BIOINFORMATICS (pp. 536-544). SciTePress. https://doi.org/10.5220/0012535900003657

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1125-1134.

Zhang, X., Yu, J., & Wang, Y. (2022). Multi-Scale Pyramid pix2pix for Enhanced Image-to-Image Translation in Biomedical Imaging. IEEE Transactions on Medical Imaging, 41(5), 1380-1392.

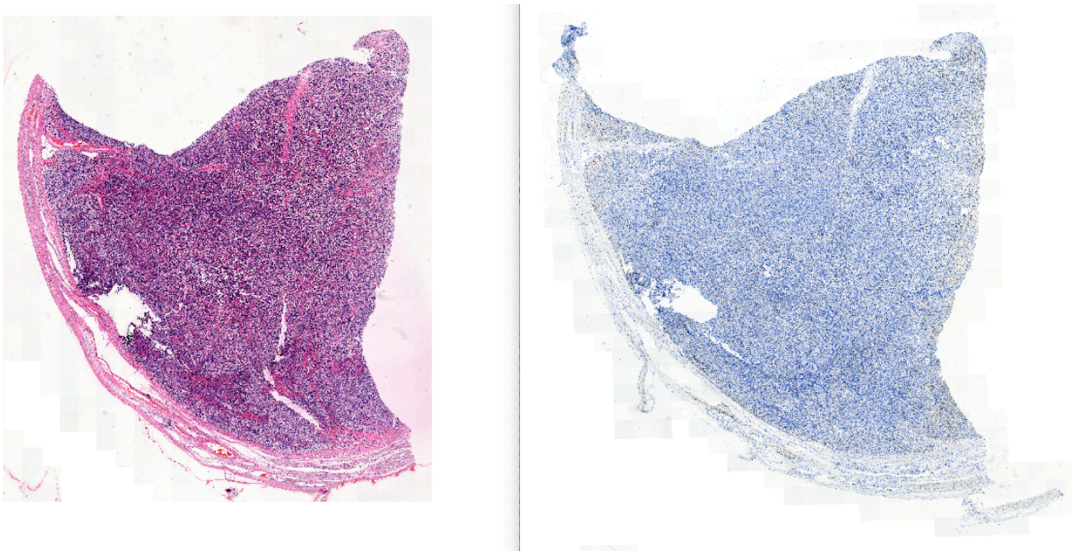
Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2223-2232.

Pipeline and Baseline:

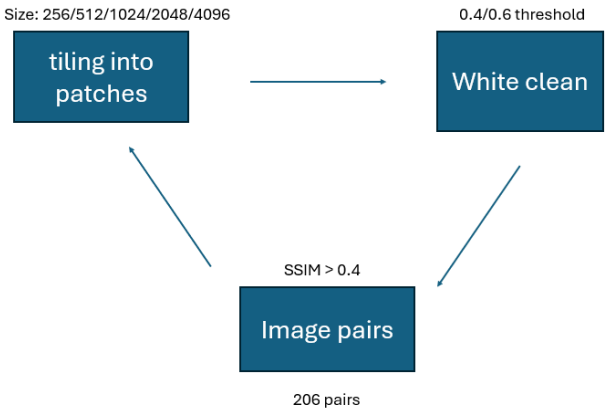
1. Data Preprocessing:

1.1 Download the Datasets

77 pairs of whole slide images of H&E staining and Ki67-stained IHC images are downloaded from https://zenodo.org/records/11218961. These slides are registered and aligned with the method described in the paper. All slides are of size 31104 pixels * 31938 pixels



1.2 Data pre-processing



The preprocessing phase consisted of several iterative steps, including tiling, white cleaning, and similarity evaluation between image pairs, aimed at ensuring high similarity between H&E and IHC image pairs. This was a critical requirement for generating IHC images from H&E slides using GAN-based models.

Smaller patch sizes, such as 256-pixel tiles used in previous studies, proved inadequate for preserving discernible similarities between the image pairs. Larger patch sizes were then used to capture meaningful features better and enhance the quality of the generated IHC images. The white cleaning step was crucial for removing patches containing mostly or entirely background areas, as these patches caused crashes when processed with StarDist.

The Structural Similarity Index Measure (SSIM) was used to evaluate the similarity between H&E and IHC image pairs. A threshold of 0.4 was applied, and only pairs exceeding this threshold were retained for further analysis. Afterward, unsuitable images were manually removed, resulting in a refined dataset of 206 image pairs for subsequent studies.

2. Developing ground truth

The original study developed ground truth by recoloring IHC slides and calculating the ratio of brown pixels to the total of brown and blue pixels. However, this approach posed two significant challenges:

1. The threshold for recoloring is subjective and can vary significantly depending on the implementation.

2. The blue channel, which stains for total cell nuclei, often gives a slight blue tint across the tissue, leading to an overestimation of the nucleus area.

To address these limitations, StarDist was used for segmentation instead. StarDist specifically calculates regions with star-convex shapes, effectively segmenting only on cell nuclei. This approach minimizes the overestimation issue and provides more accurate segmentation compared to recoloring.

The following figure illustrates the general approach for 2D images. The training data consists of corresponding pairs of input (i.e. raw) images and fully annotated label images (i.e. every pixel is labeled with a unique object id or 0 for background). A model is trained to densely predict the distances (r) to the object boundary along a fixed set of rays and object probabilities (d), which together produce an overcomplete set of candidate polygons for a given input image. The final result is obtained via non-maximum suppression (NMS) of these candidates.

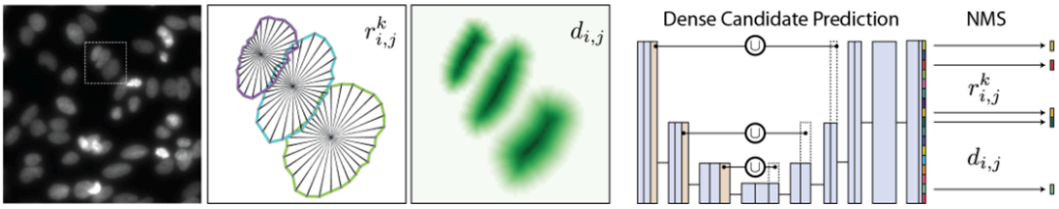
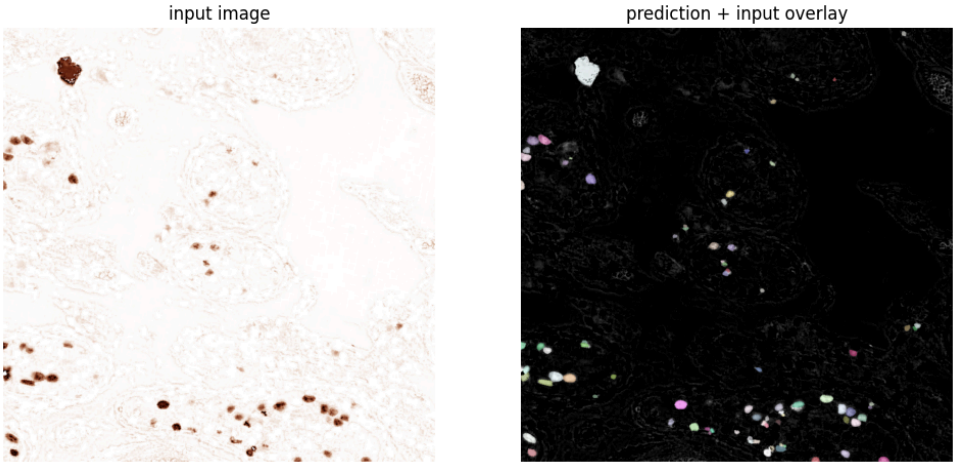


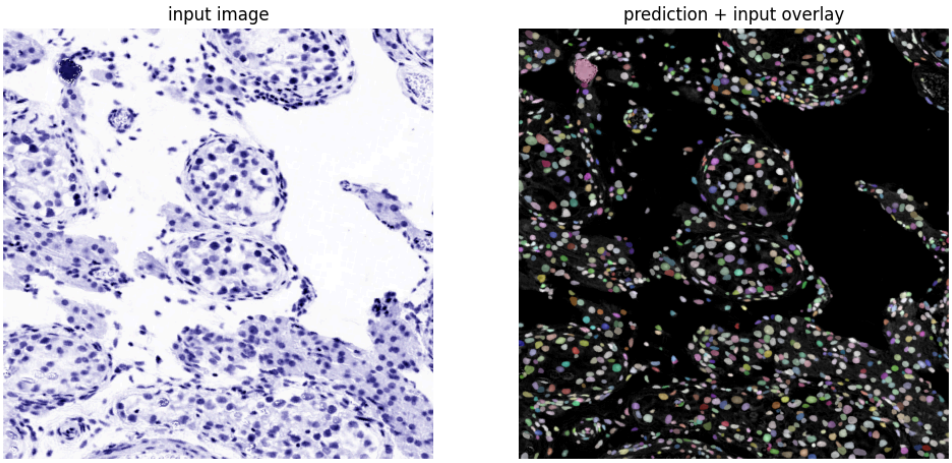
Illustration of how StarDist works from its GitHub page. Works exceptionally well with a crowded layout of objects.

The process involved converting IHC patches from RGB to HED format, separating the image into different channels. Cell nuclei were segmented in both channels, and the ratio of tumour to total tissue was calculated by summing up the segmented regions. While simple thresholding methods did not yield satisfactory results, StarDist provided accurate segmentation that was visually validated. The segmentation code was adapted from a YouTube tutorialhttps://www.youtube.com/watch?v=L3dZ6fgmIII using a pre-trained StarDist model for efficient and precise cell segmentation.

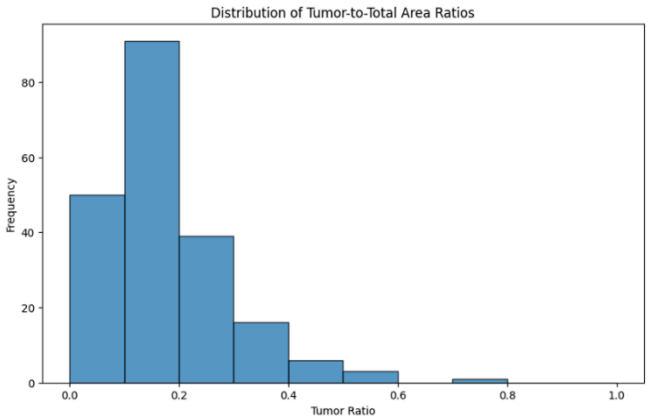
Ki-67 (tumour) channel segmentation result:



H (total tissue) channel segmentation result:



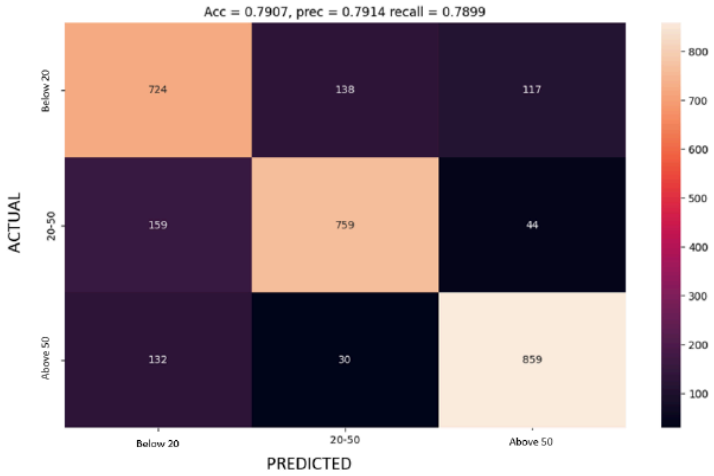
Tumour ratio with StarDist segmentation:



For H&E images, the same approach didn't work. The final decision is influenced by multiple factors beyond just colour. Nuclear size, colour balance, and nuclear arrangement are key contributors to identifying and classifying cell types and regions. Unlike IHC images, where specific stain colours (e.g., brown and blue) play a central role, H&E relies on a more holistic evaluation of morphological and spatial features.

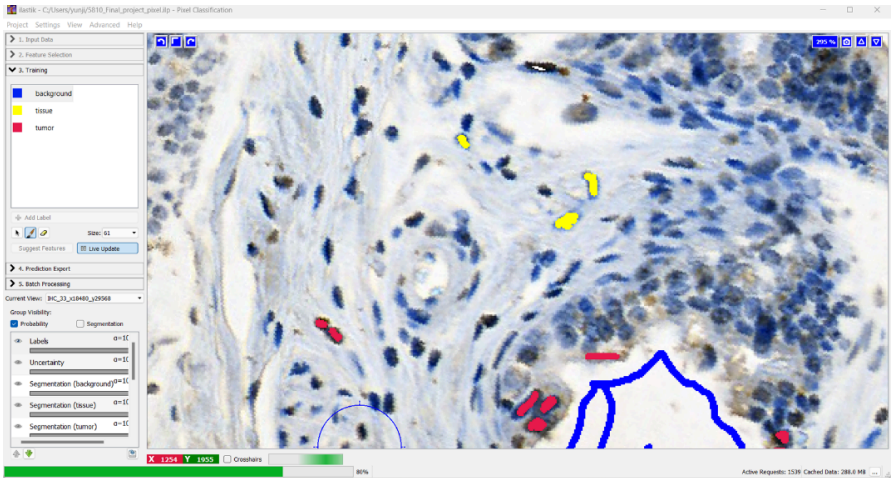
3. Baseline development with ilastik

The original paper divided IHC patches into three groups based on their tumour area ratio: 0-0.2, 0.2-0.5, and 0.5-1. They then trained H&E patches using ResNet with these labels, achieving a maximum accuracy of 0.79 after optimization. However, this method simplifies the problem by categorizing tumour area ratios into just three classes. To achieve more precise outputs from H&E slides for our research, a different approach was necessary.



Result from the original paper using ResNet18

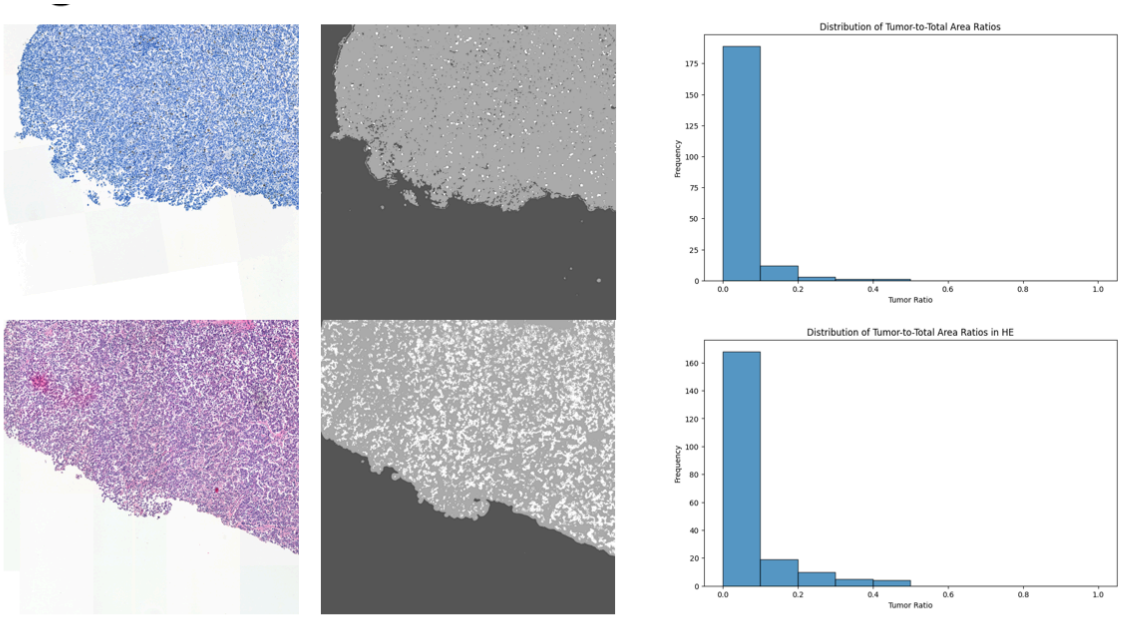
3.1 Using *ilastik* for Segmentation



Ilastik API screenshot

To improve accuracy, we employed *ilastik*, a versatile tool for image analysis and segmentation. Ilastik provides a user-friendly interface that allows manual data labelling for segmentation tasks. By utilizing a random forest algorithm, *ilastik* classifies pixels into user-defined groups with high accuracy.

We selected 10 patches from both groups and manually labelled background, nucleus, and cell regions. This method should produce a tumour ratio distribution that more closely aligns with the results presented in the original paper. Our IHC segmentation results confirmed this, demonstrating the effectiveness of *ilastik* in achieving accurate tumour-to-tissue ratio measurements.

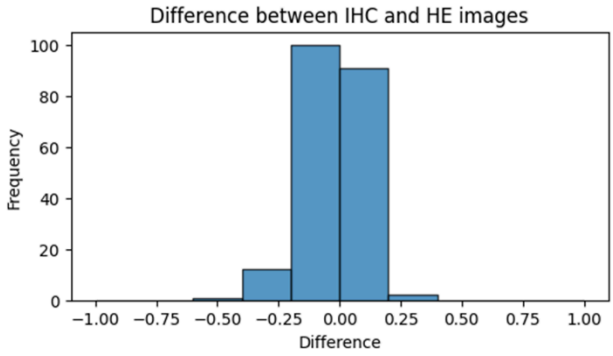


Top row, left to right: IHC patch, segmented result, tumour area ratio from 206 pairs

Bottom row, left to right: H&E patch, segmented result, tumour area ratio from 206 pairs.

When labelling H&E-stained slides, we noticed that these slides contain numerous micro-metastases, which are notoriously challenging to detect. The labelling process can be subjective and somewhat arbitrary. This subjectivity introduces variability, making accurate segmentation and classification more difficult. So, we calculated the difference between H&E and IHC results.

3.2 Accuracy of baseline



If we threshold the **difference at 0.05** between paired images, the **accuracy is 0.7524**. If we relax the threshold to **0.1**, accuracy is at **0.8447**, and a further relaxation to **0.2** renders an accuracy level of **0.9272**. **MSE for the 206 pairs is 0.0084, and MAE is 0.0455**.

The results demonstrated relatively high accuracy compared to the original report. However, with our approach, the tumour ratio distribution was heavily skewed toward the first bin (0-0.2). As a result, the high accuracy does not necessarily indicate a superior method, as it reflects the imbalanced distribution rather than true robustness. Since our goal is to determine the tumour ratio from H&E-stained slides accurately, we opted to improve both accuracy and robustness using a different approach.

4. Generate IHC images from H&E images.

There have been numerous studies focused on detecting tumour areas in H&E slides. However, most of these approaches either provide binary outputs (tumour positive or tumour negative) or classify results into cancer stages based on the slides. To achieve more precise outputs, we propose generating IHC images from H&E slides for the following reasons:

1. Reduced dependence on expert annotation: Segmentation on H&E slides heavily depends on morphological features like nuclear size and shape, which can vary widely across tissues and cell types, and requires experienced pathologists for manual annotation.
2. Alignment with ground truth of IHC: Direct segmentation on H&E lacks the direct correlation with such ground truth, making validation and comparison difficult.

We plan to experiment with two GAN-based models to generate IHC images from H&E-stained slides. This approach deals with imbalanced tumour ratio distributions by offering a more balanced and accurate method for generating IHC images and analyzing tumour areas. By leveraging these advanced generative models, we aim to improve the precision and reliability of tumour ratio estimation while minimizing the influence of distributional biases in the data.

4.1 pix2pix

Pix2Pix is a GAN-based model that incorporates an additional Mean Absolute Error (MAE) loss function alongside the GAN loss to enhance training by minimizing the pixel-level differences between the generated and target images. However, the lack of pixel-perfect paired image data presents a significant limitation for training Pix2Pix models effectively in our case. The MAE loss function, which relies on direct pixel-to-pixel comparisons, becomes ineffective without paired datasets.

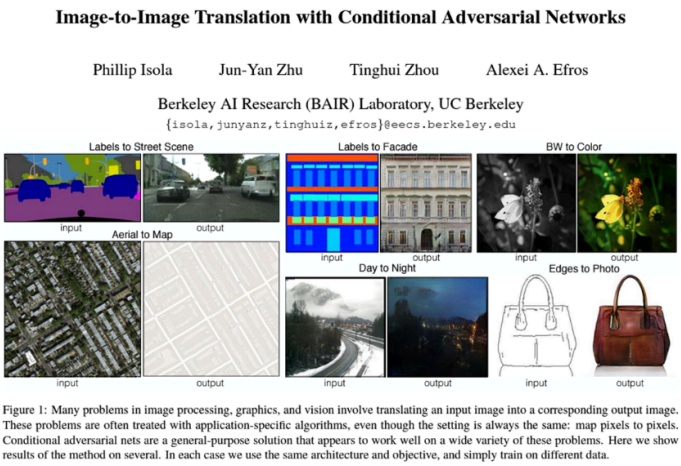
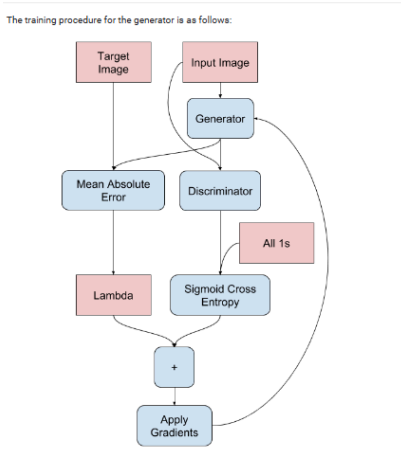


Figure 1: Many problems in image processing, graphics, and vision involve translating an input image into a corresponding output image. These problems are often treated with application-specific algorithms, even though the setting is always the same: map pixels to pixels. Conditional adversarial nets are a general-purpose solution that appears to work well on a wide variety of these problems. Here we show results of the method on several. In each case we use the same architecture and objective, and simply train on different data.



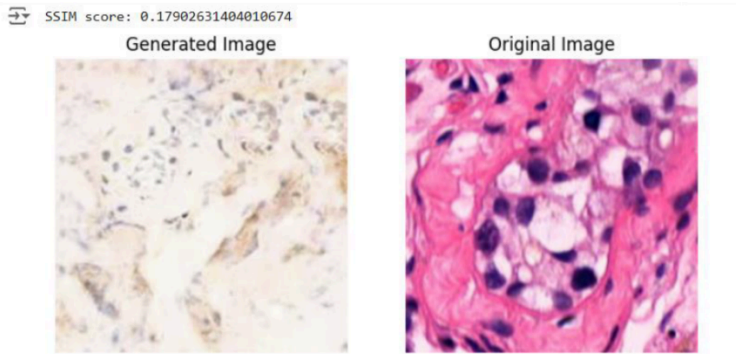
To address this, we utilized a modified version of the Pix2Pix model, cloned from a GitHub repository. https://github.com/bupt-ai-cz/BCI This Pyramid Pix2Pix model comes equipped with pre-trained weights. This adaptation made it a suitable choice for our task, given the unavailability of paired training images.

To address the issue of perfect alignment, the pyramid Pix2Pix model applies multiple rounds of Gaussian blurring and down-sampling to progressively reduce the resolution of both the ground truth and generated images. This approach relaxes the original model's loss function, making it less sensitive to small misalignments in the training data. By focusing on lower-resolution representations, the model captures overall structure and context rather than relying on precise pixel-wise accuracy to improve its generalization and robustness.

However, the pre-trained model didn't work well for at least two reasons:

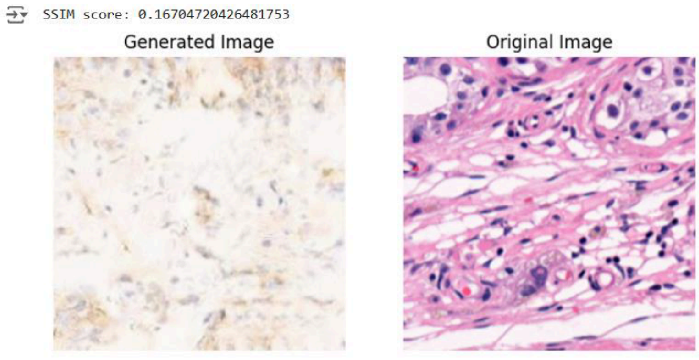
1. It was trained on human breast cancer tissue with HER2-labelled IHC slides, which is a different tissue source: human breast cancer versus mouse testicular seminoma.
2. HER2 staining highlights the cytoplasm, unlike our dataset, where the focus is on the cell nucleus.

As a result, the generated IHC images had a low Structural Similarity Index Measure (SSIM) when compared to their corresponding IHC patches.



Generated image from original input

The different cell nucleus size was evident from the generated image, and we went back to the BCI database used for training of the pre-trained model and realized that they used 20X magnification for BCI instead of 40X used in our case. So, the images were resized and went through the process again. Unfortunately, the result still lacks similarity, probably because of the reasons mentioned above.



Generated image from resized input

The Structural Similarity Index Measure (SSIM) for the generated images remains very low, indicating poor resemblance to the target images. Moreover, the pyramid pix2pix paper gave a final accuracy of 0.4, and a different method without the need for paired images seems necessary in our case.

4.2 cycleGAN

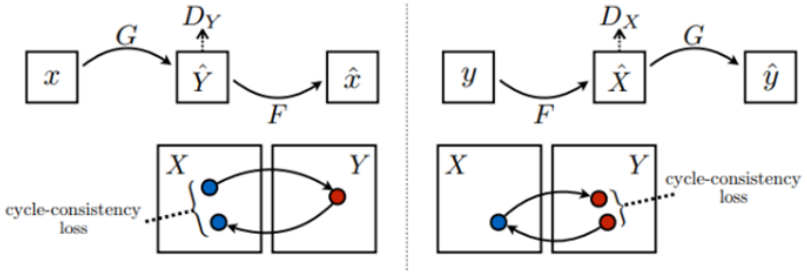
Next, we explored CycleGAN, an alternative GAN-based model that does not require paired images for training, making it well-suited for our task. CycleGAN operates by translating images between two domains using a combination of loss functions:

1. Cycle-Consistency Loss to maintain the overall structure and integrity of the image.
2. Identity Loss to preserve key features of the input.
3. GAN Loss

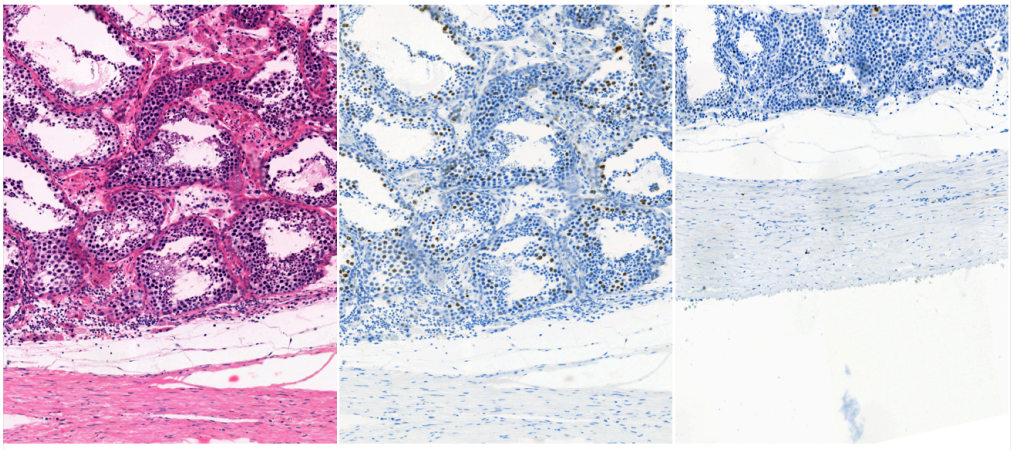
These features make CycleGAN particularly effective in our case, where paired datasets are unavailable.

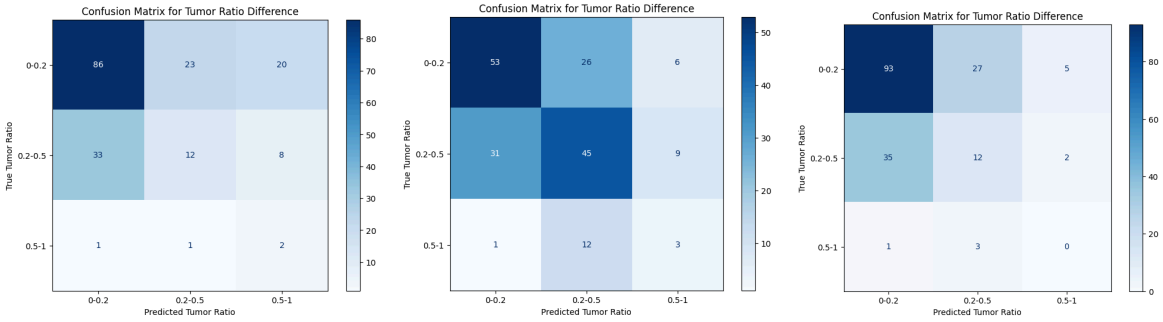
$$forward\ cycle\ consistency\ loss: X -> G(X) -> F(G(X)) \sim \hat{X}$$

$$backward\ cycle\ consistency\ loss: Y -> F(Y) -> G(F(Y)) \sim \hat{Y}$$



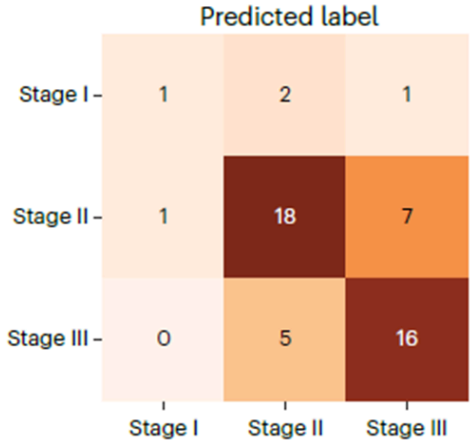
Here we cloned the repository https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix and trained 389 images from both groups. Limited to the T4 GPU offered by Google Colab, with 15GB of memory, we cannot change the crop size for training. Therefore, different resizing settings were tried for the optimal results. Load size of 4096 without resizing ended up with the highest accuracy of 0.6 and lowest MAE of 0.13.





Confusion matrix with different load sizes. left to right: 256/2048/4096 pixels

The highest accuracy achieved in our current setup is approximately 0.6, which may initially seem low. However, the previous pix2pix model gave an accuracy of 0.4, and a CUT-based model listed below achieves an accuracy of 0.68 in determining cancer stages using the generated images. This highlights that the accuracy of machine learning models, even with relatively complex architectures, still cannot compare to the performance of a trained pathologist.

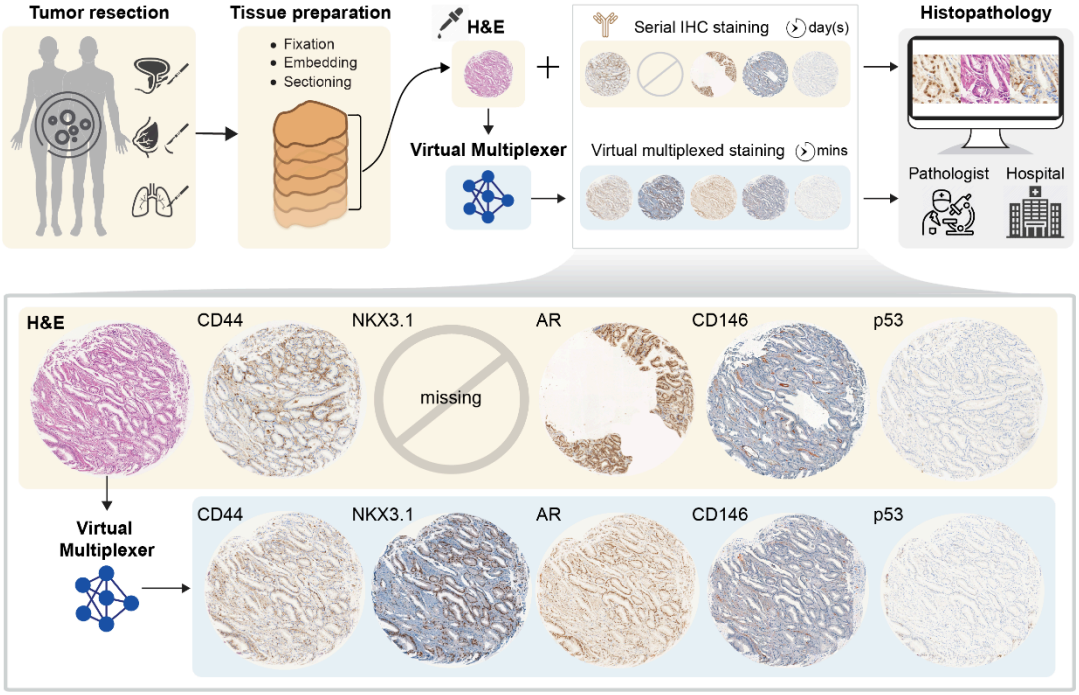


VirtualMultiplexer generated confusion matrix, accuracy at 0.68

To improve accuracy further, we can consider employing more advanced generator designs or integrating state-of-the-art backbone architectures. Additionally, increasing the number of training epochs may enhance the model's ability to learn detailed features and achieve better convergence, potentially narrowing the gap between machine learning methods and expert pathologists.

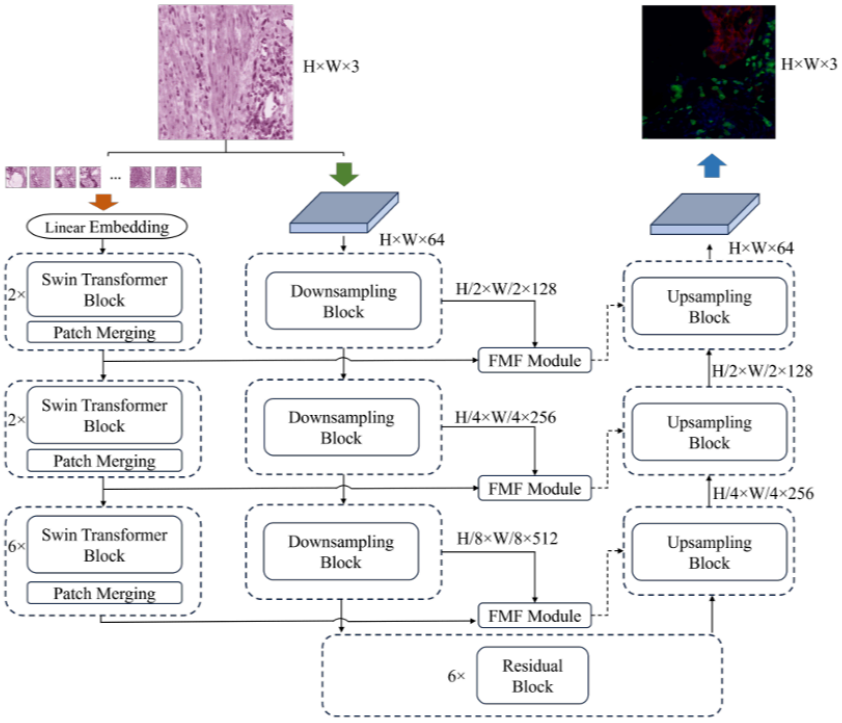
Future direction: Many other studies are attempting to generate IHC images from H&E slides, and there are many we wanted to try if there is more time. One particularly interesting model that also doesn't require perfectly paired images is called VirtualMultiplexer (https://www.biorxiv.org/content/10.1101/2023.11.29.568996v1), also a pix2pix-based model. Instead of employing two GANs to create a cycle, it uses the Contrastive Unpaired Translation (CUT) approach, which brings positive features closer together while pushing negative features further apart. This method requires only a single GAN, making it more lightweight and easier to train. As mentioned above in the results part of cycleGAN, the model achieved an accuracy of 0.68 based on the confusion matrix. However, the provided repository from the authors was not functional

(https://github.com/AI4SCR/VirtualMultiplexer), and an updated, debugged version is yet to be made available on GitHub (see issues part of the repository).



VirtualMultiplexer generated mIHC

Another pix2pix-based approach worth noticing uses a more advanced Swin transformer layer for feature extraction (https://arxiv.org/abs/2403.18501). The study combines residual CNNs and Swin Transformers within the Pix2Pix framework to effectively capture both global contextual information and local spatial details. Their results demonstrate improved accuracy and quality of the generated images compared to previous methods. Additionally, this study introduces the HEMIT dataset, which includes identical structures from multiplex immunohistochemistry (mIHC) and H&E staining on the same tissue slide. This alignment at the pixel level enables high-precision image translation and provides a valuable resource for future computational pathology research.



Structure of Transformer-based model

Lesson learned: This project allows me to delve deep into advancements in H&E segmentation and introduces me to various deep learning models, including Pix2Pix and cycleGAN. It has inspired me to further explore this field and consider applying the knowledge gained to other interdisciplinary areas.

Source Code:
https://colab.research.google.com/drive/1186jB6UVvfQLMS3v5PxYtJBT0O4Vn4x-#scrollTo=Xkvhje9un6RP

DEMO video: https://drive.google.com/file/d/12IoIveg_qdps3BgCb3HgsSlf1vMamIky/view?usp=drive_link