Machine Learning for Medical Image Analysis – Registration, Segmentation, Characterization and Tracking

Archana Machireddy
M. S., Indian Institute of Technology Madras

Presented to the
Computer Science & Engineering Education Program
within the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree
Doctor of Philosophy
in
Computer Science & Engineering

June 2022

Computer Science & Engineering Education Program
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Ph. D. dissertation of

Archana Machireddy

has been approved.

Xubo Song, Thesis Advisor
Professor

Meysam Asgari
Assistant Professor

Guillaume Thibault
Research Assistant Professor

Peter Heeman
Associate Professor

Acknowledgements

I would like to express my sincere gratitude to my advisor Xubo Song for her continuous support, guidance and encouragement throughout the doctoral degree. I have learned a lot from her, both technical and personal. She has taught me how to methodologically break down and analyze a problem to always look at all the possible ways in which it can be tackled before choosing a particular path. I will always cherish the time we spent brainstorming on multiple projects we have worked on.

I would like to thank Peter Heeman for his support through out and especially over the last few months in guiding and helping me through the process of writing this dissertation. I appreciate everything he has done for me over the years to ensure I could continue my research this far. I would also like to thank Guillaume Thibault for all his advice and feedback on the numerous projects we have collaborated on.

I would like to thank Jan van Santen, Wei Huang, Fergus Coakley, Joe Gray and Jessica Riesterer for their collaboration and input on various projects. I would like to thank Julianne Myers, Cecilia Bueno and Hannah Smith in helping with collecting and pre-processing data. I would like to thank Patricia Dickerson for helping me with administrative work through all these years.

I would like to thank my thesis committee members, Meysam Asgari, Guillaume Thibault and Peter Heeman for their insightful discussion and comments. I would also like to thank Alexander Kain and Steven Bedrick for all their encouragement and feedback over the years.

Finally, this dissertation would not have been possible without the support and patience of my husband Vineet and the constant encouragement from my parents and sister. I would also like to thank Vineet's family for their continuous support over the years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Medical images provide a non-invasive means of assessing structural and functional changes within the human body [47]. They are a rich source of information on the internal body structures and play a vital role in helping doctors diagnose, monitor and treat various medical conditions. Each imaging technique is unique in terms of the image it captures and the information it provides regarding the area of the body under study [159]. Such information helps in providing diagnosis of a disease without the need for expensive invasive procedures, sparing patients risks and reducing treatment costs [159]. It also helps in early diagnosis of several diseases, including cancer and Alzheimer's disease, providing an early opportunity to start treatment [18, 121, 163].

With recent advancements in imaging techniques, the volume of medical images being generated is growing exponentially. Manual interpretation of these images is a time-consuming and tedious task and the results are dependent on human expertise. Therefore, in order to achieve accurate and timely diagnosis, there is a need to automate medical image analysis. Over the years, several image processing algorithms have improved the diagnostic interpretability of medical images [14]. These algorithms help in processing, analyzing and extracting useful information from images. Recent advances in image processing algorithms leverage machine learning models to improve their performance on various tasks. Machine learning models learn to perform a task, without being explicitly programmed to do so, based on the data provided to train the model.

Typically, analysis of medical images using machine learning algorithms follows a specific pipeline. An image processing pipeline is a set of tasks executed in a sequence to transform an image into a desired result. A generic example of this pipeline is shown in Figure 1.1. Each stage in the pipeline employs one or more algorithms to improve the interpretabilty of the image. The acquired image is first subjected to suitable pre-processing techniques, such as noise reduction and contrast adjustment, to suppress undesired distortions and enhance image features relevant for further processing and analysis steps. When dealing with images from multiple modalities or

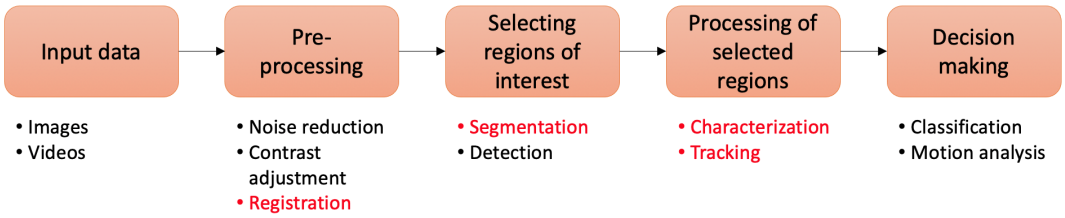


Figure 1.1: A typical image processing pipeline. Steps indicated in red are dealt with in this dissertation.

a temporal sequence of images, an additional step of aligning these images using registration algorithms is required to accurately relate information from all the images. The pre-processed images can be interpreted by humans directly or be further analyzed by machine learning algorithms to arrive at a diagnosis without human intervention.

After pre-processing, the next main step in medical image analysis is the segmentation of anatomical regions of interest within the image to aid in targeted downstream analysis. Segmentation and detection algorithms divide the image into regions and transform the representation of medical images into a meaningful form [14]. Each segmented anatomical region can be characterized by extracting texture and shape related features. The extracted features represent a compact form of the most relevant information regarding the region. By training classification models based on these feature, different medical conditions can be identified. Machine learning methods, such as deep learning, can combine multiple steps in the pipeline to directly arrive at a diagnostic decision [104]. Though images are a rich source of information, certain applications such as gait analysis and cardiac motion estimation require analysis of movement [78, 102]. In such cases, a series of images capturing the entire movement are recorded as a video. The object of interest is detected in each image using an object detection algorithm and is tracked across all images in the video using a tracking algorithm [19, 186]. Multiple studies have shown that machine learning methods perform as well, if not better, than humans in identifying features and classifying images quickly and precisely [114, 9]. Medical images may contain information that is too subtle for the human eye to see and interpret [143]. Machine learning algorithms aid in extracting such hidden insights from medical images as they develop new ways of interpreting images, sometimes in ways that humans do not understand [143]. Machine learning methods have proven to be a valuable ally to doctors by improving their productivity as well as accuracy in diagnosing and monitoring various medical conditions.

1.1 Dissertation Problem and Statement

Medical images are one of the richest and often the most complex source of information regarding patients. Analysis of medical images can provide vital insights that help doctors detect and diagnose various diseases. But as discussed in the previous section, we observe several problems in analysis of medical images.

Problem 1: Manual interpretation: With recent advancements in imaging techniques, medical images are being acquired at a pace much faster than the pace at which the human experts can interpret them. Manual interpretation is an expensive and laborious process with subjective results. Increase in the number of medical images being produced increases the burden on the radiologist to efficiently interpret them in a short amount of time. This increased reading speed could result in missing some details and making errors in the interpretation. In applications requiring real-time analysis of medical images, a radiologist would always need to be present to interpret the images. Thus, in order to achieve accurate and timely diagnosis automated image analysis methods are essential.

Problem 2: Richer information not interpreted: Medical images contain a lot of rich information that may not be visually available. Therefore, visual analysis alone may not be able to capture the heterogeneity within the structures being imaged. Machine learning methods can uncover patterns that are not evident to the naked eye and can help doctors gain richer insights. Though imaging modalities such as the magnetic resonance imaging (MRI) and computational tomography have been around for a while, they can be further explored to provide new insights by extracting features using new machine learning methods. This opens up avenues to design methods that answer unresolved problems in biology. For example, analysis of texture in dynamic contrast enhanced MRI (DCE-MRI) images (which depict the vascular functionality) can help predict the tumor's response to chemotherapy at an early stage in breast cancer patients (see Chapter 6). Early prediction of response to chemotherapy can spare patients from potential short- and long-term toxicities associated with ineffective therapies.

Problem 3: Medical images getting richer and complicated: Recent advancements in 3D and 4D medical imaging provide real-time visualization of the human body. 4D medical imaging incorporates time information in addition to the volumetric imaging data. On the other end of the spectrum, advances in electron microscopy have enabled 3D visualization of the cellular ultrastructure. This nanometer resolution views of intra and intercellular interactions can provide

a deeper understanding of the underlying cellular mechanisms and could reveal the exact cause of a medical problem (see Chapter 5). Understanding of these dynamic interactions can facilitate the design and development of efficient treatment strategies. In order to utilize the potential of these technologies and make it accessible to a doctor for clinical applications, it is essential to develop new ways to organize, process and store the data being generated.

Overall, with continuous advancements in medical imaging, there is a constant need for improving our ability to process and interpret them. In this dissertation, we illustrate the potential of machine learning in automatic processing of medical images. We develop a range of machine learning techniques along the image processing pipeline to interpret medical images addressing specific medical problems.

1.2 Contributions and Overview

In Chapter 2 we lay the foundation for our work by introducing the four steps in the image processing pipeline dealt in this dissertation - registration, segmentation, characterization and tracking. We discuss the current approaches for each of these steps. In the final section of this chapter, we introduce the different imaging modalities analyzed in this dissertation.

Registration

In Chapter 3, we propose a new point set registration algorithm that preserves both the global and local structure of the point set. The aim of point set registration is to align two sets of points and obtain the transformation that maps one to the other [128]. Most of the current registration algorithms attempt to align the point sets at a global level, paying little attention to the local shape of the point set. Shape, represented as local proximity among points, is stable even in the presence of moderate distortions, and can be helpful in differentiating objects and recovering point correspondences. In this chapter, we propose a probability density estimation framework to align the point sets by incorporating the local structural relationships among neighboring points. The experimental results demonstrate that our approach outperforms other point set registration algorithms, even when the point sets are degraded by various levels of noise, outliers, deformation, occlusion, and rotation.

Segmentation

In Chapter 4, we experiment with various modifications to the U-Net architecture to improve its performance in segmenting tumors in MRI images of human prostates. Segmentation of prostate

tumors is an important step as it helps in further downstream analysis and is also used to guide in-bore biopsy to obtain an accurate diagnosis. Therefore, automatic techniques to accurately segment tumors from prostate scans is highly desirable. As the tumor occupies a small portion of the image, there is an imbalance in the tumor and non-tumor classes. We show that the Tversky loss [147], which was specially designed for datasets with imbalanced classes, improves the segmentation performance as the weight of the false positive and false negative terms could be controlled unlike the Dice loss function. We further modify the U-Net architecture by adding attention gates [131], using auto-encoder style regularization [127] and using deep supervision [103]. The attention gates prune the feature maps to retain only features relevant for tumor detection, thus reducing false positive predictions. The auto-encoder based regularization helps when the information in the image is sparse, but degrades the segmentation performance when the image contains rich diverse information as the features required for reconstruction of rich data could be different from those required for segmentation. Deep supervision forces the features to be highly discriminative and therefore increases the segmentation performance and also leads to faster convergence. As mentioned in problem 1 in section 1.1 automating the segmentation process is a step towards quicker analysis of the MRI image.

In Chapter 5, we present a framework for segmentation of cells and sub-cellular ultrastructure in electron microscopy images of tumor biopsies. As mentioned in problem 3 in section 1.1, electron microscopy is a rich source of information that can help understand cancer at cellular level. In order to enable interpretative rendering and quantitative characterization of biologically relevant features from the electron microscopy images, cells and intra-cellular organelles needs to be accurately segmented. These quantitative features can be potentially linked to biologically or clinically relevant variables such as patient drug response in downstream analysis. The intra-cellular organelles are well distinguished from the surrounding ultrastructures enabling accurate segmentation by training deep learning models on sparse manual labels. Cell segmentation on the other hand is a complex task due to the lack of clear boundaries separating cells in the cancer tissue. We propose a multi-pronged approach combining segmentation and tracking strategies for cell segmentation. Our proposed method cuts down the time taken for cell segmentation from 8 months to 1-2 days.

Characterization

In Chapter 6, we characterize the texture of DCE-MRI parametric maps with a new spectral texture analysis method called multi-resolution fractal analysis and evaluate its potential in early prediction of breast cancer response to chemotherapy. Fractal analysis at multiple resolutions

obtains a rich representation of the heterogeneity of the texture. Multi-resolution analysis filters out irrelevant features and noise at different resolutions, rendering more emphasis on distinct features, and fractal analysis at each level captures these distinct features. The multi-resolution fractal features better predict breast cancer response to chemotherapy than those extracted with the more conventional methods. As discussed in problem 2 in section 1.1, extraction of new features can provide deeper insights from medical images than visual inspection and can be used to infer a lot more than what the DCE-MRI images are currently being used for.

Tracking

In Chapter 7, we develop a hybrid tracking system, that combines measurements from a basic camera and wearable sensors, thus benefiting from the superior spatial and temporal resolution capabilities of either modality, to accurately track motion of infant limbs. We analyze the capability of features extracted from the estimated motion to classify fidgety movements from non-fidgety movements using support vector machine. We find that the proposed approach can not only accurately estimate trajectories, including complex ones, but can also detect fidgety movements with good accuracy. Tying back to problems 1, 2 and 3, automatic analysis of richer data with new analysis methods can provide ways to capture even subtle changes and aid doctors in disease diagnosis.

Finally in Chapter 8, we summarize our findings and present possible future directions.

Chapter 2

Background

In the first four sections of this chapter we introduce the four steps in the image processing pipeline dealt with in this dissertation - registration, segmentation, characterization and tracking. We present the current approaches and discuss the algorithms referenced throughout this dissertation. In the final section, we provide a brief introduction of the imaging modalities on which the proposed methods are applied.

2.1 Registration

Image registration is a fundamental component in computer vision that aligns two images and recovers the transformation mapping from one to another. The two images are of the same object but taken at different times, using different sensors, or from different perspectives. Its main application domains include medical imaging, remote sensing, and motion tracking [79, 22, 149]. Rigid image registration methods align the images by performing rotation, scaling and translation, while non-rigid image registration methods align the images through non-rigid geometric transformations as shown in Figure 2.1. Non-rigid transformations are common in medical imaging due to soft-tissue deformations and modeling them is a challenging task. Even though there has been extensive work done in this field, non-rigid image registration is still an open problem [81].

Image registration methods are classified into two categories: intensity-based and feature-based. Intensity-based methods compare the intensity values of different pixels in the images and try to find a transformation between them. Feature-based methods try to find a transformation by aligning features, such as corners and boundary points, extracted from the image. A set of features extracted from the image is called a point set. The process of finding the alignment between two point sets is called point set registration. An example of point set registration is shown in Figure 2.2. In this dissertation, we focus on point set registration.

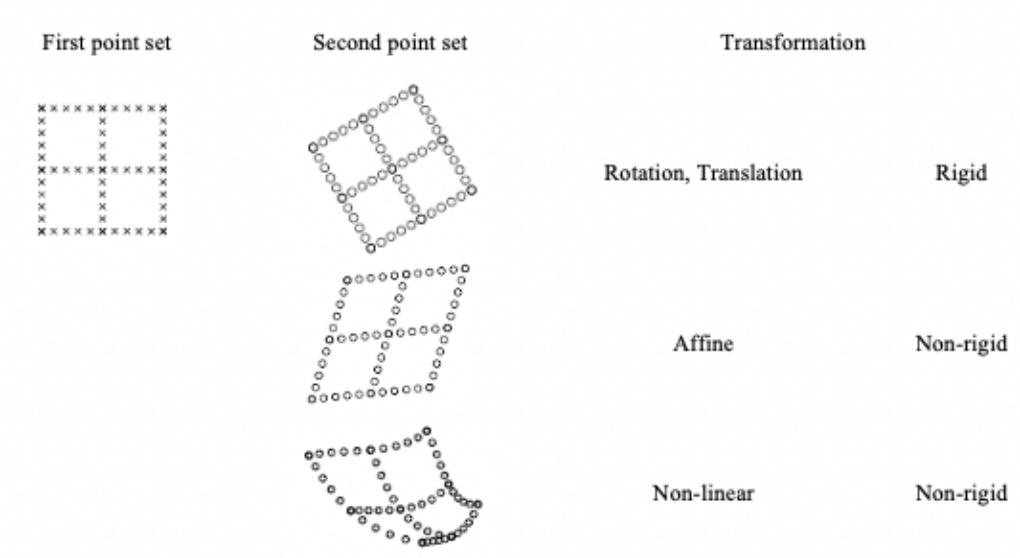


Figure 2.1: Examples of rigid and non-rigid transformations

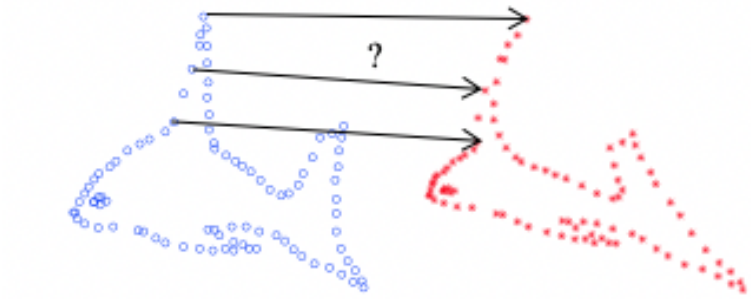


Figure 2.2: The point set registration problem: Finding a transformation that maps one point set to the other

2.1.1 Current Point Set Registration Approaches

The iterative closest point (ICP) algorithm is the most popular rigid point matching algorithm due to its simplicity and low computational complexity [24]. For each point in the first point set, the ICP algorithm finds the closest point in the second set. Then the algorithm uses a mean squared error cost function to estimate the rotation and translation values iteratively, until the two matched points are aligned. However, the ICP algorithm requires the points to be adequately close to each other.

To overcome this problem, probabilistic approaches were introduced [65, 40, 88, 128]. Among these approaches, the most popular is the robust point matching (RPM) algorithm introduced by

Gold et al. [65]. In their work, points in one set are represented as centroids of Gaussian mixture models, while points in the other set are treated as data points. The alignment of points is estimated by alternating between soft-assignment, and using the expectation-maximization algorithm to estimate the new location of the centroids. The expectation-maximization algorithm consists of two steps: the expectation step, which computes the probabilities, and the maximization step, which evaluates the transformation. In most of the probability-based methods, a uniform distribution is added to account for noise and outliers. Outliers are incorrectly extracted points that do not correspond with any point in the other point set. The addition of this distribution makes probability-based methods perform better than the ICP algorithm in the presence of noise and outliers. Chui and Rangarajan used the thin-plate splines (TPS) parameterization on RPM (resulting in the TPS-RPM algorithm) to find the transformation and the correspondence between sets of points simultaneously [40, 39]. Jain and Vemuri proposed an algorithm (GMMREG) that represents both point sets as centroids of Gaussians [87, 88]. One of the point sets is parameterized as TPS, and the transformation parameters are estimated by minimizing the L2 norm [37] between the distributions. The TPS formulation does not exist for dimensions greater than three, which limits the generalization of this approach to multidimensional cases [128].

More recently, Myronenko and Song introduced the Coherent Point Drift (CPD) algorithm [128]. They also treated point set registration as a density estimation problem and used a variational approach to derive a multidimensional solution. CPD uses radial basis functions instead of TPS to simultaneously find both the transformation and correspondence between the two point sets. The CPD algorithm uses a motion coherence constraint, introduced by Yuille et al. [195], that ensures that all the points move in a coherent fashion, preserving the topological structure. It assumes equal membership probability for each GMM component, and only considers the distance to calculate similarity and does not consider the local structure of the point set.

Though the probability-based algorithms perform well, there are still few shortcomings. They do not work well in the presence of large in-plane rotations. Typically, the point sets are compensated for rotation by other techniques before applying the registration algorithms. In most cases, the weight of the outlier distribution has to be predetermined by the user.

All the algorithms discussed so far preserve the global topological structure. As an attempt to include local structural information, Ge and Fan included local linear embedding (LLE) and Laplacian coordinate based energy terms as additional regularization terms into the CPD formulation [58]. The LLE formulation preserves the structure by retaining neighborhood relationships, but is insensitive to scaling. Ma and colleagues showed that the addition of local structural information to the CPD algorithm can be helpful in recovering point correspondences [116]. They used shape

context [21] as a local structural descriptor and the Hungarian method to match the points [135]. Using this approach slows down the CPD algorithm as the Hungarian method has cubic time complexity ($O(n^3)$). Also, during the process of matching, a fixed probability is assigned in an ad-hoc manner to matched and unmatched points. The calculation of shape context is complex in three or higher dimensions. To overcome these problems, in Chapter 3, we propose an algorithm to preserve both the global and local structure of the point set without significant increase in the time complexity. We compare our algorithm with GMMREG, CPD and Ma and coll algorithm which we call CPD-SC.

2.2 Segmentation

Image segmentation is a sub-domain of computer vision which aims at dividing an image into different regions. Grouping of similar pixels simplifies the information represented in the image and allows for efficient analysis. It is the step following pre-processing in most image processing pipelines. Segmentation therefore finds applications in various fields including robotics, autonomous vehicles and medical imaging [153, 176, 131]. Previously, image processing methods such as region growing, edge detection and thresholding often coupled with optimization algorithms were used to segment images. Watershed segmentation is one such region growing method used for separating different objects in an image [16, 25]. It represents the gray scale image as a topographic relief, flooded with water, where watershed are lines dividing areas of water from different basins. The basic idea of watershed algorithm consists of placing a water source in each local minimum in the relief and flooding the entire relief from these sources. Watershed is built in places where waters from different sources meet. These watershed lines represent the separation between the objects.

Recently, deep convolutional neural networks (CNN) with their rich feature representation capability have achieved extraordinary performance in image segmentation [117, 8, 124]. Specifically, U-Net, introduced by Ronneberger et al., with an encoder-decoder architecture is the most prominent network and U-Net along with its variants are widely used in medical imaging applications [145].

U-Net consists of a contracting path that encodes the input image into feature representations and a symmetric expanding path which projects the discriminative features learned by the contracting path back to the spatial resolution of the input image to obtain a dense segmentation map. This unique architecture captures global image context while preserving spatial accuracy, thereby facilitating high-precision image segmentation. Each layer in the encoding path consists of two 3x3 convolutions each followed by a rectified linear unit (ReLU) as the activation function and a 2x2

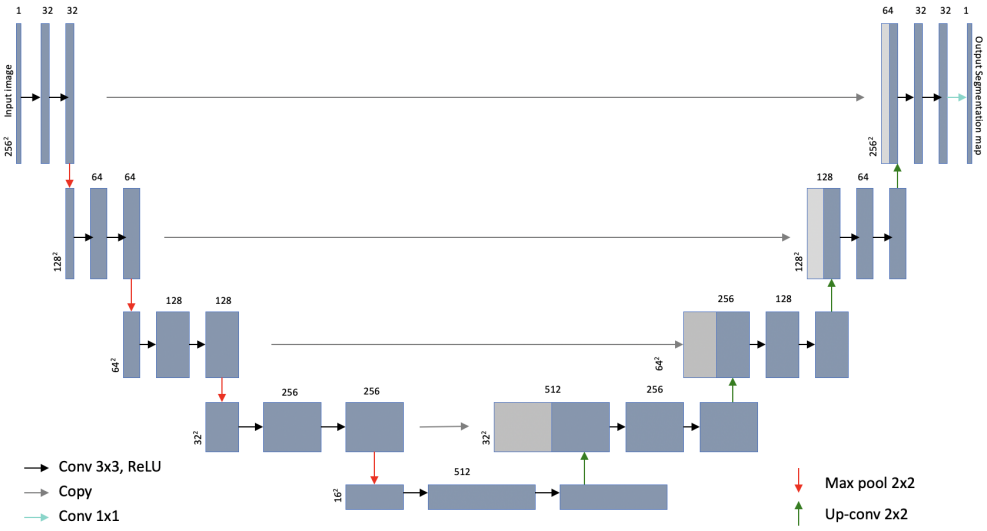


Figure 2.3: U-net architecture. Input size is written on the side of each box. The number of feature maps in each layer is written on top of each box.

max-pooling operation to down-sample the feature maps (except the last layer). The number of feature maps is doubled along each successive layer in the encoding path to enable richer feature extraction. Each layer along the decoding path consists of up-sampling feature maps followed by a 2x2 convolution (to half the number of feature maps), a concatenation of feature maps from the encoding path and two 3x3 convolutions each followed by a Rectified Linear Unit (ReLU). An architecture of U-Net is shown in Figure 2.3.

The skip connections in U-Net brings across many redundant low-level features to the decoder as feature representation is poor in the initial layers of the encoder. Oktay et al. proposed an attention mechanism to highlight only the relevant features in the encoder layers to pass through the skip connections [131]. Attention mechanism is incorporated in the form of attention gates integrated into the U-Net architecture before feature concatenation. The attention gates prune the feature maps by generating an attention coefficient, which identifies relevant features. An attention coefficient $\alpha_i^l \in [0, 1]$ is generated for each pixel i in layer l based on the feature map from the encoding layer x_i^l and a gating vector g_i from a coarser scale. The gating signal is used to remove irrelevant and noisy responses in the feature map from the encoding layer. The output of the attention mechanism is obtained by multiplying attention coefficients element-wise with the encoder feature map x_i^l thus highlighting only the relevant regions in the feature maps. The attention coefficients are obtained using the following equations.

$$q_{att}^l = \psi^T \left(\sigma_1 \left(W_x^T x_i^l + W_g^T g_i^l + b_g \right) \right) + b_\psi \tag{2.1}$$

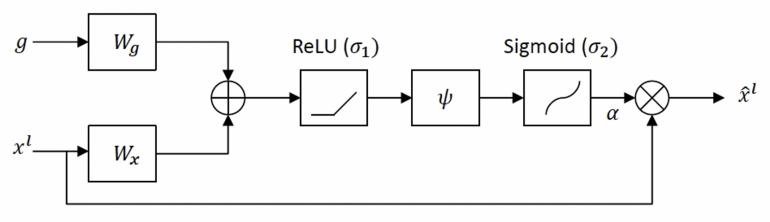


Figure 2.4: Schematic of the attention gate block. Input feature map x^l are multiplied with the attention coefficients (α) to generate a new map $\hat{x}^l$ in which only the relevant features are highlighted

$$\alpha_i^l = \sigma_2(q_{att}^l(x_i^l, g_i)) \tag{2.2}$$

where σ_1 is the ReLU activation function, σ_2 is the sigmoid activation function, W_x, W_g, ψ, b_g and b_ψ are the linear transformation parameters. The schematic of the attention gate is shown in Figure 2.4.

However, as the depth of the U-Net increases, the gradient information propagated through the layers to update the network weights gets vanishingly small, thus preventing effective training of the network [64]. The residual learning technique was proposed to solve this problem by adding a shortcut mapping to reuse data across subsequent layers such that each layer only fits a residual mapping. The gradients could easily backpropagate through the shortcut connection making optimization of deeper networks feasible and faster [75]. The resulting ResUNet model architecture combines the strengths of feature concatenation from the contracting and expanding structure of U-Net and residual connections, facilitating propagation of local spatial information and ease the training of the neural network.

Though the convolution operation in CNN helps capture good features it still has a few drawbacks. Due to the small size of the convolution filter, the network does not encode the relative position of different features. Large filters are required to encode long-range dependencies within an image. However, using large filters will negate the computational and statistical efficiency obtained by using local convolutional operations. In order to increase the receptive field without greatly increasing computational efficiency, Yu and Kolten introduced dilated convolutions [193]. The dilated convolution operator applies the same convolution filter with different dilation rates resulting in a lager receptive field. But even this operator increases the receptive field only to a certain range and does not provide context from the entire image.

In the field of natural language processing, Vaswani et al. introduced Transformers - a self-attention-based architecture - that allows every element in a sequence to interact with all other elements and make an informed decision on whom to pay more attention to [180]. The self-attention

mechanism helps update each component of the sequence by aggregating global information from the complete input sequence. Inspired by their success in natural language processing, Dosovitskiy et al. applied Transformers on images by splitting an image into patches and providing a linear embedding of these patches as an input sequence to the Transformer [48]. They observed that Transformers outperformed CNNs in image recognition tasks in the presence of large datasets. However they did not generalize well on smaller datasets as Transformers lack some of the inductive biases inherent to CNNs, such as locality and translational equivariance. Chen et al. designed a hybrid CNN-Transformer architecture, called TransUNet, to leverage both spatial information from CNN features and global context from Transformers [34]. They demonstrated improved results with TransUNet over pure convolutional-based ResUNet and pure attention-based vision transformer [48] in segmenting multiple organs in CT images. Zheng et al. designed a pure Transformer-based encoder combined with a simple decoder to yield a powerful segmentation model, termed SEgmentation TRansformer (SETR), achieving state-of-the-art results on several large image segmentation datasets [200]. In Chapters 4 and 5 we look at the application of U-Net and its variants and Transformers in different medical image segmentation problems.

2.3 Characterization

The next step in the image processing pipeline after segmenting a region of interest is to characterize the region. Characterization refers to the process of defining a set of features that most efficiently represent the information essential for downstream analysis. Color, shape and texture are the three commonly used features to characterize regions in an image. As most medical images are in gray scale, color can often not be used as a distinguishing feature. Therefore, shape and texture are the features predominantly used to characterize medical images. The shape and texture features used in this dissertation are described in Sections 2.3.1 - 2.3.4. In all our applications we characterize 3D volumes of different objects and therefore all features are calculated in 3D.

2.3.1 Shape Features

Shape plays an important role in the perception of objects and shape features capturing the holistic structure can provide a powerful clue in distinguishing different objects. Shape features can be broadly classified into two groups - contour-based and region-based. Contour-based features such as perimeter, curvature, bending energy and shape signature [61] represent the shape using only the information present in the contour of the object. On the other hand, region-based features such as area and eccentricity, use the entire area of the object for shape description.

In this dissertation we use shape features to characterize the roundness of nuclei and nucleoli in cells. Solidity, sperificity and circular variance are three shape features that characterize roundness of objects. *Solidity* quantifies the concavities of a surface and is calculated as the ratio of the volume of the object to the volume of the convex hull (smallest encompassing convex polygon) of the object.

$$\text{Solidity} = \frac{\text{Volume}}{\text{Convex Hull Volume}} \tag{2.3}$$

Sphericity is a measure of how close an object resembles a sphere. It is defined as the ratio of the surface area of a sphere with same volume as the object to the surface area of the object.

$$\text{Sphericity} = \frac{\pi^{1/3}(6V)^{2/3}}{A} \tag{2.4}$$

where V is the volume of the object and A is the surface area of the object [184]. Its value is 1 for a perfect sphere and decreases as the shape varies from a sphere. *Circularity variance* provides a measure of the spread of radii across the volume. A radius here denotes the distance between a point on the contour and the centroid (geometric center) of the volume. The lower the value, the tighter the clustering about a single mean.

$$\text{Circularity varaince } (O) = \frac{1}{|Surf(O)|\mu_r^2} \sum_{p \in Surf(O)} (\|p - C\| - \mu_r)^2 \tag{2.5}$$

where $Surf$ is the surface contour of the object O, μ_r its mean radius and C its centroid.

2.3.2 Texture Features

Texture features measure the variations in the spatial arrangement of intensity values in an image to quantify properties such as regularity and smoothness. Based on the domain from which the texture features are extracted they are broadly divided into spatial and spectral texture features. Spatial texture features are extracted by computing statistics on local pixel structures in the original image. Spectral texture features transform an image into the frequency domain and calculate features in the transformed domain. The spatial texture features - grey level co-occurance matrix (GLCM), run-length matrix (RLM), size-zone matrix (SZM) and power spectrum - are discussed in Section 2.3.3 and the spectral texture features - wavelet and fractal analysis - are discussed in Section 2.3.4.

2.3.3 Spatial Texture Features

Grey Level Co-occurrence Matrix (GLCM)

A GLCM is created by calculating how often pairs of pixels with specified values and in a specified spatial relationship occur in an image. GLCM summarizes the number of times two pixel values

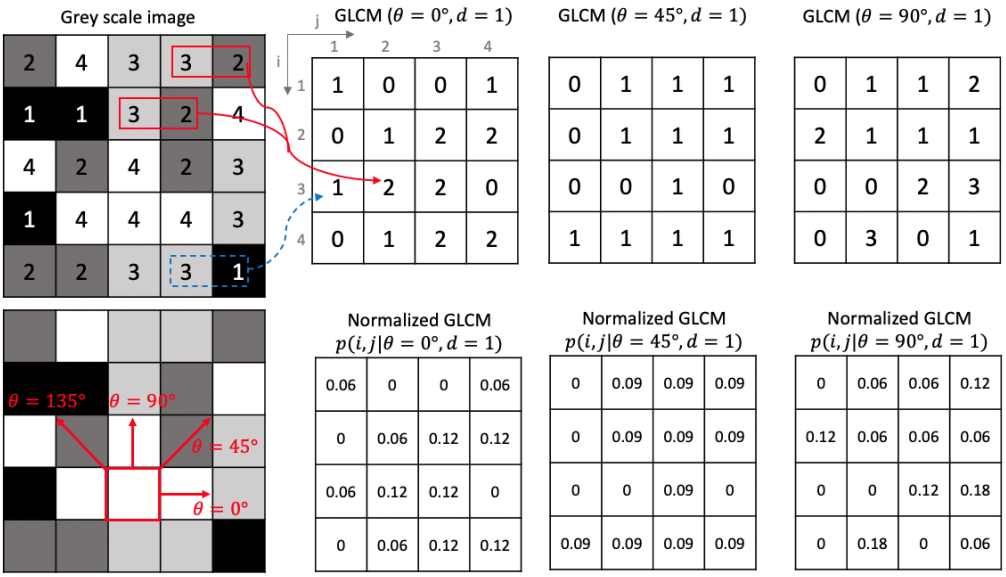


Figure 2.5: Top row shows a 4×4 grey-scale image, its corresponding GLCM for distance $\delta = 1$ and angle $\theta = 0°$, the normalized GLCM and some examples of Haralick texture features. The bottom row shows the construction of GLCM for distance $\delta = 1$ and angles $\theta = 45°, 90°$ and $135°$.

occur together in a given direction θ and at a given distance δ. GLCMs are constructed for different values of distances δ and angle θ. If N_g is the number of gray levels in an image, a GLCM of size $N_g \times N_g$ is constructed such that, the $(i,j)^{th}$ element of this matrix represents the number of times the combination of levels i and j occur in two pixels in the image, that are separated by a distance of δ pixels along angle θ. The GLCM is then normalized to represent the probability distribution $p(i,j|\delta,\theta)$ of the gray-level pairs in the image. Haralick texture features are then computed from the normalized GLCM using equations in Table 2.1. Haralick features are extracted from each of the GLCM constructed for different values of distances δ and angle θ. The final value of a feature is taken as the average across all its values calculated from the individual GLCMs over different values of δ and θ. An example of constructing a GLCM from an image having 4 grey level values for distance $\delta = 1$ and angles $\theta = 0°$, $45°$ and $90°$ are shown in Figure 2.5. Haralick texture features computed from the GLCM are widely used to characterize textures as they are easy to calculate and result in interpretable texture descriptors [73].

Run-Length Matrix (RLM)

A gray level Run Length Matrix (RLM) quantifies gray level runs, which are defined as the number of consecutive pixels that have the same gray level value in a given direction [57]. In a RLM, $P(i,r|\theta)$ is defined as the number of pixels with gray-level i and run-length r, along a given

Table 2.1: Haralick texture features.*

Feature	Equation	What it measures
Energy	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)^2$	Uniformity of grey level distribution
Contrast	$\sum_{k=0}^{N_g-1} k^2 \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \delta_{\lvert i-j\rvert,k} p(i,j)$	Amount of local gray level variations
Correlation	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i\cdot j)p(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y}$	Linear dependancy of grey levels on neighboring pixels
Variance	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-j)^2 p(i,j)$	Dispersion of gray level distributions
Homogeneity	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{p(i,j)}{1+(i-j)^2}$	Smoothness of the gray level distribution
Sum average	$\sum_{i=2}^{2N_g} i p_{x+y}(i)$	Mean of the gray level sum distribution
Sum variance	$\sum_{i=2}^{2N_g}\left(i - \left[\sum_{i=2}^{2N_g} i p_{x+y}(i)\right]\right)^2$	Dispersion of the gray level sum distribution
Sum entropy	$-\sum_{i=2}^{2N_g} p_{x+y}(i)\log p_{x+y}(i)$	Disorder related to gray level sum distribution
Entropy	$-\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log[p(i,j)]$	Degree of disorder among pixels
Difference variance	$\sum_{i=2}^{2N_g}\left(i - \left[\sum_{i=2}^{2N_g} i p_{x-y}(i)\right]\right)^2$	Dispersion of gray level difference distribution
Difference entropy	$-\sum_{i=2}^{2N_g} p_{x-y}(i)\log p_{x-y}(i)$	Disorder related to the grey level difference distribution
Maximal correlation coefficient	$\sqrt{2^{\text{nd}} \text{ largest eigenvalue of } \sum_k \frac{p(i,k)p(j,k)}{p_x(i)p_{y(k)}}}$	Complexity of the texture

*N_g is the number of discrete intensity levels in the image, μ_x, σ_x and μ_y, σ_y are the mean and standard deviation of p_x and p_y respectively, where p_x and p_y are the marginal row and column probabilities. $p_{x+y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)$, where $i+j=k$ and $k = 2,3,...,2N_g$. $p_{x-y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)$, where $|i-j| = k$ and $k = 0,1,...,N_g-1$.

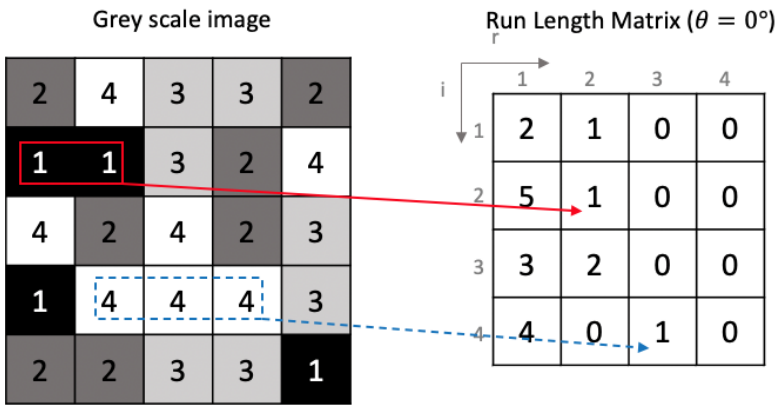


Figure 2.6: Construction of RLM

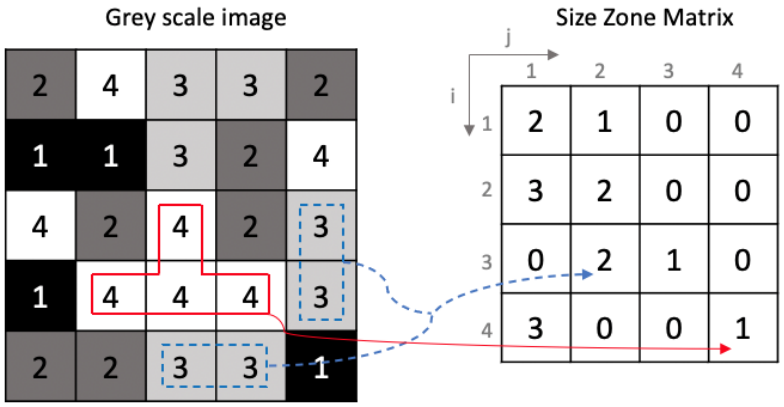


Figure 2.7: Construction of SZM

direction θ. Features defining the underlying texture are extracted from RLM by using equations in Table 2.2. Similar to features in GLCM, the final value of each feature is computed as the mean of all values calculated along each angle separately. An example of constructing a RLM from a 5×5 image having 4 discrete values for angle $\theta = 0°$ is shown in Figure 2.6.

Size Zone Matrix (SZM)

A gray level Size Zone Matrix (SZM) provides a statistical representation of the clusters of gray levels in the image [168, 169, 167]. Each element $P(i, j)$ in a SZM is equal to the number of zones of size j with gray level i. Unlike GLCM and RLM, SZM is rotation invariant. Features are extracted from SZM by using equations in Tables 2.3 and 2.4. An example of constructing a SZM for a 5×5 image having 4 discrete values is shown in Figure 2.7. SZM is useful in characterizing texture homogeneity [167, **?**].

Table 2.2: Run Length Matrix features.*

Feature	Equation	What it measures
Short Run Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\|\theta)}{j^2}}{N_r(\theta)}$	Distribution of short run lengths
Long Run Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}P(i,j\|\theta)j^2}{N_r(\theta)}$	Distribution of long run lengths
Gray Level Non-Uniformity	$\frac{\sum_{i=1}^{N_g}\left(\sum_{j=1}^{N_r}P(i,j\|\theta)\right)^2}{N_r(\theta)}$	Similarity of gray-level intensity values
Run Length Non-Uniformity	$\frac{\sum_{j=1}^{N_r}\left(\sum_{i=1}^{N_g}P(i,j\|\theta)\right)^2}{N_r(\theta)}$	Similarity of run lengths
Run Percentage	$\frac{N_r(\theta)}{N_p}$	Coarseness of the texture
Gray Level Variance	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,j\|\theta)(i-\mu)^2$	Variance in gray level intensity for the runs
Run Variance	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,j\|\theta)(j-\mu)^2$	Variance in runs for the run lengths
Low Gray Level Run Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\|\theta)}{i^2}}{N_r(\theta)}$	Distribution of low gray-level values
High Gray Level Run Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}P(i,j\|\theta)i^2}{N_r(\theta)}$	Distribution of higher gray-level values
Short Run Low Gray Level Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\|\theta)}{i^2j^2}}{N_r(\theta)}$	Joint distribution of shorter run lengths with lower gray-level values
Short Run High Gray Level Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\|\theta)i^2}{j^2}}{N_r(\theta)}$	Joint distribution of shorter run lengths with higher gray-level values
Long Run Low Gray Level Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\|\theta)j^2}{i^2}}{N_r(\theta)}$	Joint distribution of long run lengths with lower gray-level values
Long Run High Gray Level Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}P(i,j\|\theta)i^2j^2}{N_r(\theta)}$	Joint distribution of long run lengths with higher gray-level values

*N_g is the number of discrete intensity values, N_r is the number of discrete run lengths, N_p is the number of voxels in the image, $N_r(\theta)$ is the number of runs in the image along angle θ, $P(i,j|\theta)$ is the RLM for an arbitrary direction θ and $p(i,j|\theta)$ is the normalized RLM defined as $p(i,j|\theta) = \frac{P(i,j|\theta)}{N_r(\theta)}$.

Table 2.3: Size Zone Matrix features - Part 1.*

Feature	Equation	What it measures
Short Area Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{P(i,j)}{j^2}}{N_z}$	Distribution of small size zones
Large Area Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}P(i,j)j^2}{N_z}$	Distribution of large area size zones
Gray Level Non-Uniformity	$\frac{\sum_{i=1}^{N_g}\left(\sum_{j=1}^{N_s}P(i,j)\right)^2}{N_z}$	Variability of gray-level intensity values
Gray Level Non-Uniformity Normalized	$\frac{\sum_{i=1}^{N_g}\left(\sum_{j=1}^{N_s}P(i,j)\right)^2}{N_z^2}$	Variability of gray-level intensity values
Size-Zone Non-Uniformity	$\frac{\sum_{j=1}^{N_s}\left(\sum_{i=1}^{N_g}P(i,j)\right)^2}{N_z}$	Variability of size zone volumes
Size-Zone Non-Uniformity Normalized	$\frac{\sum_{j=1}^{N_s}\left(\sum_{i=1}^{N_g}P(i,j)\right)^2}{N_z^2}$	Variability of size zone volumes
Zone Percentage	$\frac{N_z}{N_p}$	Coarseness of the texture
Gray Level Variance	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}p(i,j)(i-\mu)^2$	Variance in gray level intensities for the zones
Zone Variance	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}p(i,j)(j-\mu)^2$	Variance in zone size volumes for the zones
Zone Entropy	$-\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}p(i,j)\log_2(p(i,j)+\epsilon)$	Uncertainty in the distribution of zone sizes and gray levels.
Low Gray Level Zone Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{P(i,j)}{i^2}}{N_z}$	Distribution of lower gray-level size zones
High Gray Level Zone Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}P(i,j)i^2}{N_z}$	Distribution of higher gray-level size zones

*N_g is the number of discrete intensity values, N_s is the number of discrete zone sizes, N_p is the number of voxels in the image, N_z is the number of zones in the image, $P(i,j)$ is the SZM and $p(i,j|\theta)$ is the normalized SZM defined as $p(i,j) = \frac{P(i,j)}{N_z}$.

Table 2.4: Size Zone Matrix features - Part 2

Feature	Equation	What it measures
Small Area Low Gray Level Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{P(i,j)}{i^2j^2}}{N_z}$	Joint distribution of smaller size zones with lower gray-level values
Small Area High Gray Level Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{P(i,j)i^2}{j^2}}{N_z}$	Joint distribution of smaller size zones with higher gray-level values
Large Area Low Gray Level Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{P(i,j)j^2}{i^2}}{N_z}$	Joint distribution of larger size zones with lower gray-level values
Large Area High Gray Level Emphasis	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}P(i,j)i^2j^2}{N_z}$	Joint distribution of larger size zones with higher gray-level values

Power spectrum

Power spectrum is a gray level rotation and intensity invariant texture descriptor [170]. It describes the shape and size of structures in an image. It is calculated by iteratively opening and closing an image with morphological operators and recording the resulting areas. Morphological operators are a set of image processing operators that process images based on shapes. They apply a structuring element to an input image, creating an output image of the same size. A structuring element is a matrix with the size and shape of the object that needs to be analyzed in the input image. For example to find a line, the structuring element represents a line. If n is the size factor for a structuring element, then the family of morphological openings $\Gamma = (\gamma_n)_{n\geq 0}$ and morphological closings $\Phi = (\phi_n)_{n\geq 0}$ can help in the study of all the object sizes present in an image. The analysis of an image f with respect to an operator for example Γ, involves evaluating each opening of size n with a measurement $\int \gamma_n(f)$. The power spectrum curve of an image f with respect to Γ and Φ is defined as the following.

$$PS_n(f) = \frac{1}{\int f}\begin{cases}\int \gamma_n(f) - \int \gamma_{n+1}(f) \text{ for } n \geq 0 \\ \int \phi_n(f) - \int \phi_{n-1}(f) \text{ for } n < 0\end{cases} \tag{2.6}$$

It defines a probability distribution and the moments of this distribution are employed as signature patterns to characterize textures. A peak in power spectrum at a given scale n indicates the presence of many image structures of this scale or size.

2.3.4 Spectral Texture Features

Wavelet Analysis

Wavelet analysis is used to decompose a signal into a set of frequency sub-bands based on small basis functions of varying frequency and limited time duration, called wavelets. It can decompose special patterns hidden in data and is used extensively to extract information from audio signals and images. The wavelet is scaled and translated to cover the time-frequency domain which enables study of local characteristics in this domain. Wavelet transforms take a signal and express it in terms of scaled and translated wavelets. The resulting wavelet transform is a representation of the signal at different scales. Decomposition of image using wavelets enables the characterization of texture as it analyzes the image at different resolutions [185]. The discrete wavelet transform for a function $f(x, y, z)$ of size (M, N, K) can be represented as

$$W_\phi(j_0, m, n, k) = \frac{1}{\sqrt{MNK}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \sum_{z=0}^{K-1} f(x, y, z) \phi_{j_0,m,n,k}(x, y, z), \tag{2.7}$$

$$W_\phi(j_0, m, n, k) = \frac{1}{\sqrt{MNK}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \sum_{z=0}^{K-1} f(x, y, z) \psi^i_{j,m,n,k}(x, y, z), \quad i = \{H, V, D\}, \tag{2.8}$$

where, $W_\phi(j_0, m, n, k)$ is the approximation of $f(x, y, z)$ at scale j_0, $W_\phi(j_0, m, n, k)$ coefficients define the horizontal (H), vertical (V), and diagonal (D) details for scales $j \geq j_0$, ϕ is the scaling function and ψ is the wavelet function. The wavelet transform depends mainly on the scaling (ϕ) and wavelet (ψ) functions, but it is not necessary to define their explicit form. Instead, a low pass and high pass filter that characterize the interaction of these functions are used.

Fractal Analysis

Textures can also be characterized by fractals, which describe irregular structures that show self-similarity at various scales. Fractal based texture analysis correlates texture heterogeneity to fractal dimension (FD), which is a mathematical descriptor of a structure's geometrical complexity, based on the concept of spatial pattern self-similarity. The fractal dimension is calculated based on the power spectrum analysis of the Fourier transformation of the 3D volume [97]. The 3D discrete Fourier transform is defined as

$$F(x, y, z) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \sum_{z=0}^{K-1} I(m, n, k) e^{-j2\pi\left(x\frac{m}{M} + y\frac{n}{N} + z\frac{k}{K}\right)}, \tag{2.9}$$

where $I(m, n, k)$ is the 3D volume of size (M, N, K) and $x,\ y$ and z are the spatial frequencies. The power spectral density (P) is estimated as the following.

$$P(x, y, z) = |F(x, y, z)|^2, \tag{2.10}$$

The frequency space is evenly divided into 12 zenith and 24 azimuth directions, and 30 points are uniformly sampled along the radial component in each of these directions. The power spectral density is plotted against sampled radial frequency in a log-log plot. The slope (β) of a least-squares regression line of the log-log plot is related to the fractal dimension as

$$FD = \frac{11 - \beta}{2}. \tag{2.11}$$

Convolutional Neural Networks

All the above mentioned methods aim at extracting features defining a specific attribute, shape or texture, from the image. On the other hand, convolutional neural networks have demonstrated powerful overall feature extraction capabilities for downstream analysis such as classification or segmentation. A CNN has multiple convolutional layers extracting features from the inputs. Each convolutional layer has multiple filter banks that compute different features from the input. The encoder part of the U-Net described in Section 2.2 acts as a feature extractor. Unlike other texture analysis methods which have pre-designed filter banks, CNN can learn appropriate features from the data provided through forward and backward propagation during training. Given a set of training images x_i and their labels y_i, CNN learns the filter banks (W) by solving an optimization problem

$$\min_{W} \sum_{i=1}^{m} l(f(W; x_i), y_i) \tag{2.12}$$

where $f(W; x_i)$ is the prediction of the network for input x_i, and l is the log-likelihood loss function defined in terms of the normalized soft-max probability.

Though standalone CNN features cannot be used to describe specific attributes of the image, they are more powerful in capturing the differences between images than the shape and texture features as they are learnt with the objective to accurately differentiate the images belonging to different classes. We use the features described in this section to characterize different medical images in Chapters 5 and 6.

2.4 Tracking

A sequence of images is a richer source of information than a still image as it captures the motion in the scene. Tracking the motion of objects can provide objective measurements characterizing their movement. In this section, we delve into the fourth component of the image analysis pipeline dealt in this dissertation - Tracking. Tracking objects in a video is an important step in many computer vision applications such as autonomous driving, surveillance and robotics [105, 5, 178].

In the medical field, tracking algorithms are used in many applications including tracking of cells in microscopy images, instruments in surgery videos [43] and movements of eye in a video. Eye-tracking is used as a diagnostic tool for neural disorders such as schizophrenia and Autism [74]. We discuss two commonly used tracking algorithms - optical flow and Kalman filter in Sections 2.4.1 and 2.4.2.

2.4.1 Optical Flow

Optical flow is a widely used approach to track the movement of objects. It gives the flow vectors representing the apparent motion of individual pixels between two images [54]. There are multiple methods of estimating optical flow between two frames, including differential, energy, phase and region-based methods. The differential method is the most commonly used method and is based on the assumption that for really small space-time steps the brightness of each pixel in two consecutive image frames remains constant. Given a sequence of images, if $I(x, y, t)$ represents the intensity of a pixel (x, y) of the t^{th} frame and $(\delta x, \delta y, \delta t)$ is the small change in movement between two consecutive frames, the brightness constancy assumption can be expressed as the following.

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t) \tag{2.13}$$

The differential methods expand the above equation using Taylor series expansion to arrive at the brightness constancy equation

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \tag{2.14}$$

where V_x, V_y are the x and y components of the velocity or optical flow and $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y}$ and $\frac{\partial I}{\partial t}$ are the derivatives of the image at (x, y, t) in the corresponding directions. Since there are two unknowns per pixel but only one equation, the problem is under constrained. Horn and Schunck solve this problem by assuming that the underlying optical flow is smooth and add a smoothness constraint leading to the minimization of the energy functional given by the following equation [198].

$$J(u, v) = \int\int \left(\left(\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} \right)^2 + \lambda(u_x^2 + u_y^2 + v_x^2 + v_y^2) \right) dxdy \tag{2.15}$$

Lucas and Kanade, on the other hand, solve the brightness constancy equation by assuming that the optical flow is constant within a small window [198]. They use a least square approach to solve for optical flow by combining together several brightness constancy equations for a given window $(N \times N)$, i.e. they minimize the following quantity.

$$J(u, v) = \sum_{i=1}^{N^2} \left(\frac{\partial I}{\partial x_i}V_x + \frac{\partial I}{\partial y_i}V_y + \frac{\partial I}{\partial t_i} \right)^2 \tag{2.16}$$

Lucas-Kanade method is a local method as it solves several small problems independently (and therefore is more robust to noise), while the Horn and Schunck method is a global method as it solves the problem globally to yield dense flow fields.

Alternatively, Farnebäck proposed an optical flow algorithm based on polynomial expansion to obtain dense optical flow [52]. The algorithm first approximates the neighborhood of each pixel with a polynomial

$$f_1(x) = x^T A_1 x + b_1^T x + c_1 \tag{2.17}$$

where A is a symmetrix matrix, b a vector and c a scalar. The coefficients of the polynomial are estimated by a least square fit to the intensity values in the neighborhood considered. If the displacement between two images is represented as d, the new pixel is constructed as the following.

$$\begin{aligned} f_2(x) &= f_1(x-d) \\ &= (x-d)^T A_1 (x-d) + b_1^T (x-d) + c_1 \\ &= x^T A_1 x + (b_1 - 2A_1 d)^T x + d^T A_1 d - b_1^T d + c_1 \\ &= x^T A_2 x + b_2^T x + c_2 \end{aligned} \tag{2.18}$$

Equating the coefficients of both the polynomials yields the following equations.

$$A_2 = A_1 \tag{2.19}$$

$$b_2 = b_1 - 2A_1 d \tag{2.20}$$

$$c_2 = d^T A_1 d - b_1^T d + c_1 \tag{2.21}$$

The global displacement d can be solved by using Equation 2.20.

$$2A_1 d = -(b_2 - b_1) \tag{2.22}$$

$$d = -\frac{1}{2} A_1^{-1} (b_2 - b_1) \tag{2.23}$$

Since it is unrealistic to assume an entire image to be represented by a single polynomial, the global variables in Equation 2.17 are replaced by local polynomial approximations, giving coefficients $A_1(x), b_1(x)$ and $c_1(x)$ for the first image and $A_2(x), b_2(x)$ and $c_2(x)$ for the second image. Ideally $A_2 = A_1$ according to Equation 2.19, but in practice it is approximated as $A(x) = \frac{A_1(x)+A_2(x)}{2}$. Also $-\frac{1}{2}(b_2(x) - b_1(x))$ is represented as $\Delta b(x)$ to obtain the primary constraint

$$A(x)d(x) = \Delta b(x) \tag{2.24}$$

where $d(x)$ indicates the spatially varying displacement field instead of the global displacement field d in Equation 2.18. Farnebäck assumes that the displacement field is slowly varying and

therefore tries to find $d(x)$ satisfying 2.18 over a neighborhood I of the pixel x by minimizing

$$\sum_{\Delta x \in I} w(\Delta x) \left\| A(x + \Delta x)d(x) - \Delta b(x + \Delta x) \right\|^2 \tag{2.25}$$

where $w(\Delta x)$ is a weight function for points in the neighborhood. Therefore the minimum is obtained as the following.

$$d(x) = \left(\sum wA^T A \right)^{-1} \sum wA^T \Delta b \tag{2.26}$$

If the displacement field is parameterized by a motion model such as the eight parameter model given as

$$\begin{aligned} d_x(x, y) &= a_1 + a_2 x + a_3 y + a_7 x^2 + a_8 xy \\ d_y(x, y) &= a_4 + a_5 x + a_6 y + a_7 xy + a_8 y^2 \end{aligned} \tag{2.27}$$

then the displacement can be rewritten as the following.

$$d = Sp \tag{2.28}$$

$$S = \begin{pmatrix} 1 & x & y & 0 & 0 & 0 & x^2 & xy \\ 0 & 0 & 0 & 1 & x & y & xy & y^2 \end{pmatrix} \tag{2.29}$$

$$p = (a_1\ a_2\ a_3\ a_4\ a_5\ a_6\ a_7\ a_8) \tag{2.30}$$

Inserting into Equation 2.25, the weighted least square problem can be rewritten as

$$\sum_i w_i \left\| A_i S_i p - \Delta d_i \right\|^2 \tag{2.31}$$

where i is used to index the coordinates in a neighborhood. The solution for the parameter p is given as the following.

$$p = \left(\sum_i S_i^T A_i^T A_i S_i \right)^{-1} \sum_i S_i^T A_i^T \Delta b_i \tag{2.32}$$

In order to account for large displacements, the motion is first estimated at a coarser scale to obtain a rough estimate of the displacement. This displacement is propagated through finer scales to obtain increasingly accurate estimates. We use the Farnebäck's optical flow method in Chapter 5 to propagate contour along images in a stack of electron microscopy images.

2.4.2 Kalman Filter

The Kalman filter is an algorithm that estimates the state of a system from measurements observed over time [150]. It enables tracking of objects while fusing data from multiple sources. A dynamic system is a system that evolves through time and the system's state is a set of parameters that

describe the system at time k. A dynamic model of the system defines the evolution of state x from time $k-1$ to time k. The Kalman filter fuses measurement data from one or more sources with a dynamic model of a system to optimally estimate the system's state. The Kalman filter assumes a linear dynamic model defined by the process equation

$$x_k = Fx_{k-1} + Bu_k + w_k \tag{2.33}$$

where F is the state transition matrix applied to the previous state x_{k-1}, B is the control-input matrix applied to the control vector u_k and w_k is the process noise, which is assumed to be drawn from a zero mean Gaussian distribution $\mathcal{N}$, with covariance $Q_k : w_k \sim \mathcal{N}(0, Q_k)$.

An observation z_k of the true state x_k at time k is made according to the observation equation

$$z_k = Hx_k + v_k \tag{2.34}$$

where H is the observation matrix, which maps the true state into the observed state, and v_k is the observation noise, which is assumed to be zero mean Gaussian white noise with covariance $R_k : v_k \sim \mathcal{N}(0, R_k)$

The Kalman filter algorithm provides an estimate of the system's state x_k at time k, given the initial estimate x_0, the series of measurements $z_1, z_2, ..., z_k$, and the information of the model given by the matrices F, B, H, Q and R. It does so in a recursive process consisting of two steps - predict and update. The predict step, estimates the next state of the system and the update step corrects the estimated state with actual measurements. The predict and update steps are summarized by the following equations.

Predict:

Predicted state estimate $$\hat{x}_{k|k-1} = F_k\hat{x}_{k-1|k-1} + Bu_k \tag{2.35}$$

Predicted error covariance $$P_{k|k-1} = F_kP_{k-1|k-1}F_k^T + Q \tag{2.36}$$

Update:

Measurement residual $$\tilde{y} = z_k - H_k\hat{x}_{k|k-1} \tag{2.37}$$

Kalman gain $$K_k = \frac{P_{k|k-1}H_k^T}{H_kP_{k|k-1}H_k^T + R} \tag{2.38}$$

Updated state estimate $$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k\tilde{y} \tag{2.39}$$

Updated error covariance $$P_{k|k} = (I - K_kH_k)P_{k|k-1} \tag{2.40}$$

The state of the Kalman filter is represented by two variables - $\hat{x}_{k|k}$, the state estimate at time k, and $P_{k|k}$, the state error covariance matrix, which gives a measure of the estimated accuracy

of the state estimate $\hat{x}_{k|k}$. The predict step computes an estimate of the state $\hat{x}_{k|k-1}$ and $P_{k|k-1}$ based on the state transition matrix F as shown in Equations 2.35 and 2.36.

The update step, updates the estimates of the state variables $\hat{x}_{k|k-1}$ and $P_{k|k-1}$ based on the true measurement at time k. It first calculates the measurement residual given by Equation 2.37. The measurement residual is the difference between the true measurement z_k and the estimated measurement $H\hat{x}_{k|k-1}$. The estimated measurement is calculated by multiplying the estimated state $\hat{x}_{k|k-1}$ with the observation matrix H. As mentioned earlier the observation matrix maps the state space into the observed space. The residual $\tilde{y}$ is multiplied by the Kalman gain K_k to provide the correction to the state estimate $\hat{x}_{k|k-1}$ from the prediction step. The Kalman gain K_k represents the relative importance of the measurement residual $\tilde{y}$ with respect to the estimate $\hat{x}_{k|k-1}$. From Equations 2.38 and 2.39 it can be seen that the relative magnitudes of matrices $P_{k|k-1}$ and R control the relation between the filter's use of estimated state $\hat{x}_{k|k-1}$ and measurement residual $\tilde{y}$. When magnitude of the observation noise covariance matrix R is small, it implies that the measurements are accurate, and the updated state estimate $\hat{x}_{k|k}$ depends mostly on the measurement. When the state is estimated accurately by the predict step, the state error covariance matrix $P_{k|k-1}$ will be small compared to R and the Kalman gain will be close to zero. In this case, the filter will ignore the measurements and rely on the state estimates from the predict step. Finally, the update step calculates the updated error covariance $P_{k|k}$, which will be used in the subsequent predict step.

The Kalman filter assumes the system to be linear, but most systems in reality are non-linear and cannot be represented by a simple state transition matrix or observation martix as in Equations 2.33 and 2.34. Extended Kalman filter (EKF) is a non-linear version of the Kalman filter which linearizes the model about a current estimate via a Taylor series expansion. In EKF the state transition and observation models can be represented by differentiable functions.

$$x_k = f(x_{k-1}, u_k) + w_k \tag{2.41}$$

$$z_k = h(x_k) + v_k \tag{2.42}$$

The function f computes the estimate of the current state based on the previous state x_{k-1} and the control input u_k. The function h computes the predicted measurement from the current state estimate x_k. As f and h are non-linear functions, they cannot be multiplied with the covariance matrices directly as in Equations 2.36 and 2.38. Instead, the Jacobian - a matrix of first order partial derivatives - is computed to represent the state transition matrix F_k and the observation

matrix H_k.

$$F_k = \left.\frac{\partial f}{\partial x}\right|_{\hat{x}_{k-1,k-1},u_k} \tag{2.43}$$

$$H_k = \left.\frac{\partial h}{\partial x}\right|_{\hat{x}_{k,k-1}} \tag{2.44}$$

The EKF algorithm is similar to Kalman filter algorithm with the predict and update steps applied recursively. Equations 2.35 and 2.37 are replaced by the following two equations.

$$\text{Predicted state estimate} \qquad \hat{x}_{k|k-1} = f(\hat{x}_{k-1|k-1}, u_k) \tag{2.45}$$

$$\text{Measurement residual} \qquad \tilde{y} = z_k - h(\hat{x}_{k|k-1}) \tag{2.46}$$

In Chapter 7 we design a system using EKF to combine image and sensor-based signals to track movements of infants.

2.5 Imaging Techniques

The processes and structures within the human body that are largely hidden from the eye can be visualized through medical images. Anatomical visualization through structural and functional imaging has become a vital tool for early detection, diagnosis and treatment of cancer and other diseases [177]. Structural imaging techniques such as radiography, computed tomography and magnetic resonance imaging (MRI) are specialized for visualization of the anatomical structures enabling geometrical quantification, such as shape and size, of a given structure. On the other hand, functional imaging techniques such as positron emission tomography and dynamic contrast enhanced MRI (DCE-MRI) provide visualization of changes in blood flow or metabolism within the structures. Recently, high resolution microscopy techniques have offered views into the deepest structure of organs, providing scientists a means of understanding biological mechanisms down to sub-cellular level. In this dissertation, over various chapters, we work with images captured by MRI, diffusion-based MRI, DCE-MRI, electron microscopy (EM) and videos captured by a video camera.

2.5.1 Magnetic Resonance Imaging (MRI)

MRI uses a magnetic field and radio waves to produce detailed images of internal body structures [182]. The MRI technique makes use of the fact that body tissue contains lots of water, and hence protons (1H nuclei). To obtain an MRI, the person is made to lie in a scanner and a strong magnetic field is applied around the area to be imaged. The average magnetic moment of the

protons in the water molecules of the body get aligned in the direction of the applied magnetic field. A radio frequency current is briefly turned on, producing a varying electromagnetic field. The protons absorb this field and flip their spin. When the electromagnetic field is turned off, the spins of the protons return to equilibrium and the bulk magnetization becomes re-aligned with the static magnetic field. A radio frequency signal emitted during this relaxation is measured using receiver coils and processed to deduce positional information and form the final image. The rate at which the protons return to equilibrium is different for different tissues and determines the contrast in MRI images. Capturing signals at proton level results in high soft tissue contrast in MRI and can help depict anatomy in great detail. Based on the process used to characterize the return of protons to equilibrium, two different MRI images are obtained. T1-weighted image is produced by allowing magnetization to recover before measuring the signal by changing the repetition time (time between successive pulse sequences). T2-weighted image is produced by allowing magnetization to decay before measuring the signal by changing the echo time (time between delivery of pulse and receipt of echo signal). T1-weighted images highlight the fat tissue within the body, whereas T2-weighted images highlight fat tissue and water within the body. Therefore, T1-weighted images are helpful in obtaining structural information, while T2-weighted images are helpful in detecting edema, tumors and assessing anatomy of prostate and uterus.

2.5.2 Diffusion-weighted MRI

Diffusion-weighted MRI is a form of MRI imaging that uses diffusion of water molecules to generate contrast in images [182]. Various fibers and membranes in tissues alter the water diffusion pattern and therefore its visualization can reveal microscopic details of tissue architecture. To disentangle the effect of T1 or T2 weighting on the image contrast and obtain a quantitative measure of diffusion of water within a tissue, an apparent diffusion coefficient (ADC) image is calculated from diffusion-weighted MRIs. The ADC values are calculated by acquiring two diffusion-weighted images with different b values (degrees of diffusion sensitization). The change in signal between the two images is proportional to the rate of diffusion. ADC images are useful in detecting lesions that restrict diffusion (strokes, abscesses) and also in characterizing tumors.

2.5.3 Dynamic Contrast Enhanced MRI

DCE-MRI is a technique for in-vivo mapping of vascular function by acquiring sequential images during the passage of contrast agent within a tissue of interest [192]. The contrast agent uptake curves obtained from the acquired images are analyzed using compartmental modeling approaches

to infer the properties of tissue vasculature. It allows one to study microcirculation in tumors and helps evaluate tumor response to therapy *in vivo*.

2.5.4 Electron Microscopy (EM)

While images of entire tissues provide a holistic view, visualization at cellular and sub-cellular levels help in understanding the underlying biological changes causing normal cells to become diseased. Electron microscopy provides visualization of entire cellular ultrastructure at nanometer-resolution [144]. Electron microscopes use a beam of electrons as a source of illumination instead of a beam of light. With up to 100,000 times shorter wavelength than light photons, electrons have higher resolving power and can reveal structure of smaller objects. There are two main forms of EM – transmission EM (TEM) and scanning EM (SEM). TEM is analogous to the conventional microscope and generates a projection image by passing electrons through a thin specimen. TEM is used in various biological applications including visualization of interiors of cells, protein molecules in cell membranes and organization of molecules in viruses. SEM forms an image by scanning a focused electron beam onto the surface in a raster pattern and collecting secondary electrons emitted from the point of interaction between the electron beam and surface. The relative number of electrons detected translates to the brightness in the SEM image. SEM provides detailed images of the surfaces of cells and whole organisms, while TEM provides details of internal composition and morphology.

Advances in focused ion beam-scanning electron microscopy (FIB-SEM) enabled generation of detailed 3D images providing deeper insights into cellular organization and interactions. FIB-SEM imaging proceeds via serial steps of SEM imaging of a sample surface and FIB removal of a uniform thin layer of the tissue with size comparable to the spatial resolution in x-y plane, thereby revealing a new surface to be imaged. This process is fully automated and ensures that imaged data is equidimensional in all three axes, which significantly improves the accuracy of feature recognition within the dataset.

2.5.5 Other Imaging Techniques

In certain cases, regular camera photos and videos are used to arrive at medical diagnosis. Photographs of skin rashes can help identify skin conditions such as eczema, psoriasis and dermatitis. Video-based motion analysis is used by physical therapists to assess movement. Gait analysis – a comprehensive evaluation of the way an individual stands and walks - helps identify medical conditions such as Parkinson's disease, osteoarthritis, cerebral palsy and muscle dystrophy. It is

performed by analyzing a combination of data provided by video cameras and sensors.

In Chapter 3 we register lung radiography images, in Chapter 4 we segment tumors from MRI and ADC images, in Chapter 5 we segment and characterize cells and intra-cellular structures from FIB-SEM images, in Chapter 6 we extract features from DCE-MRI images to predict breast cancer therapy response and in Chapter 7 we track the movement of infant limbs by combining measurements from video and sensors to predict cerebral palsy.

Chapter 3

Registration: Point Set Registration Incorporating Local Shape Information

Registration is one of the first steps in the image processing pipeline. As seen in Chapter 2 point set registration is an important task in many computer vision problems. The purpose of point set registration is to align two sets of points and obtain the transformation that maps one to the other. Most of the current registration algorithms only try to align the point sets at a global level, paying little attention to the local shape of the point sets. Shape, represented as local proximity among points, is stable even in the presence of slight distortions, and can be helpful in differentiating between different objects and in recovering point correspondences. In this chapter, we propose a probability density estimation framework to align the point sets by incorporating the local structural relationships among neighboring points. The experimental results demonstrate that our approach outperforms other point set registration algorithms, even when the point sets are degraded by various levels of noise, outliers, deformation, occlusion, and rotation.*

3.1 Introduction

In Section 2.1 we discussed that most often the alignment of two point sets is considered as a density estimation problem and thin-plate splines (TPS) are used to model the transformation between them. There they made the assumption that the points in one set are normally distributed around the points in the other set. This process preserves the global structure of the point set.

However, even in non-rigid deformations the local structure is preserved due to some physical constraints. For example, though the human face is non-rigid, the nose, eyes and mouth cannot deform independently due to the constrain on their movements owing to the bone structure [201]. Therefore, the rough shape of the point set is preserved even in non-rigid deformations, else it

*The work in this chapter was presented as part of my qualifying exam in 2017.

would be difficult even for humans to match different objects. As a major contribution of this chapter, we introduce a new measure to define the local structure of the point set and propose an algorithm to integrate this measure into the density estimation framework.

We formulate the point set registration problem in a probabilistic framework. Points in one set are represented as centroids of a Gaussian mixture model (GMM) and points in the other set are represented as data samples drawn from the Gaussians. We define a new local structural measure that represents distances between points as probability distributions. We use the proposed local structural measure to arrive at the membership probabilities of the mixture densities. This results in different membership probabilities of the points based on their relative distances to a given point unlike the approach using the Hungarian method [21] where fixed probability values are assigned. Then the expectation maximization algorithm is used to fit the Gaussian centroids to the data samples. Incorporation of local structural information in the probabilistic framework will help preserve both the global and local structure of the point set.

The rest of the chapter is organized as follows. In Section 3.2, the proposed algorithm to preserve both the local and global structure is introduced. In Section 3.3, the performance of this algorithm is compared with other approaches. Finally in Section 3.4 we summarize and present possible future directions.

3.2 Proposed Algorithm

We formulate the point set registration as a density estimation problem using Gaussian Mixture Model (GMM) similar to Myronenko and Song [128]. We then propose a new local structural measure to calculate the membership probabilities of the mixture densities and solve the point set registration problem using expectation maximization algorithm. The coordinates of the reference point set are denoted as $\mathbf{X}_{N\times D} = [x_1, x_2, \ldots, x_N]^{\mathrm{T}}$ and the coordinates of the template point set are denoted as $\mathbf{Y}_{N\times D} = [y_1, y_2, \ldots, y_N]^{\mathrm{T}}$. The aim of the algorithm is to obtain the transformation $\mathbf{T}$ that best aligns the template point set $\mathbf{Y}$ with reference point set $\mathbf{X}$. We formulate the point set registration as a density estimation problem, where we fit GMM centroids to the data points, by maximizing the likelihood function. The points in $\mathbf{Y}$ are considered as centroids of Gaussians and the points in $\mathbf{X}$ are considered as data points. The non-rigid transformation $\mathbf{T}$ which needs to be estimated, is defined as a continuous velocity function v, which is added to the initial centroid positions $\mathbf{Y}$, such that the current position of the centroids can be given as, $\mathbf{T}(\mathbf{Y}, v) = \mathbf{Y} + v(\mathbf{Y})$.

Consider the GMM probability density function given as

$$p(x) = \sum_{m=1}^{M+1} P(m)p(x|m) \tag{3.1}$$

where, the first factor represents the membership probability and the second term defines the Gaussian

$$p(x|m) = \frac{1}{(2\pi\sigma^2)^{D/2}} e^{-\frac{\|x-T(y_m,v)\|^2}{2\sigma^2}} \tag{3.2}$$

with equal isotropic covariances σ^2. Here y represents the centroids of the GMM and x represent the data point. We add an additional uniform distribution $p(x|M+1) = \frac{1}{N}$ to Equation 3.2, to account for noise and outliers. The weight of the uniform distribution is denoted by w, $0 \leq w \leq 1$. The Gaussian mixture model with the addition of uniform distribution and the transformation function expanded as $\mathbf{T}(\mathbf{Y}, v) = \mathbf{Y} + v(\mathbf{Y})$ has the following form.

$$p(x) = w\frac{1}{N} + (1-w)\sum_{m=1}^{M} \frac{P(m)}{(2\pi\sigma^2)^{D/2}} e^{-\frac{\|x-y_m-v(y_m))\|^2}{2\sigma^2}} \tag{3.3}$$

Assuming all Gaussian components are independently distributed, the joint GMM probability density function, also called the likelihood function, that needs to be maximized can be written as the following.

$$p(X) = \prod_{n=1}^{N} p(x_n) = \prod_{n=1}^{N} \sum_{m=1}^{M+1} P(m)p(x_n|m) \tag{3.4}$$

As maximizing a likelihood function is equivalent to minimizing its negative log-likelihood, taking the negative log of Equation 3.4 gives the energy function that needs to be minimized.

$$E = -\sum_{n=1}^{N} log \sum_{m=1}^{M+1} P(m)p(x_n|m) \tag{3.5}$$

3.2.1 Incorporation of Local Similarity Measure

As mentioned in Section 2.1, most of the algorithms assume equal membership probabilities $P(m)$ for all GMM components irrespective of how similar or dissimilar the neighborhoods of the points in one point set are to the other. They do not look into the local structural similarities among the point sets. However, lot of information can be gained by incorporating this knowledge into the registration framework. Therefore, instead of using equal membership probabilities, we propose a measure based on local structural information to estimate the membership probabilities. We assume that the spatial proximities between points in a local neighborhood of a point set remain the same, even in the presence of moderate deformation and noise.

Figure 3.1 shows two fishes where the red fish is a deformed version of the blue fish. It can be seen from Figures 3.1a and 3.1b that the distribution of distances between the eye and some

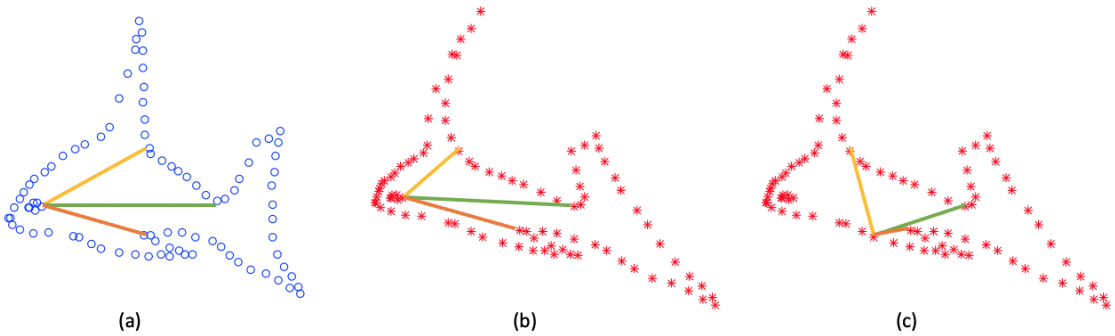


Figure 3.1: The red fish is a deformed version of the blue fish. Comparing (a) and (b) it can be seen that the relative distances between points are maintained even in the presence of non-rigid deformation. Comparing (a) and (c) it can be seen that the distribution of distances is different around different points.

random points in both the fish is similar. In contrast, from Figure 3.1c it can be seen that the distribution of distances from another point to the same three random points is different. Therefore, the distribution of distances around a point can give a rough idea of the overall shape of the object. Here we use only the distance and not the angle to keep the measure simple.

We introduce a distance metric D_{ij}^X which measures the distance between points x_i and x_j in a given point set $\mathbf{X}$, and is defined as

$$D_{ij}^X = \frac{e^{-\frac{\|x_i - x_j\|^2}{2c^2}}}{\sum_k e^{-\frac{\|x_i - x_k\|^2}{2c^2}}} \tag{3.6}$$

such that $\sum_j D_{ij} = 1$ and c^2 is the variance of the Gaussian and is set to 0.01. In the matrix D^X, d_{ij} represents the probability that x_i will pick x_j as a neighbor. This probability is proportional to the probability density of x_j under a Gaussian centered at x_i. Therefore, this probability is high for nearby points and lower for points further away. Each row in D^X represents a distribution of distances between one point and the remaining $N-1$ points. This gives an approximate representation of the shape of the point set relative to that point. Similarly, D^Y represents the distribution of distances between points in set Y.

The more similar the local structure around two points x_n and y_m, the more likely is the correspondence between these two points. This similarity between local structures is measured by calculating the Kullback-Leibler (KL) distance between the distance metrics corresponding to points x_n and y_m.

$$KL(D_n^X || D_m^Y) = \sum_j D_{nj}^X log \frac{D_{nj}^X}{D_{mj}^Y} \tag{3.7}$$

where, D_n^X is the row of D^X representing distances from point x_n to all other points in point set X, and D_m^Y is the row of D^Y representing distances from point y_m to all other points in point set Y. In order to use the KL distances as membership probabilities in GMM, they need to be converted into a probability measure. To achieve this, we used Gibb's measure, which defines the probability of points x_n and y_m matching to be proportional to the exponential of the KL_{nm} distance. After normalizing over all the points, the similarity measure is given as

$$S_{nm} = \frac{e^{-\beta \cdot KL_{nm}}}{\sum_{m=1}^{M} e^{-\beta \cdot KL_{nm}}} \tag{3.8}$$

where β controls the spatial width of the distribution. The term S_{nm} represents the similarity between the local structure of a point x_n to that of a point y_m. This is a point-wise similarity measure that is sensitive to its local neighborhood. For points with similar neighborhoods in the two datasets this measure is high, and it decreases exponentially as the local structure between the two points changes. This eliminates the need to manually set a matching value as done while using shape context [116]. This measure is inherently invariant to rotation and translation as it is calculated with the help of relative point locations. It can be made scale invariant by normalizing the distances by the mean distance between all point pairs. It can be used for dimensions greater than two, since the measure depends only on the distance between the points. The proposed local structural similarity measure is used to represent the membership probability $P(m)$ in Equation 3.5.

$$\begin{aligned} E(v, \sigma^2, w) &= -\sum_{n=1}^{N} log \sum_{m=1}^{M+1} S_{nm} p(x_n|m) \\ &= -\sum_{n=1}^{N} log \left[w\frac{1}{N} + (1-w) \sum_{m=1}^{M} \frac{S_{nm}}{(2\pi\sigma^2)^{D/2}} e^{-\frac{\|x - y_m - v(y_m))\|^2}{2\sigma^2}} \right] \end{aligned} \tag{3.9}$$

There are three unknown parameters in this formulation: v, the velocity field; σ^2, the variance of the Gaussians and w, weight of the outliers, that need to be estimated.

3.2.2 Regularization by Incorporating Motion Coherence

The velocity field v is non-rigid and unknown, and therefore can have a broad class of solutions, making its determination an ill-posed problem. In order to choose one particular solution, a regularization term containing knowledge regarding the desired properties of the velocity fields need to be added. Here we want the velocity field v created by the displacement of the template point set to be smooth, which captures the idea that points close to one other tend to move coherently. Smoothness is a measure of the oscillatory behavior of a function. A function is said to be smooth if it exhibits few oscillations; in frequency domain it translates to less energy in the high

frequency range. The high frequency content of a function can be estimated by high-pass filtering the function, and then measuring the resultant power, which is the L2 norm, of the result. This power is used as the regularization function and is expressed as, $\phi(v) = \|v\|_H^2 = \int_{R^d} \frac{|\tilde{v}(s)|^2}{\tilde{G}(s)} ds$, where $\tilde{v}$ indicates the Fourier transform of the velocity and $\tilde{G}$ is the Fourier transform of a function G such that it approaches zero as $\|s\| \to \infty$, so that $\frac{1}{\tilde{G}}$ represents a high pass filter. Here G is chosen to be a Gaussian low-pass filter. The Fourier transform of a Gaussian is another Gaussian, where $\tilde{G}(s)$ decreases exponentially fast with s. So if v has to have a finite value, its Fourier transform must drop even faster than that of G. This effectively results in v having only a few low frequency components with high weights. A function with only low frequency components does not oscillate much and is smooth. The size of the Gaussian filter controls the range of frequencies filtered and thus the amount of spatial smoothness. Thus, the regularization function $\phi(v)$, ensures that the velocity field is smooth. This process is analogous to taking the norm of a function v in the Reproducing Kernel Hilbert Space (RKHS) with a Gaussian as the reproducing kernel [20]. If λ is the weight given to the regularization term, Equation 3.9 is extended to give the following.

$$E(v, \sigma^2, w) = -\sum_{n=1}^{N} log \sum_{m=1}^{M+1} S_{nm} p(x_n|m) + \frac{\lambda}{2}\phi(v) \tag{3.10}$$

3.2.3 Optimization using the Expectation Maximization Algorithm

The three unknown parameters v, σ^2 and w are estimated by using the expectation maximization algorithm. In the expectation step, the posterior probability of the mixture components is evaluated. In the maximization step, the new values of the parameters are estimated by minimizing the expectation of the negative log likelihood function.

E-step: We first evaluate the membership probabilities between the two point sets using Equations 3.6 - 3.8. We then compute the posterior probability $P^{old}(m|x_n)$ using the old values of v, σ^2 and w and applying Bayes rule.

$$P^{\text{old}}(m|x_n) = \frac{S_{mn} e^{-\frac{\|x_n - y_m - v(y_m)\|^2}{2\sigma^2}}}{\sum_{k=1}^{M} S_{mn} e^{-\frac{\|x_n - y_m - v(y_m)\|^2}{2\sigma^2}} + \frac{w}{(1-w)N}(2\pi\sigma^2)^{D/2}}. \tag{3.11}$$

This posterior probability indicates to what degree the centroid y_m coincides with the data point x_n under the estimated transformation $\mathbf{T}(\mathbf{Y}, v) = \mathbf{Y} + v(\mathbf{Y})$. To obtain the objective function the negative log likelihood function is multiplied and divided with $P^{old}(m|x_n)$ inside the logarithm.

$$L = -\sum_{n=1}^{N} log \sum_{m=1}^{M+1} S_{nm} p(x_n|m) \frac{P^{old}(m|x_n)}{P^{old}(m|x_n)} \tag{3.12}$$

Jensen's inequality takes the form $\varphi \sum_i a_i b_i \leq \sum_i a_i(\varphi(b_i))$ for a convex function φ and positive weights a_i, such that $\sum_i a_i = 1$ [98]. Here the negative logarithm is a convex function and

$P^{old}(m|x_n)$ is a probability distribution. Applying Jensen's inequality to Equation 3.10 gives the following.

$$L \leq -\sum_{m,n=1}^{M+1,N} S_{nm} P^{old}(m|x_n) log \frac{S_{nm} p(x_n|m)}{P^{old}(m|x_n)} \tag{3.13}$$

$$L \leq -\sum_{m,n=1}^{M+1,N} S_{nm} P^{old}(m|x_n) log(S_{nm} p(x_n|m)) - \sum_{m,n=1}^{M+1,N} S_{nm} P^{old}(m|x_n) log\left(P^{old}(m|x_n)\right) \tag{3.14}$$

The upper bound on the negative log-likelihood function is the objective function Q that is minimized in the M-step.

$$Q = -\sum_{n=1}^{N} \sum_{m=1}^{M+1} P^{old}(m|x_n) log[S_{nm} p(x_n|m)] \tag{3.15}$$

Here the second term is ignored as $P^{old}(m|x_n)$ are old posterior probabilities and are already known and need not be estimated.

M-step: In this step the new parameter values are obtained by minimizing the objective function given by Equation 3.15. Ignoring the constants independent of v, σ^2 and w, the objective function including the regularization term $\phi(v)$ is expanded.

$$\begin{aligned} Q(v, \sigma^2, w) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} \sum_{m=1}^{M} P^{old}(m|x_n) \|x_n - y_m - v(y_m)\|^2 + \frac{N_p D}{2} \log \sigma^2 \\ - N_p \log(1-w) - (N - N_P) \log w + \frac{\lambda}{2} \|v\|_H^2 \end{aligned} \tag{3.16}$$

where, $N_p = \sum_{n=1}^{N} \sum_{m=1}^{M} P^{\text{old}}(m|x_n) \leq N$. Taking the derivative of Equation 3.16 with respect to w and σ^2, and equating it to zero, we obtain the following.

$$w = 1 - \frac{N_p}{N} \tag{3.17}$$

$$\sigma^2 = \frac{1}{N_p D} \sum_{n=1}^{N} \sum_{m=1}^{M} \|x_n - y_m - v(y_m)\|^2 \tag{3.18}$$

Next we look at the solution for the velocity field v. As mentioned previously, it is modeled to lie within a Reproducing Kernel Hilbert Space H, defined by a diagonal Gaussian matrix kernel $\boldsymbol{G}$, where $g_{ij} = e^{-\frac{\|y_i - y_j\|^2}{2\beta^2}}$, and its square norm is used as the regularization term i.e. $\phi(v) = \|v\|_H^2$. The objective function given by Equation 3.16 is a special form of the regularization functional introduced by Tikhonov to solve ill-posed problems [187, 37]. According to the Representer theorem [123], the optimal solution for velocity field v, which minimizes Equation 3.16, is unique and is given as

$$v(y) = \sum_{m=1}^{M} w_m \boldsymbol{G}(y, y_m) \tag{3.19}$$

where $\boldsymbol{G}$ is a kernel matrix with elements $G_{ij} = e^{-\frac{\|y_i - y_j\|^2}{2\beta^2}}$, and w_m are Gaussian kernel weights (the proof of the theorem is given in Appendix A). The regularization term is taken as the square norm of the velocity field and can be expressed using the inner product as

$$\begin{aligned}\phi(v) = \|v\|_H^2 = \langle v, v \rangle_H &= \left\langle \sum_{i=1}^{M} w_i \boldsymbol{G}(., y_i), \sum_{j=1}^{M} w_j \boldsymbol{G}(., y_j) \right\rangle_H \\ &= \sum_{i,j=1}^{M} w_i w_j \boldsymbol{G}(y_i, y_j) = \boldsymbol{W}^T \boldsymbol{G} \boldsymbol{W} \end{aligned} \tag{3.20}$$

where $\boldsymbol{W} = (w_1, w_2, ..., w_m)^T$ are the Gaussian kernel weights. Next we substitute the solution from Equation 3.20 into Equation 3.16.

$$\begin{aligned} Q(v, \sigma^2, w) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} \sum_{m=1}^{M} P^{old}(y_m|x_n) \|x_n - y_m - v(y_m)\|^2 + \frac{N_p D}{2} \log \sigma^2 \\ - N_p \log(1-w) - (N - N_P) \log w + \frac{\lambda}{2} (\boldsymbol{W}^T \boldsymbol{G} \boldsymbol{W}) \end{aligned} \tag{3.21}$$

We next take the the derivative of Equation 3.21 with respect to $\boldsymbol{W}$.

$$\frac{\partial Q}{\partial \boldsymbol{W}} = -\frac{1}{\sigma^2} \boldsymbol{G} \left[\sum_{n=1}^{N} \sum_{m=1}^{M} P^{old}(m|x_n)(x_n - y_m - \boldsymbol{G}\boldsymbol{W}) \right] + \lambda \boldsymbol{G} \boldsymbol{W} \tag{3.22}$$

$$= \frac{1}{\sigma^2} \boldsymbol{G} \left[d(\boldsymbol{P}\boldsymbol{1})(\boldsymbol{Y} + \boldsymbol{G}\boldsymbol{W}) - \boldsymbol{P}\boldsymbol{X} \right] + \lambda \boldsymbol{G} \boldsymbol{W} = 0 \tag{3.23}$$

Rewriting Equation 3.22 we obtain a linear system of equations

$$(\boldsymbol{G} + \lambda \sigma^2 d(\boldsymbol{P}\boldsymbol{1})^{-1}) \boldsymbol{W} = d(\boldsymbol{P}\boldsymbol{1})^{-1} \boldsymbol{P}\boldsymbol{X} - \boldsymbol{Y} \tag{3.24}$$

which are solved to obtain the weights $\boldsymbol{W} = (w_1, w_2, ..., w_m)^T$. $\boldsymbol{G}$ is the Gaussian kernel matrix, $\boldsymbol{P}$ is the posterior probability matrix $P^{old}(m|x_n)$, $\boldsymbol{1}$ is a column vector of ones and $d(\cdot)$ is the diagonal matrix. The proposed algorithm is summarized in Algorithm 1. The algorithm has two free parameters β and λ. The regularization weight λ acts as a tradeoff to determine how smooth we want the solution to be and β determines the width of the smoothing kernel.

3.3 Experimental Results

To evaluate the performance of the proposed algorithm, we performed several experiments on 2D points, 3D points and real biomedical data. The experiments were implemented in Matlab and run on a 2.4 GHz Intel Core i5 CPU with 8GB memory.

Algorithm 1 Registration Algorithm

Input: Two point sets $\mathbf{X}$ and $\mathbf{Y}$, parameters $\beta,\ \lambda$.
Output: Optimal transformation $\mathbf{T}$.

Initialize : $\beta > 0,\ \lambda > 0, 0 \geq w \geq 1,\ \mathbf{W} = 0, \sigma^2 = \frac{1}{DNM}\sum_{m,n=1}^{M,N} \|x_n - y_m\|^2$.

Construct $\mathbf{G} : g_{ij} = e^{-\frac{1}{2\beta^2}\|y_i - y_j\|^2}$
Compute feature matrix $\mathbf{A}$ for point set $\mathbf{X}$.
EM optimization, repeat until convergence.
  E-step:
    Compute feature matrix $\mathbf{B}$, for point set $\mathbf{Y}$.
    Calculate KL divergence between $\mathbf{A}$ and $\mathbf{B}$.
    Initialize S_{mn} according to Equation 3.8.
    Compute posterior probability $\mathbf{P}$ according to Equation 3.11.
  M-step:
    Solve $\left(\mathbf{G} + \lambda\sigma^2 d(\mathbf{P}\mathbf{1}^{-1})\right)\mathbf{W} = d(\mathbf{P}\mathbf{1}^{-1})\mathbf{P}\mathbf{X} - \mathbf{Y}$
    $\mathbf{T} = \mathbf{Y} + \mathbf{G}\mathbf{W}$
    $w = 1 - \frac{N_p}{N}$
    $\sigma^2 = \frac{1}{N_p D}\left(tr(\mathbf{X}^{\mathrm{T}} d(\mathbf{P}^{\mathrm{T}}\mathbf{1})\mathbf{X})\right) - 2tr((\mathbf{P}\mathbf{X})^{\mathrm{T}}\mathbf{T}) + tr\left(\mathbf{T}^{\mathrm{T}} d(\mathbf{P}\mathbf{1})\mathbf{T}\right)$
The aligned point set is $\mathbf{T} = \mathbf{Y} + \mathbf{G}\mathbf{W}$.
The probability of correspondences is given by $\mathbf{P}$.

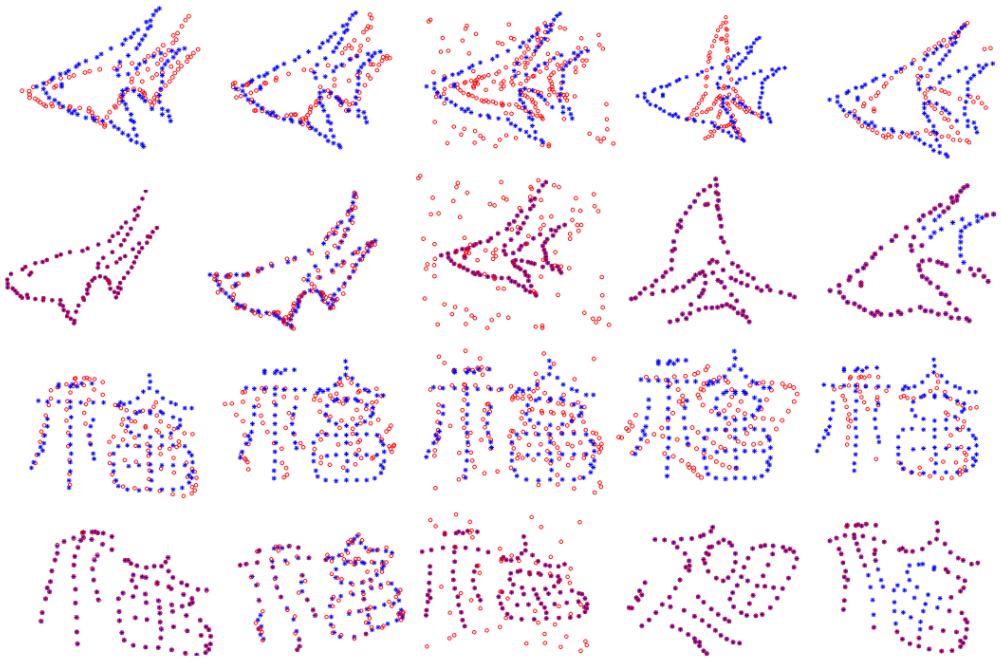


Figure 3.2: Results of the proposed algorithm on fish (top two rows) and Chinese character (bottom two rows) shapes, with deformation, noise, outliers, rotation and occlusion (left to right). The template point set (blue) is aligned with the reference point set (red). For each shape the top row represents the initialization and bottom row represents the registration result.

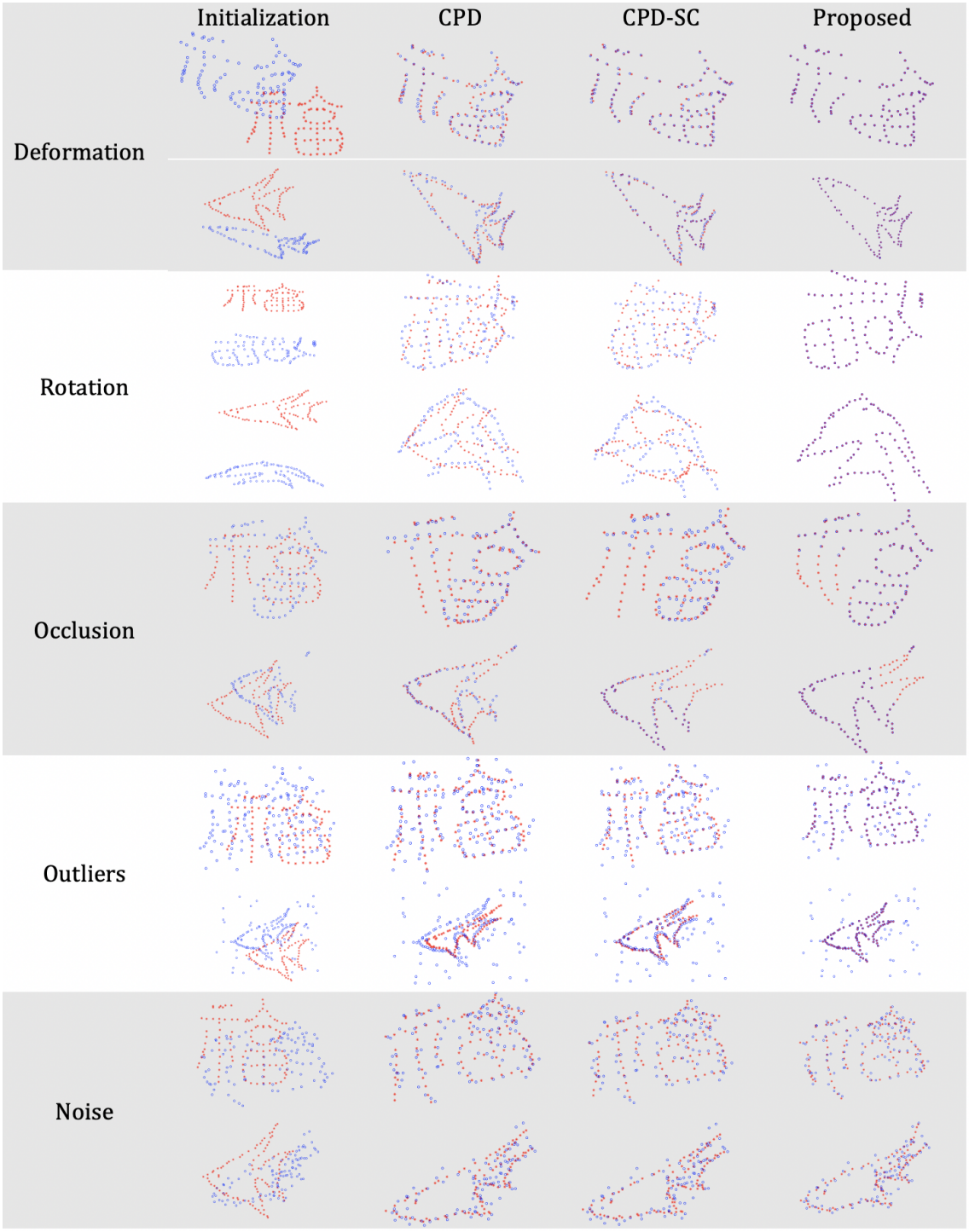


Figure 3.3: Registration results of on Chinese character and fish in the presence of (a) deformation, (b) rotation, (c) occlusion, (d) outliers and (e) noise. From left to right, point set initialization, result of CPD, CPD-SC and proposed algorithm.

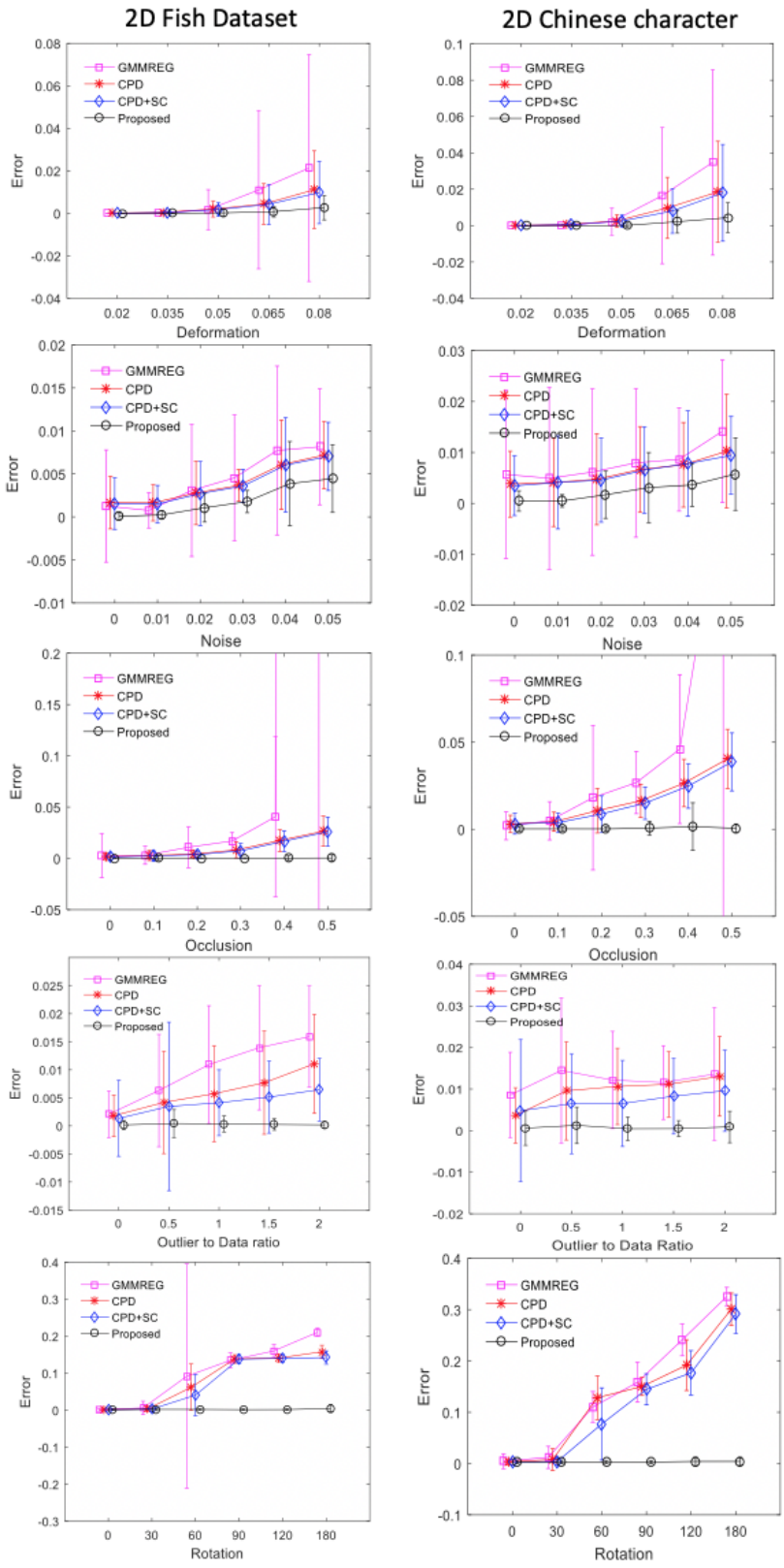


Figure 3.4: Performance of the proposed algorithm (black) in comparison with CPD (red), CPD+SC (blue) and GMMREG (magenta) on 2D fish dataset (left) and 2D Chinese character (right) in the presence of (top to bottom) (a) deformation, (b) noise, (c) occlusion, (d) outliers and (e) rotation.

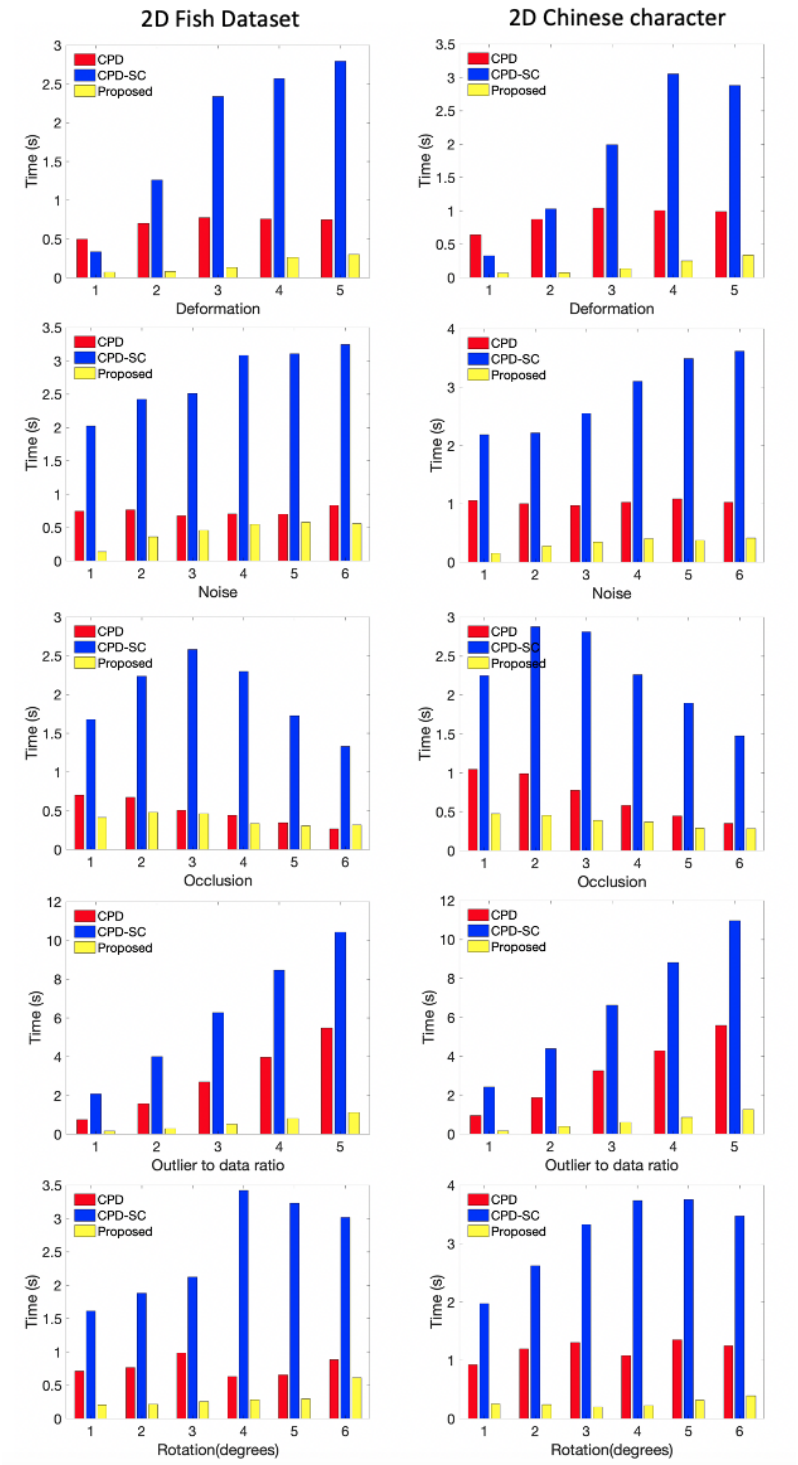


Figure 3.5: Comparison between CPD (red), CPD+SC (blue) and proposed algorithm (yellow) based on of time taken on on 2D fish dataset (left) and 2D Chinese character (right) in the presence of (top to bottom) (a) deformation, (b) noise, (c) occlusion, (d) outliers and (e) rotation

3.3.1 2D Fish and Chinese Character Dataset

To test the performance in the 2D case, we used the synthesized datasets of the fish and the Chinese character generated by Chui and Rangarajan [40]. For both the fish and Chinese character, there are five sets of data to measure the robustness of the algorithm in the presence of different levels of noise, outliers, rotations, occlusions and deformations. Each set at each level had 100 representative example created by the following modifications. The addition of zero-mean white noise with standard deviation ranging from 0 to 0.05 to the point set produced the five levels of noise. To produce the outlier dataset, points generated randomly from a normal zero-mean distribution were added, with outlier to data ratio ranging from 0 to 2. The data points were rotated by 0, 30, 60, 90, 120, and 180 degrees to generate the rotation dataset. Deletion of data points, with the ratio of deleted points to original points ranging from 0 to 0.05 generated the occlusion dataset. The deformations were generated by parameterizing the data points by a mesh of control points, perturbing these control points and using splines to interpolate the deformation. The mesh points were perturbed by using Gaussians of mean zero and standard deviations varying from 0.02 to 0.08. The results obtained after registration for different settings for both fish and Chinese character are shown in Figure 3.2

To test the performance for purely rigid deformations, we took a fish point set and applied various levels of rotations, scaling and translations. The proposed algorithm converged in two iterations while CPD took around twenty iterations, as the proposed local structural measure is rotation, translation and scale invariant. Further CPD does not perform well for rotations greater than 60 degrees, while the proposed similarity measure can handle any angle of rotation.

3.3.2 Results on 2D Data

Analysis of Registration Error

To evaluate the performance of the proposed algorithm, we compared it with GMMREG [88], CPD [128] and CPD-SC [116] algorithms described in 2.1. Here we are referring to the work by Ma and colleagues as CPD-SC. The GMMREG and CPD algorithms had the source code available, but for CPD-SC we re-implemented it based on the authors' description of their algorithm as the source code was not available. The results obtained after registration for different settings in the non-rigid case for both the fish and Chinese character using CPD, CPD-SC and the proposed algorithm are shown in Figure 3.3. In the experiments, we set the value of $\lambda = 1$, and $\beta = 2$. The registration performance of GMMREG, CPD, CPD-SC and the proposed algorithm for different scenarios on 2D fish and Chinese character are compared in Figure 3.4. The mean squared distance between

corresponding points after the registration is taken as the error measure. The error-bars indicate the results for each test averaged over 100 runs. From these figures it can be seen that the proposed algorithm outperforms the other three algorithms in all scenarios. The first row of Figure 3.4 shows that addition of local structural information results in lower error compared to other algorithms in the presence of deformation. The advantage of including our structural similarity measure is more significant as the level of deformation increases. In addition to the mean value of error being decreased, the standard deviation of error is also reduced using the proposed algorithm.

A similar trend can be observed in the experiments involving various levels of occlusion, rotation, noise, and outliers. In CPD-SC, the shape context similarity measure is calculated only once for every ten iterations as it is a computationally expensive step. Therefore, in CPD-SC if a wrong correspondence is given high probability at the beginning it will be difficult to change in further iterations. Further in CPD-SC the matched points are assigned fixed high probability and all other points are assigned a fixed low probability, which are set by the user. In our proposed algorithm the membership probabilities vary depending on the local structural similarity. Matching points have the maximum probability and the probability keeps decreasing as the local structure changes. In this setting, the probability of non-matching far off points is assigned a much smaller value compared to the one used in CPD-SC as in the proposed measure the value decreases exponentially according to Equation 3.8. This leads to a faster decrease in the variance of the Gaussians when compared to CPD-SC as seen in Figure 3.5 (c).

Analysis of Computational Efficiency

We also compared the computational efficiency of these algorithms. The time taken to converge by CPD, CDP-SC and the proposed algorithm for 2D fish and Chinese character dataset under various settings are shown in Figure 3.5. Even though the shape context similarity measure is calculated only once for every ten iterations, CPD-SC algorithm takes longer time than both CPD and the proposed algorithm. This increase in time is due to the Hungarian algorithm used to calculate the correspondences in CPD-SC. The proposed algorithm converges in far less number of iterations when compared to the other two algorithms. This is because the proposed similarity measure forces similar points to align, leading to faster convergence. Though calculating the similarity measure in each iteration takes extra time, as the number of iterations to convergence is much smaller than CPD, the overall runtime of the proposed algorithm is smaller.

To examine how the proposed measure helps improve the performance, we considered an example from the fish dataset with the smallest level of deformation. It can be seen from Figure 3.6 that CPD, CPD-SC and the proposed algorithm all lead to good correspondences, but the proposed

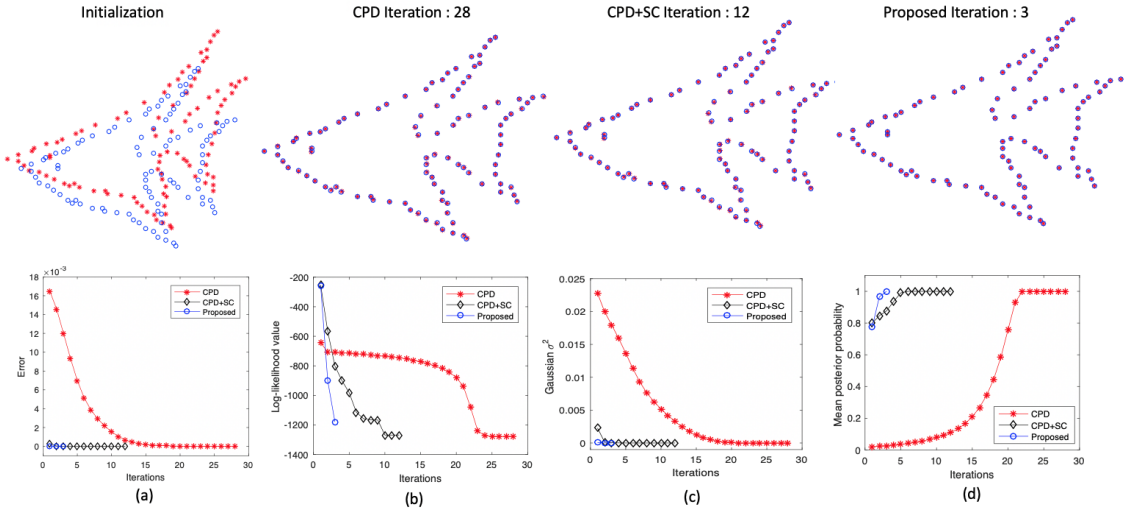
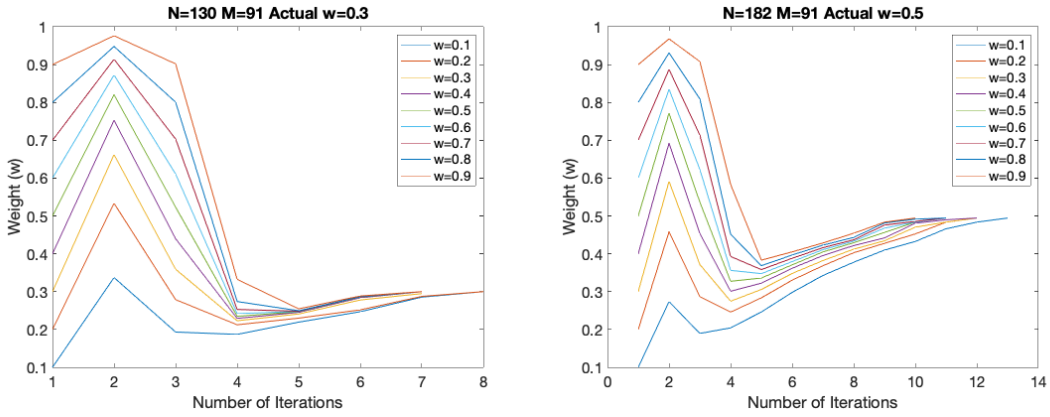


Figure 3.6: Top row shows the initialization and registration results for CPD, CPD-SC and the proposed algorithm. The graphs compare the variation of (a) Mean square error, (b) Log-likelihood value, (c) Variance of the Gaussians in the GMM, and (d) Mean posterior probability along different iterations.

algorithm converges in 3 iterations, whereas CPD+SC takes 12 and CPD 28 iterations to reach convergence. Figure 3.6(a) shows the variation of mean square error with iterations. As CPD-SC and the proposed algorithm use local structural measure to define the membership probabilities, the initial error is much lower when compared to CPD. As the local membership probability narrows the range of corresponding points the width of the Gaussians is also reduced as compared to CPD as seen in Figure 3.6(c).

The variation of the log-likelihood value, which is minimized to obtain registration result, is shown in Figure 3.6(b). It can be seen that addition of local structural measure helps in steeper reduction of the function, and the proposed method has much steeper reduction as the membership probability is varied according to the structural similarity, instead of keeping it constant as in CPD-SC. As the true correspondences are known, we calculated the mean posterior probability for the corresponding points at each iteration. From figure 3.6(d) it can be seen that the proposed method finds the correct matches (probability one) much faster than the other two algorithms.

Unlike CPD where the weight of the outliers w, was set by the user, we have solved for w as an unknown parameter in the EM framework. The variation of the weight of the outliers w, is examined over different iterations. From Figure 3.7 it can be seen that irrespective of the initialization, the weight eventually converged to the correct percentage of outliers.



(a) (b)

Figure 3.7: Variation of weight (w) along iterations for different initializations

3.3.3 Non-Rigid 3D Race Registration

To test the performance of the algorithm on 3D data, we took the 3D face point set and introduced various levels of noise and rotations. The initialization and final alignment at different settings are shown in Figure 3.8. The proposed algorithm performs better than CPD in the presence of various levels of both noise and rotation as seen from Figure 3.9. Our method of calculating probabilistic distances is computationally less complex than calculating shape context for 3D case.

3.3.4 Non-Rigid 2D Tools Registration

We evaluated the performance of the proposed algorithm on 2D tools dataset [29]. This dataset consists of articulated shapes for shape similarity experiments. We selected boundary points from the silhouettes of the different shapes and performed point set registration using different algorithms. It can be observed from Figure 3.10 that the addition of local structural information led to better performance. The first row shows the two shapes to be registered and the remaining rows show the results of registration using different algorithms. The proposed algorithm results in better correspondences between points than the other three. The negative log-likelihood value which is the objective function we are minimizing was much lower for the proposed algorithm when compared to CPD or CPD+SC suggesting better correspondence between the points.

3.3.5 Non-Rigid 2D Lung Point Set Registration

We also evaluated our algorithm on real data from chest radiography images. We used chest radiographs from the Japanese Society of Thoracic Radiology database [152], which had manual

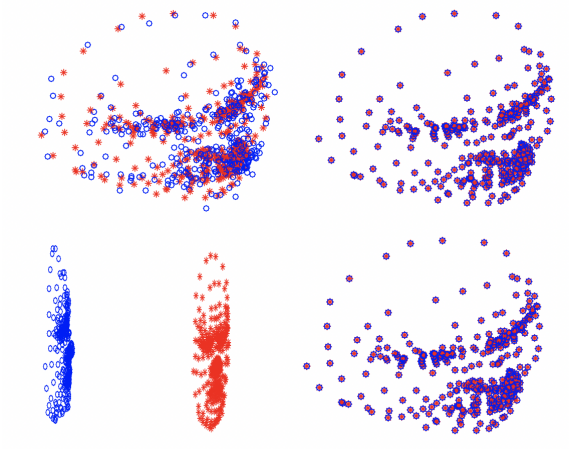
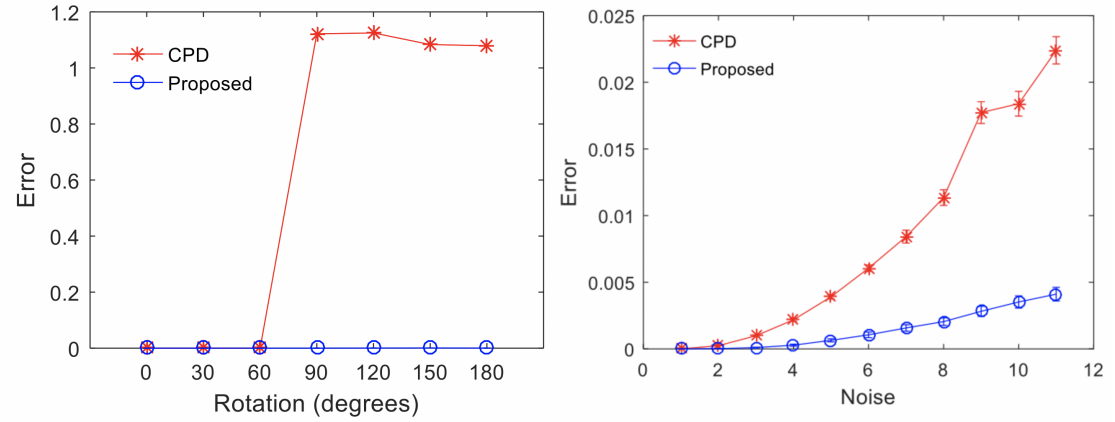


Figure 3.8: Initialization (left) and final alignment (right) of 3D face point sets in the presence of noise (top) and rotation (bottom)



(a) (b)

Figure 3.9: Comparison of performance between CPD and proposed algorithm in the presence of rotation and noise in 3D face dataset.

segmentations of the lungs, heart and clavicles for each image [179]. The registration results show that inclusion of local structural information during registration helps in better alignment of point sets. This can be observed in the alignment of the right lung in Figure 3.11.

3.4 Conclusion

In this chapter, we proposed a new algorithm for point set registration. The key feature of this algorithm is that it helps preserve the local structure as well as the global structure of the point set. We introduced a new measure to capture the local structure of the point set. This measure is used

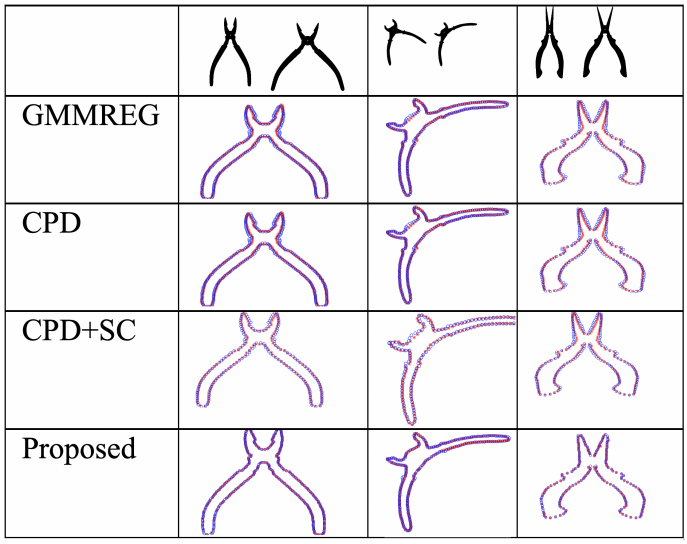
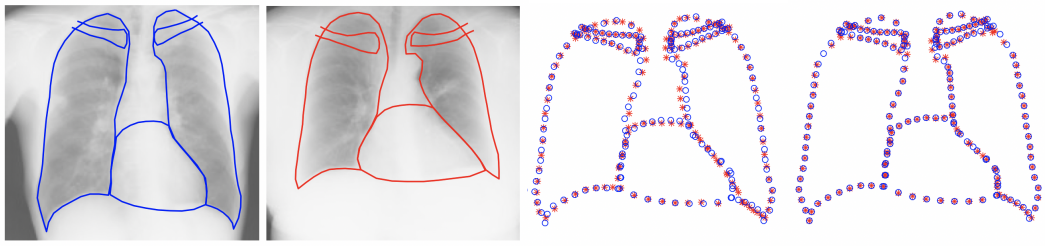


Figure 3.10: Comparison of performance between GMMREG, CPD, CPD+SC and proposed algorithm on articulated tools dataset



(a) (b) (c)

Figure 3.11: Comparison of performance between CPD and proposed algorithm on chest radiography images (a) Initializations (b) Registration with CPD (c) Registration with proposed algorithm

to represent the membership probabilities of GMM, which in turn helps in faster convergence of the algorithm. We derived the weight of the outlier distribution by considering it to be an unknown in the expectation maximization framework. We tested this algorithm on 2D, 3D and real datasets. The proposed algorithm yields better results even in the presence of noise, deformation, outliers, occlusion and rotation than the existing algorithms.

Having observed that the proposed local structural measure reduces the computational time and error in pairwise point set registration, the next step is to apply this measure in group-wise point set registration setting to see if it results in comparable outcome. The present group-wise registration algorithms take a few hours to obtain results on small datasets [28]. Reducing computational time in this setting would be a prominent contribution. Further in group-wise registration of sequence

of images (e.g. Heart MRI sequence), there is temporal information that can be modeled into the registration algorithm, to give better registration results.

Noise and outliers are presently accounted by a uniform distribution. This can probably be improved by modeling noise and outliers using better fitting models. In addition to improving registration performance, a local structural measure can also help in modeling noise and outliers. When we see an image containing an object along with outliers and noise, we may be able to roughly estimate which of the points belong to the object and which among them are outliers. We can identify two different distributions of points. If we come up with a measure to model these two distributions mathematically, then the error in the registration process can be further reduced. The regularization in the proposed algorithm only makes sure that the motion is coherent. A regularization method conserving local structure along with motion coherence can help achieve better registration preserving the overall structure. Further, the proposed algorithm is still slow when applied to large datasets. There is a need to come up with a new framework that can speed up the registration process to greater extent.

Chapter 4

Segmentation: Modifications to Enhance the Performance of U-Net

4.1 Introduction

Once the images are pre-processed using registration, the next step in most image processing pipelines is segmenting the region of interest. In Section 2.2 we discussed that U-Net is the most commonly used segmentation architecture. It consists of an encoding path that extracts features, and a decoding path that up-samples the extracted features to obtain the final segmentation. Though U-Net performs well in segmenting most objects, modifying its design to better suit the problem in hand could further improve its performance. In this chapter, we experiment with several modifications to the U-Net architecture to improve its performance in segmenting tumors in MRI images of human prostates. As tumor occupies a very small region in the entire image, using a method designed to deal with such class imbalance could improve the segmentation performance. We experiment with Tversky loss and Focal loss, which are specially designed for datasets with imbalanced classes. We explore the effect of pruning the extracted features with attention gates so as to only retain features relevant for segmentation. We also explore the effect of adding auxillary functions, such as an auto-encoder branch and companion objective functions at each level of the U-Net, to provide additional guidance in extracting representative features.*

*The work in this chapter related to attention gates has been published in A. Machireddy, N. Meermeier, F. Coakley, and X. Song, 2020, Malignancy detection in prostate multi-parametric MR images using U-Net with attention, in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 1520-1523).

4.2 Background and Related Work

4.2.1 Prostate Cancer Diagnosis

Prostate cancer is the most common cancer diagnosed among men in the United States and is the second leading cause of death due to cancer [154]. Prostate cancer detected when still localized provides the highest chance of being cured [31]. Current screening tests include a digital rectal exam and a blood test to detect the level of prostate specific antigen (PSA) - a protein made by the cells in the prostate [126]. An elevated level of PSA is indicative of prostate cancer or other prostate related problems. A transrectal ultrasound guided biopsy is performed to confirm a diagnosis of prostate cancer when the rectal exam or the blood test results are abnormal. The transrectal ultrasound guided biopsy randomly extracts 12 specimens from several areas of the prostate [35]. Due to random sampling, it misses clinically significant tumors in 30-40% of the cases [83] and sometimes underestimates the aggressiveness and extent of the cancer [99].

An elevated PSA but negative transrectal ultrasound guided biopsy could indicate that the tumor might have been missed and such patients are referred to undergo multi-parametric MRI imaging. As mentioned in Section 2.5, MRI scans can be programmed for several different pulse sequences, or parameters, to obtain images that highlight specific differences between healthy and unhealthy tissue. When two or more parameters are used, it is called multi-parametric MRI imaging. For localizing prostate tumors, multi-parametric MRI includes T2-weighted images, diffusion weighted images and DCE-MRI images. The T2-weighted image is a high-resolution image, which are helpful in defining anatomical features. A Malignant tumor appears as a homogeneous low-intensity region against the high-signal-intensity prostate tissue in a T2 image [194]. From diffusion weighted images we derive apparent diffusion coefficient (ADC) images and use the diffusion of water molecules to generate the contrast in MRI as discussed in Section 2.5. A malignant tumor appears as a darker region in ADC images due to restricted diffusion in the tumor regions because cells in tumors are densely packed and do not allow for free diffusion of water [194].

Recent studies have shown that multi-parametric magnetic resonance imaging can provide accurate localization and staging of prostate cancer [126, 60, 17, 126]. Staging is the process of determining how much cancer is within the body (tumor size) and how far it has spread from where it originated. Interpreting MRI sequences manually requires expertise from the radiologist and is time consuming and error prone. Segmentation of prostate tumors is an important step as it helps in further downstream analysis and is also used to guide in-bore biopsy to obtain an accurate diagnosis. Therefore, an automated technique for segmenting tumors from MRI images is highly desirable.

4.2.2 Related Work

Several studies have attempted to segment malignant tumors from multi-parametric MRI images of prostate. Artan et al. classifying pixels of the peripheral zone of prostate as malignant vs. non-tumor by combining support vector machines with conditional random fields and achieved a Dice score of 0.56 using T2 and ADC images and the k_{ep} parametric map from DCE-MRI images [10]. Dice score is an overlap index most commonly used to validate medical segmentation results [145]. It ranges between 0 and 1, where Dice score of 1 denotes perfect overlap and 0 denotes no overlap. Chan et al. used the anatomical location of the tumor and texture features (gray level concurrence matrix and discrete cosine transform) from T2-weighted, T2-mapping and line scan diffusion imaging to classify pixels in the peripheral zone of the prostate as malignant vs. non-tumor using SVM and Fisher's linear discriminant and achieved area under the receiver operating curve (ROC AUC) values of 0.76 and 0.84 respectively [32]. Liu et al. achieved a Dice score of 0.62 using fuzzy Markov random fields for malignant tumor segmentation by combining MR spectroscopic images, T2, DWI and DCE-MRI [115]. Guo et al. used fuzzy c-means clustering and Bayesian modeling to produce a membership degree map, which is used to make a decision on cancer regions and achieved a Dice score of 0.76 by combining MR spectroscopic imaging, T2, DWI and DCE-MRI images [70]. Lavasani et al. extracted semi-quantitative and wavelet-based features from DCE-MRI voxel time courses and used fuzzy c-means clustering to segment malignant tumors and achieved a Dice score of 0.83 [100]. Yang et al. generated a cancer response map at the last convolutional layer of a network trained to classify MRI images as cancerous or benign and achieved a Dice score of 0.66 by combining T2 and ADC images [188]. Kohl et al. used adversarial networks to segment malignant tumors from T2, ADC and DWI images to achieve a Dice score of 0.41 [95].

Several benign conditions, such as fibrosis, post-biopsy hemorrhage and prostatitis, appear as low T2 signal mimicking prostate cancer [94]. Due to similar image properties, over-segmentation of tumor tends to occur in prostate multi-parametric MRI images. In this chapter we explore multiple strategies to reduce the over-segmentation.

4.3 Proposed Approach

4.3.1 Segmentation network

We first design a network based on the U-Net architecture defined in Section 2.2 to segment tumors from T2 and ADC images. We experiment with different depths of the network to determine the best architecture. In order to reduce over-segmentation, we experiment with multiple strategies

that can be broadly classified into two types - modifying the loss function and modifying the U-Net architecture. In addition to the commonly used Dice loss, we experiment with Tversky loss and Focal loss, which are specially designed for datasets with imbalanced classes. We further modify the U-Net architecture by adding attention gates, adding an auto-encoder branch and using deep supervision.

Loss Functions

We experiment with three loss functions - Dice loss, Tversky loss and Focal loss. Dice loss is a standard loss function used for segmentation. Tversky loss and Focal loss were introduced to deal with class imbalance problems. As the tumor occupies a small region in the entire image, we look into the ability of Tversky and Focal losses in improving the segmentation results.

Dice Loss: As described in Section 4.2.2, dice coefficient is an overlap index most commonly used to validate medical segmentation results [145]. It ranges between 0 and 1, where Dice coefficient of 1 denotes perfect overlap and 0 denotes no overlap. Dice coefficient is defined as

$$\text{Dice} = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} = \frac{2TP}{2TP + FP + FN} \tag{4.1}$$

where N is the total number of pixels, p_i and g_i are the binary labels from the predicted segmentation and ground truth labels respectively, TP is true positives, FP is false positives and FN is false negatives. As neural networks require a loss function that can be minimized, we use $1 - Dice$ as the loss function.

$$\text{Dice loss} = 1 - \text{Dice} \tag{4.2}$$

Tversky loss: Tversky loss is a generalization of Dice loss that allows for better control over trade-off between precision and recall [147].

$$\text{Tversky Index} = \frac{|PG|}{|PG| + \alpha|P\backslash G| + \beta|G\backslash P|} = \frac{TP}{TP + \alpha FP + \beta FN} \tag{4.3}$$

Here P and G represent the set of predicted and ground truth binary labels respectively, PG represents the set of pixels labeled as tumor in both the predicted segmentation and ground truth, $P\backslash G$ represents the set of pixels labeled as tumor in the predicted segmentation and non-tumor in the ground truth and $G\backslash P$ represents the set of pixels labeled as tumor in the ground truth and non-tumor in the predicted segmentation. The values of α and β control the magnitude of penalties for false positives and false negatives. In the case of $\alpha = \beta = 0.5$ the Tversky index simplifies to

be the same as the Dice coefficient in Equation 4.1. Similar to Dice loss, $1 - TverskyIndex$ is used as the loss function to train U-Net.

$$\text{Tversky loss} = 1 - \text{Tversky Index} \tag{4.4}$$

Focal loss: Focal loss emphasizes learning hard misclassified examples by applying a modulating term on the cross entropy loss function [111].

$$\text{Focal loss} = -\alpha_t(1 - p_t)^{\gamma} log(p_t) \tag{4.5}$$

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \tag{4.6}$$

Here γ is the focusing parameter, α is a weighting factor to balance classes and p is the predicted probability for the class with label $y = 1$. If an example is classified correctly, its predicted probability p will be close to 1 and therefore the value of p_t will also be close to 1. When p_t is close to 1, the value of the modulation factor $(1 - p_t)^{\gamma}$ becomes close to 0 and therefore the loss for well classified examples is down-weighted. The focusing parameter γ controls the rate at which the easy examples are down-weighted. Therefore, α balances the importance given to positive vs. negative examples, while γ balances the importance given to easy vs. hard examples. If the weight of one class is set to α, the weight of the other class is set to $1 - \alpha$.

U-Net Modifications

Attention U-Net: We design a network similar to U-Net and incorporate attention gates to highlight relevant features for tumor segmentation. The architecture of a basic U-Net and the attention gates were discussed in Section 2.2. The proposed network architecture is shown in Figure 4.1. In the conventional U-Net, feature maps of similar resolution from the encoding path are directly concatenated with the up-sampled feature maps in the decoding path. The tumor only occupies a small portion of the entire image and exhibits large shape variability among patients. Due to successive down sampling in the encoding path, the information regarding the precise location of tumor is lost at coarser spatial levels. Direct concatenation of feature maps to the decoding path lacks locational precision and tends to increase the false positive predictions [131]. Similar to work done by Oktay [131] as discussed in Section 2.2, we modify the U-Net architecture by incorporating attention gates to prune the feature maps before concatenation to preserve only the activations relevant for tumor segmentation and suppress the irrelevant background regions.

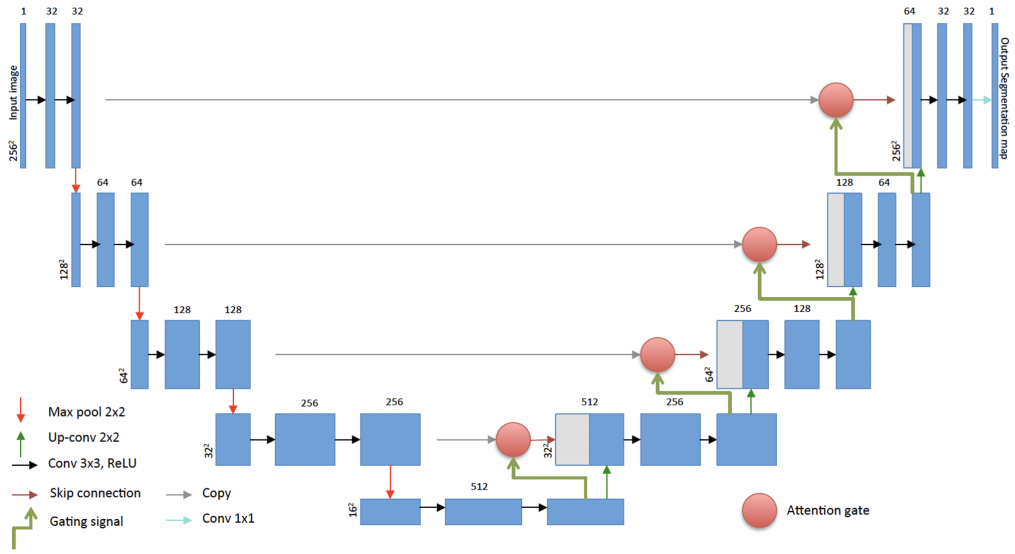


Figure 4.1: U-Net architecture with attention gates.

Auto-encoder branch: Inspired by the work of Myronenko et al. [127], we incorporate an additional variational auto-encoder branch to the decoder of the U-Net to provide added guidance and also regularize the shared encoder as shown in Figure 4.2. By providing an extra task, the encoder is forced to learn richer feature representations. The variational auto-encoder maps the features of the encoder endpoint to a multivariate latent distribution representative of the training images. The output of the encoder endpoint is reduced to a low-dimensional latent space of $2N$ (N to represent mean and N to represent standard deviation). Then a sample is drawn from the latent distribution to reconstruct the input image following the same architecture as the decoder in the U-Net. The loss function here is the sum of the losses from the decoder branch and the VAE branch. The first term of the VAE loss represents the reconstruction error and the second term represents the regularization which ensures that the learnt distribution is similar to a unit Gaussian distribution.

$$L = L_{dice} + \lambda L_{vae} \tag{4.7}$$

$$L_{vae} = \|I_{input} - I_{pred}\|^2 + \frac{1}{N}\sum \mu^2 + \sigma^2 - \log \sigma^2 - 1 \tag{4.8}$$

Deep supervision: Chen et al. introduced the concept of deeply supervised nets for classification by introducing companion objective functions at each hidden layer in addition to the overall objective function [103]. Inspired by this work, we down-sample the final output to the dimensions of the decoder outputs at multiple intermediate levels and calculate the Dice loss at each level of

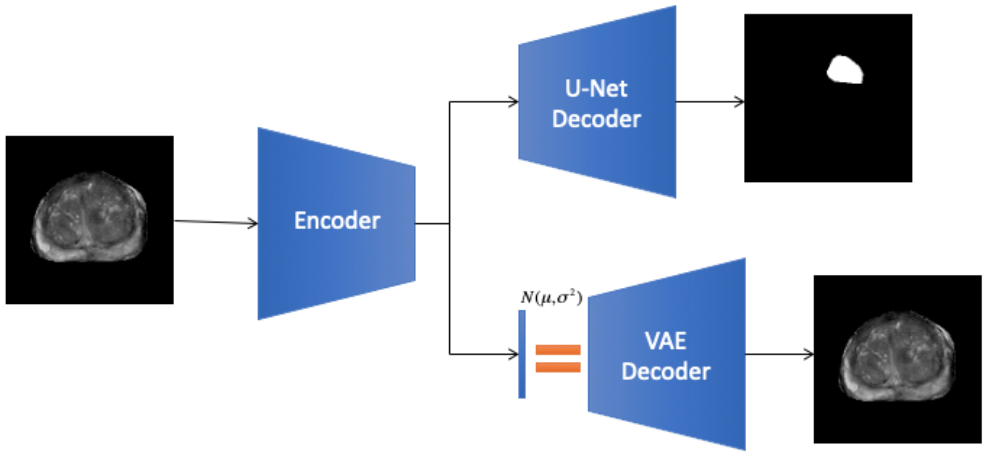


Figure 4.2: U-Net architecture with the added auto-encoder branch.

the U-Net in addition to the final level. The proposed network with deep-supervision is shown in Figure 4.3. Deep supervision serves as a proxy for the discriminativeness of the feature maps at each level and to the quality of the upper layer feature map. It influences the weight update process to favor highly discriminative feature maps.

4.4 Methods

4.4.1 Patient Cohort and Image Acquisition

A number of subjects with elevated PSA but negative transrectal ultrasound guided biopsy were referred to undergo multi-parametric MRI imaging. The T2, ADC and DCE-MRI images were acquired using a Philips system at Oregon Health and Science University in Portland, USA. The in-plane resolution and spacing between slices for T2 were 0.321 mm and 3 mm respectively, and for ADC 1.023 mm and 3.3 mm respectively. Based on the information from all three scans, a radiologist identified tumors in seventy eight subjects and manually outlined the prostate tumor regions on the T2 images. Based on these manually labeled tumor regions, an image guided in-bore biopsy was performed. A specimen obtained from the targeted biopsy provides a more reliable ground truth than that obtained using transrectal ultrasound guided biopsy. Based on the in-bore biopsy, 55 tumors were found to be malignant while 23 were benign.

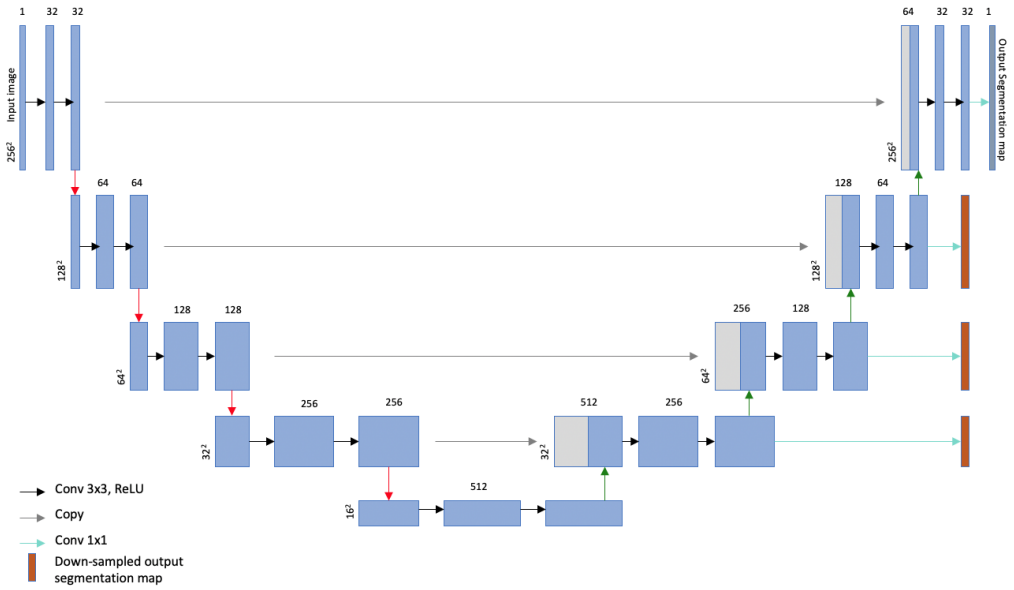


Figure 4.3: U-Net architecture with deep-supervision.

4.4.2 Mapping between T2 and ADC Images

The tumor regions are manually outlined only on T2 images. Due to the difference in in-plane resolution and spacing between slices of T2 and ADC images, the manual segmentation cannot be directly transferred to the ADC images. The ADC images need to be registered with the T2 images in order to transfer the manual segmentation. As seen from Figure 4.4, the T2 and ADC images are too dissimilar to apply any intensity or mutual information based registration algorithms. Therefore, we use the attributes in the DICOM header file to map the 3D coordinates of T2 and ADC images. The DICOM standard defines a reference coordinate system that is patient oriented. The reference coordinate system allows us to measure the position and orientation of an image with respect to the patient. Three attributes in the DICOM header file that provide this information are - *image position* (S, the x, y, and z coordinates of the upper left hand corner of the image, in mm), *pixel spacing* (Δi, Δj the adjacent row and column spacing in mm) and *image orientation* (X, Y the direction cosines of the first row and the first column with respect to the patient). Using these three attributes the T2 and ADC image stacks are converted to a common

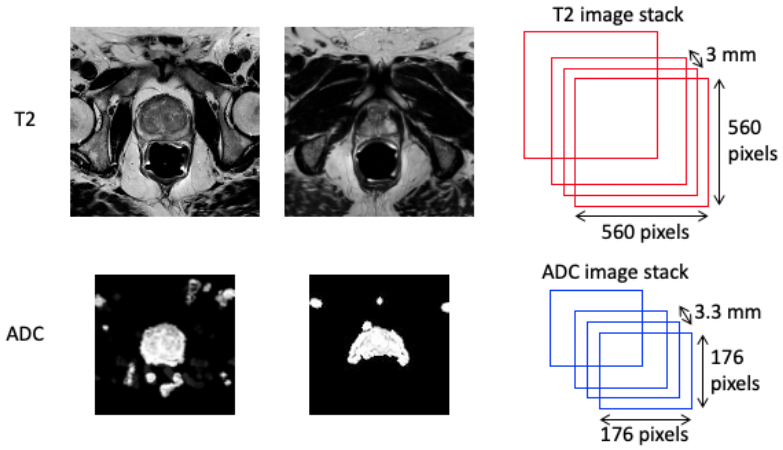


Figure 4.4: T2 and ADC images are dissimilar to apply any intensity or mutual information based registration algorithms.

reference coordinate system based on the following equation.

$$\begin{bmatrix} P_x \\ P_y \\ P_z \\ 1 \end{bmatrix} = \begin{bmatrix} X_x\Delta i & Y_x\Delta j & 0 & S_x \\ X_y\Delta i & Y_y\Delta j & 0 & S_y \\ X_z\Delta i & Y_z\Delta j & 0 & S_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} i \\ j \\ 0 \\ 1 \end{bmatrix} \tag{4.9}$$

Here P_{xyz} represents the 3D coordinates of the pixel (i, j) in the reference coordinate system, S_{xyz} represents the three values of *image position* (which gives the position of the center of the first voxel in the reference coordinate system), X_{xyz} and Y_{xyz} represent the X and Y direction cosines of the *image orientation* and Δi and Δj represent the *pixel spacing*. The 3D coordinates of pixels in T2 and ADC image stacks in the reference coordinate system are obtained using Equation 4.9. The ADC images are up-sampled by interpolating the 3D ADC data and querying for its value corresponding to each (x, y, z) location from the T2 image stack. The process of mapping T2 and ADC images is represented in Figure 4.5 and the results after mapping images from both modalities are shown in Figure 4.6.

4.4.3 Implementation details

The proposed and the baseline U-Nets are implemented using the Keras and TensorFlow [1] libraries. We use an Adam optimizer with a learning rate of 1e-4. We normalize all input images to have zero mean and unit standard deviation. We augment data using rotation (0-180 degree), horizontal mirror flips, zooming (0-20%), and horizontal and vertical shifts (0-20%) of the image. We use a batch size of 1, drawing each training image once per epoch in a random order.

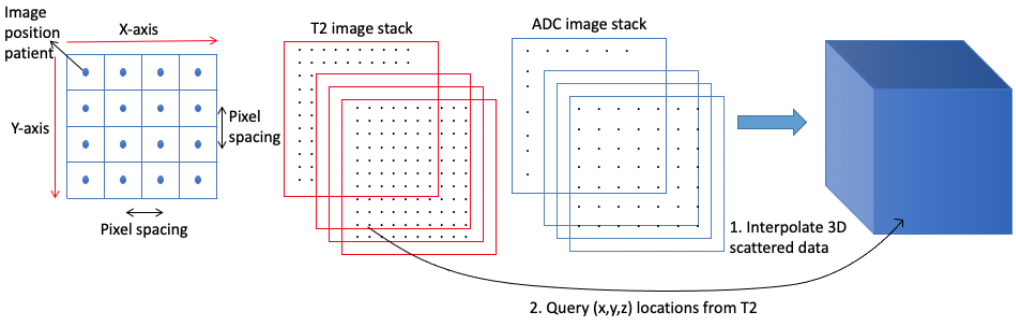


Figure 4.5: Process of establishing correspondences between image stacks.

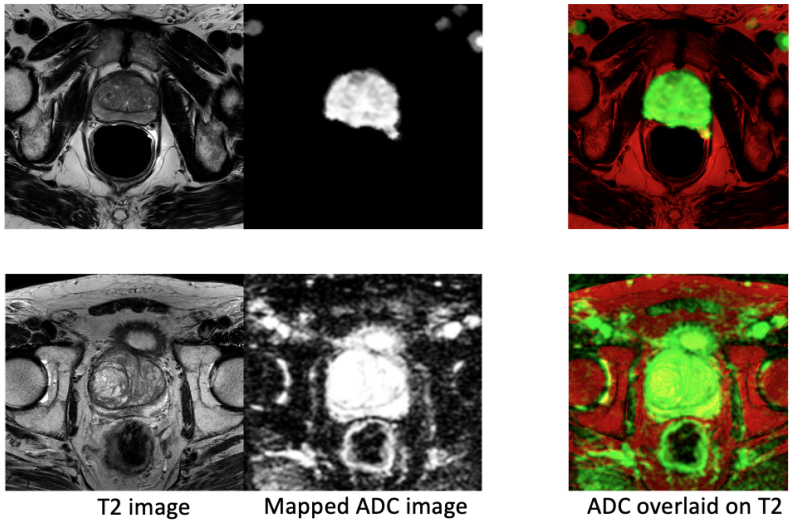


Figure 4.6: Mapped T2 and ADC images

The T2 and ADC image slices of the 78 patients are divided into test and train sets as shown in Table 4.1. All 2D images from a given patient would go to either train or test set but not both. For the experiments each slice is considered as a separate case. The total number of T2 and ADC slices are 165 and 132 for malignant and 65 and 48 for the benign cases respectively. The number of slices in ADC is lower than T2 as ADC signal is distorted in few cases due to the presence of metal implants. We use U-Net as the baseline to compare the performance of all other modifications in segmenting tumors from T2 and ADC images of the prostate.

4.4.4 Evaluation metric

The tumor segmentation is evaluated using three metrics: Dice coefficient, recall, precision. The Dice coefficient provides a measure of overlap between the segmentation output and the ground truth. Recall is the fraction of tumor pixels correctly predicted as tumor and precision is the

Table 4.1: Train and test data in each category by subjects and slices.

		Total	Train	Test
Malignant	Number of subjects	55	43	12
	Number of T2 slices	156	117	38
	Number of ADC slices	132	102	30
Benign	Number of subjects	23	18	5
	Number of T2 slices	65	51	14
	Number of ADC slices	48	38	10

fraction of the predicted pixels that belong to the tumor. Here the recall and precision are metrics calculated on the tumor pixels and not on the fraction of patients. If P is the predicted output and G the ground truth, the Dice, recall and precision measures are represented as the following:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} \tag{4.10}$$

$$\text{Recall} = \frac{|P \cap G|}{|G|} \tag{4.11}$$

$$\text{Precision} = \frac{|P \cap G|}{|P|} \tag{4.12}$$

4.5 Results

We concentrate on segmenting the tumor (from both malignant and benign cases), as tumor segmentation is required to guide the in-bore biopsy. Segmentation results using U-Net are presented in Section 4.5.1, while results on different loss functions and U-Net modifications are presented in Sections 4.5.2 and 4.5.3 respectively. In Section 4.5.4, we briefly explore the segmentation of malignant tumors alone. In Section 4.5.5, we explore two ways of combining data from T2 and ADC images to improve the segmentation performance.

4.5.1 Tumor Segmentation using U-Net

We first present the tumor segmentation results obtained using U-Net with Dice loss function. We initially use the entire T2 and ADC images to segment the tumors as shown in Figure 4.7. But we observe that in a few cases, regions outside the prostate are also segmented as tumors, as in the last

column in Figure 4.7. There is an automatic prostate segmentation algorithm already employed in the system capturing the T2 and ADC images. We therefore can segment the prostate region from the T2 and ADC images and perform tumor segmentation only on the prostate region. The qualitative results of performing segmentation on prostate region alone are shown in Figure 4.8. Using the prostate alone to segment the tumors increases the segmentation performance as shown in Table 4.2.

The overall segmentation performance was better on ADC images than on T2, due to the highly discriminable presentation of the malignancy in the ADC images as a darker region against the generally bright appearance of prostate. T2 images have rich anatomical information and present a higher chance of other regions mimicking the low intensity signal characteristic of tumor tissue.

We experiment with the depth of the U-Net to design a network suitable for the tumor segmentation problem. We observe that a 4 level deep network with 512 filters in the deepest layer (Figure 2.3) resulted in the best Dice score for T2 images, while a 5 layer network resulted in best performance for ADC images as seen in Table 4.3. We start with 32 filters and double the number of filters as we go down each level. A network with 64 filters represents a one level U-Net. For all evaluations in the following sections we used the 4-level deep U-Net with 512 filters at its deepest layer as the base against which the performance of the modified networks are evaluated. This is also the network used to obtain the results in Table 4.2. Therefore the best Dice, precision and recall values obtained for this setting for T2 images are 0.49, 0.77 and 0.36 respectively and for ADC images are 0.61, 0.88 and 0.46 respectively.

As seen in Table 4.4, 50% of the tumor pixels are segmented in 88% of slices in T2 and 85% of the slices in ADC. Using U-Net the overall recall is high, i.e. the tumor pixels are segmented correctly, but the precision is low, i.e. a portion of the predicted pixels are not tumorous, in other words there is over-segmentation, as can be seen in Figure 4.8 and Table 4.3. We employ several techniques to reduce the over-segmentation and improve the overall segmentation accuracy. We present the results of using different loss functions in Section 4.5.2 and different modifications of the U-Net architecture in Section 4.5.3.

4.5.2 Loss Functions

We use the 4-level deep U-Net with 512 filters at its deepest layer trained with Dice loss as the base and perform all modifications on this architecture. To improve the segmentation performance we employ 2 loss functions that are designed for imbalanced datasets. The tumor occupies a very small region on the whole image and therefore the number of tumor pixels are much lower than the number of non-tumor pixels.

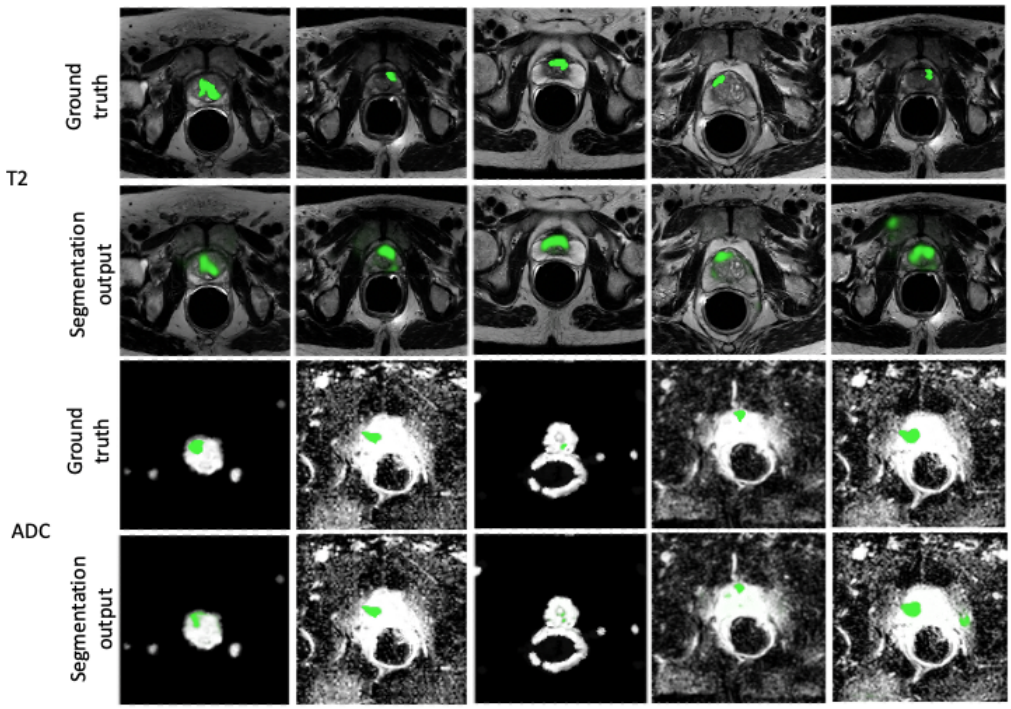


Figure 4.7: Tumor segmentation from the full T2 and ADC images using U-Net. First and third rows show T2 and ADC images respectively, with ground truth labels of tumors marked in green. Second and fourth rows show the same T2 and ADC images with segmentation result from U-Net marked in green.

We first look at the Tversky loss (Equation 4.3). A higher value of α in the Tversky loss will shift the emphasis to reduce false positives and boost the precision. The results are shown in Table 4.5 using a range of values for α and β. With for $\alpha = 0.75$ and $\beta = 0.25$ in the Tversky loss, dice score improved from 0.49 to 0.55 in T2 images and 0.61 to 0.67 in ADC images. But with the increase in precision value we observe a slight decrease in the recall value. This could be due to the fact that some non-tumor regions display features similar to that of the tumor and when the network is encouraged to avoid such regions, certain regions of the tumor also get avoided, leading to the reduced recall value.

We next look at the Focal loss (Equation 4.5). We experiment with values of focusing parameter γ and weighting factor α in the Focal loss. The results for various values of the parameters are shown in Table 4.6. As we increase the focusing parameter γ, we observe an increase in the recall value but a decrease in the precision value. It seems to focus on the misclassified tumor pixels but not on the misclassified non-tumor pixels. There is no improvement in Dice score in T2 images but in ADC images we obtain a Dice score of 0.67 with Focal loss compared to 0.61 with Dice loss.

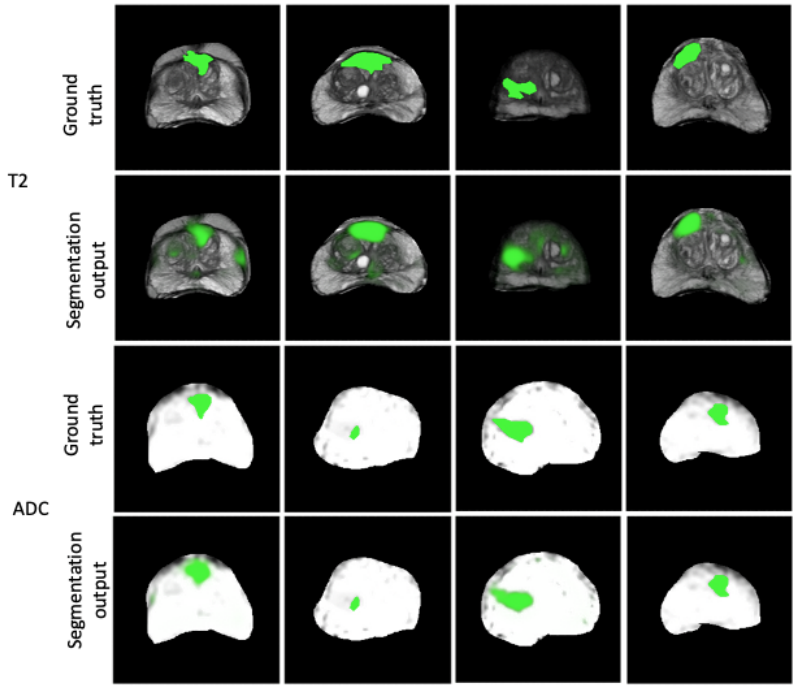


Figure 4.8: Tumor segmentation from segmented prostate regions in T2 and ADC images using U-Net. First and third rows show T2 and ADC images respectively, with ground truth labels of tumors marked in green. Second and fourth rows show the same T2 and ADC images with segmentation result from U-Net marked in green.

On closer observation, this increase is not due to the focusing effect of the Focal loss but rather due to better balancing of weights between the two classes as seen from Table 4.6 with $\gamma = 0$ and $\alpha = 0.75$. When $\gamma = 0$, Focal loss is equivalent to weighted cross entropy loss. Weighting the tumor pixel more allows the network to pay more attention to pixels representing tumor. This weighting is helpful as there are very few pixels representing tumor compared to non-tumor in the image, and cross-entropy loss without this weighting would be dominated by the contribution from the non-tumor pixels.

Overall, we observe that the Tversky loss improves the segmentation performance as the weight of the false positive and false negative terms could be controlled unlike in the Dice loss function. Experiments using Focal loss show that modifying the weights of the classes in the loss function to represent the class imbalance improves the results in ADC images but did not help in T2 images.

4.5.3 U-Net Modifications

Though we observe better performance using different loss functions, in this section we still use the 4-level deep U-Net with 512 filters at its deepest layer trained with Dice loss as the base. Table 4.7

Table 4.2: Quantitative results comparing tumor segmentation from full image and prostate alone.

Modality	Image type	Dice	Recall	Precision
T2	Full	0.43	0.63	0.33
	Prostate	0.49	0.77	0.36
ADC	Full	0.54	0.70	0.44
	Prostate	0.61	0.88	0.46

Table 4.3: Dice score for different network sizes for tumor segmentation.

Number of filters in deepest encoder level	T2			ADC		
	Dice	Recall	Precision	Dice	Recall	Precision
64	0.33	0.82	0.21	0.58	0.83	0.44
128	0.45	0.75	0.32	0.60	0.84	0.47
256	0.49	0.79	0.36	0.62	0.84	0.49
512	**0.49**	**0.77**	**0.36**	0.61	0.88	0.46
1024	0.39	0.82	0.26	**0.64**	**0.80**	**0.54**

presents the segmentation performance using the three modifications on the U-Net architecture. Integrating attention gates appears to increase the precision and the overall Dice score in both T2 and ADC images. The attention gates prune the feature maps to suppress features of irrelevant background regions from showing up in the final segmentation map. This reduces the false positive predictions and thus improves the overall precision.

Addition of the auto-encoder branch slightly helps ADC images but not T2 images. The main idea behind adding the auto-encoder branch is to provide an auxillary task that can offer extra guidance to the encoder in extracting better representative features. However, with the addition of the auto-encoder branch, the encoder needs to learn features that are both representive of the input image to reconstruct it and discriminative enough to enable segmentation of the tumor. We assume the poor performance in T2 images is due to its rich representation of anatomical information. The features required to differentiate the tumor pixels from the non-tumor pixels could be different from the features required to reconstruct the rich anatomical information in T2

Table 4.4: Percentage of slices tumor was identified based on recall.

	Recall > 5%	Recall > 10%	Recall > 50%	Recall > 80%
T2	100%	100%	88%	78%
ADC	100%	98%	85%	65%

Table 4.5: Quantitative results comparing tumor segmentation for different setting of α and β in the Tversky loss function.

		T2			ADC		
α	β	Dice	Recall	Precision	Dice	Recall	Precision
0.5	0.5	0.49	0.77	0.36	0.61	0.88	0.46
0.67	0.33	0.50	0.54	0.47	0.66	0.69	0.63
0.75	0.25	0.55	0.60	0.51	0.67	0.66	0.69
0.8	0.2	0.53	0.56	0.51	0.63	0.61	0.65
0.83	0.17	0.49	0.54	0.46	0.63	0.58	0.69

Table 4.6: Quantitative results for prostate tumor segmentation for different setting of γ and α in the Focal loss function.

		T2			ADC		
γ	α	Dice	Recall	Precision	Dice	Recall	Precision
0	0.75	0.49	0.79	0.35	0.67	0.88	0.54
0.1	0.75	0.40	0.86	0.26	0.60	0.86	0.51
0.5	0.75	0.43	0.85	0.29	0.61	0.89	0.47
0.2	0.5	0.42	0.86	0.28	0.59	0.88	0.47
0.5	0.5	0.49	0.77	0.36	0.60	0.88	0.46
1	0.25	0.43	0.86	0.29	0.59	0.90	0.44
2	0.25	0.30	0.96	0.18	0.45	0.94	0.30

Table 4.7: Quantitative results for prostate tumor segmentation for different modifications to U-Net.

Network type	T2			ADC		
	Dice	Recall	Precision	Dice	Recall	Precision
U-Net	0.49	0.77	0.36	0.605	0.88	0.46
Attention U-Net	0.52	0.55	0.49	0.644	0.67	0.62
U-Net VAE	0.38	0.76	0.25	0.606	0.85	0.47
U-Net with deep supervision	0.50	0.80	0.36	0.637	0.88	0.50

images. This mismatch in the feature set required by the two tasks could be the reason for the reduced performance. On the other hand, ADC images contain simpler image content, and the features required to reconstruct these images could be simpler and more aligned with the features required for the segmentation task. Therefore, in ADC images the auto-encoder branch works as a complementary task enhancing the performance while in T2 images it degrades the performance.

Deep supervision increases the Dice score as it forces the network to learn better features at each level of the U-Net. Providing feedback at each level offers an extra input to correct the weights and forces the network to learn better features at each level.

4.5.4 Prostate Cancer Segmentation

So far we have been segmenting tumors - both malignant and benign - from the T2 and ADC images. The results so far have not been as strong as what we hoped for. The low performance could be due to the presence of benign tumors that present themselves with features similar to malignant sometimes and non-tumor regions sometimes as seen in Figure 4.9. We therefore try to access the segmentation performance on malignant tumors alone.

Segmentation of Malignant Tumors

Here we remove all the benign cases from the dataset and analyze the performance of the network in segmenting tumors only in the malignant cases. Examples of segmentation results obtained on malignant tumors in T2 and ADC images using U-Net and U-Net with attention mechanism trained with Dice loss function are shown in Figure 4.10. Using U-Net the mean Dice, recall and precision values obtained for T2 images are 0.58, 0.70 and 0.49 respectively and for ADC images

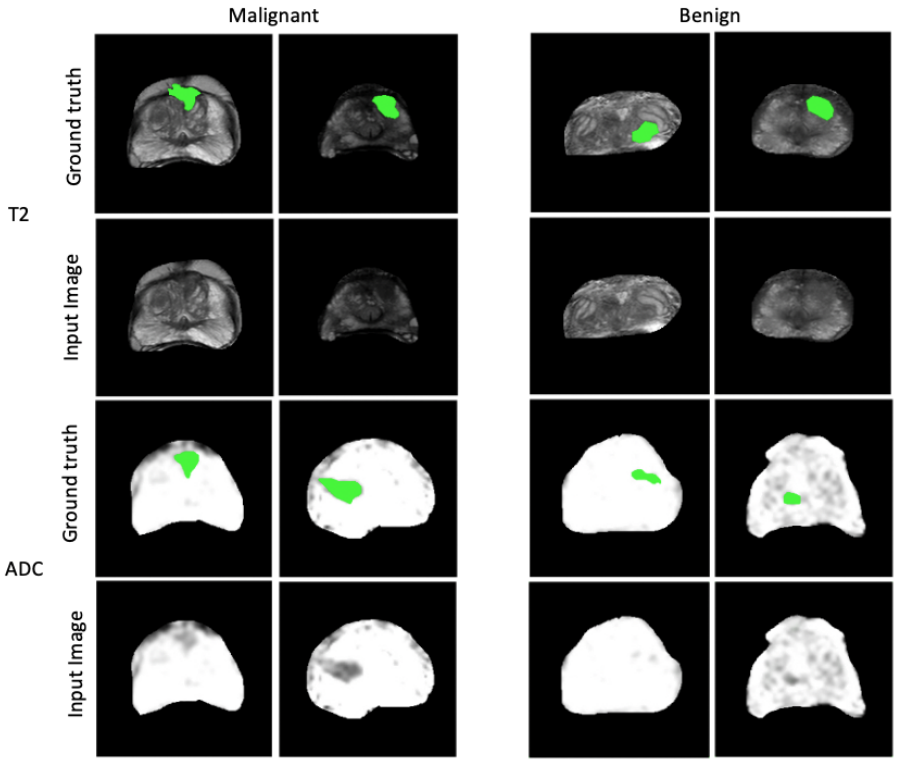


Figure 4.9: Example of malignant and benign tumors. The benign tumors did not have significant distinguishing image features.

are 0.71, 0.70 and 0.71 respectively. The Dice score increased from 0.49 to 0.58 for T2 images and 0.60 to 0.71 in ADC images by removing the images with benign tumors. Integrating attention units appears to increase the precision and the overall Dice score in both T2 and ADC images as observed from Table 4.8. Percentage of T2 and ADC slices with greater than 50% of the malignant pixels correctly detected increased with the addition of attention mechanism as seen in Table 4.9. The Dice score obtained in malignancy segmentation analyzing each modality separately is higher than some of the studies using a combination of MRI images. Combining features from both T2 and ADC images could further improve accuracy of malignancy segmentation.

4.5.5 Multi-Parametric Segmentation - Combining T2 and ADC

The Dice score obtained in malignancy segmentation analyzing each modality separately is higher than some of the studies using a combination of MRI images. Combining features from both T2 and ADC images could further improve accuracy of malignancy segmentation. We experiment with two methods of combining the information from T2 and ADC images to improve the segmentation performance. In the first method, we stack T2 and ADC images and provide that as the input

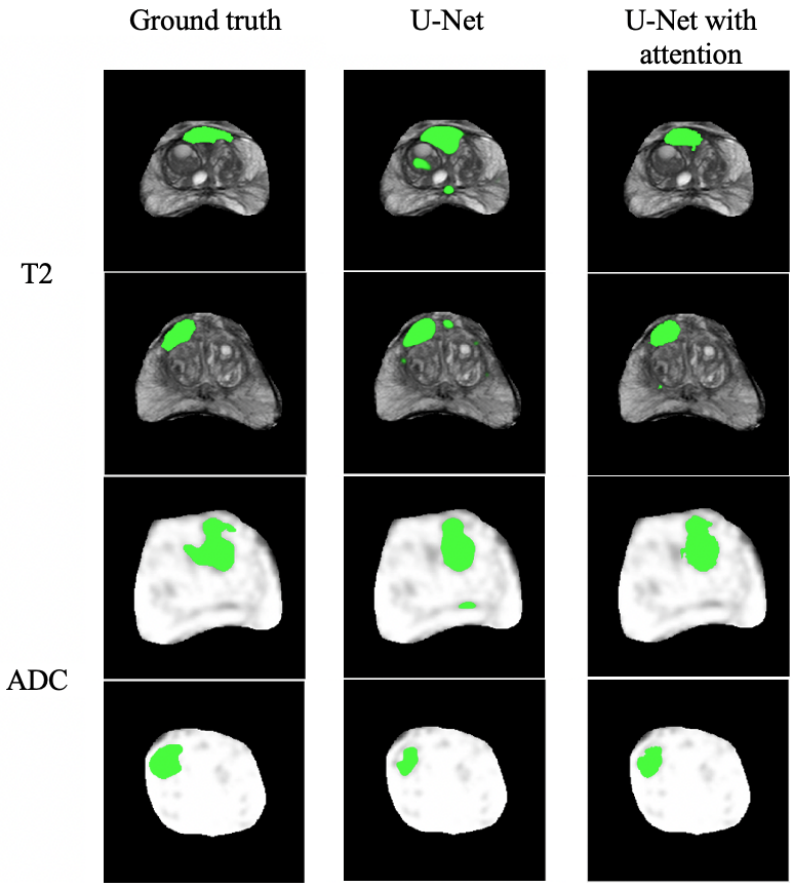


Figure 4.10: Malignant tumor segmentation on T2 and ADC images using classic and attention U-Nets. The green region in the ground truth column indicates the prostate cancer manually outlined by the radiologist and in the U-Net and U-Net with attention columns indicate the predicted prostate cancer regions.

Table 4.8: Comparison of Dice score, recall and precision using classic U-Net and U-Net with attention gates for malignancy segmentation.

	T2			ADC		
Network type	Dice	Recall	Precision	Dice	Recall	Precision
U-Net	0.58	0.70	0.49	0.71	0.70	0.71
Attention U-Net	0.59	0.69	0.51	0.74	0.77	0.72

Table 4.9: Percentage of slices malignant tumor was identified on based on recall using classic U-Net and U-Net with attention gates.

		Recall > 50%	Recall > 90%
T2	U-Net	76%	18%
	U-Net with attention	84%	18%
ADC	U-Net	77%	17%
	U-Net with attention	90%	43%

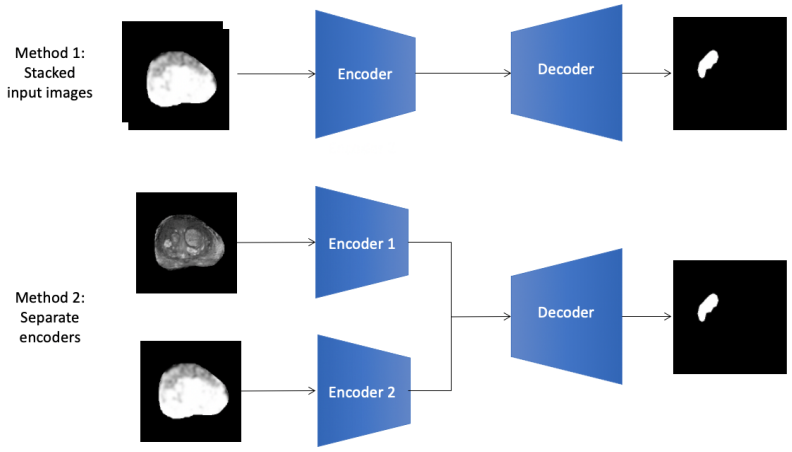


Figure 4.11: Methods of combining T2 and ADC image features for tumor segmentation.

to the U-Net and in the second method we design a separate encoder for each ADC and T2 and concatenate their features and pass them through a single decoder. Both methods are illustrated in Figure 4.11. The use T2 and ADC images aligned using the method in Section 4.4.2 as inputs. As seen from Table 4.10 and Figure 4.12, stacking T2 and ADC images did not improve the segmentation accuracy. Encoding T2 and ADC images with separate encoders improves the segmentation performance over considering T2 images alone but is much lower than the performance obtained by considering ADC images alone.

4.6 Conclusion

In this chapter we improve the tumor segmentation performance in T2 and ADC images of prostate by incorporating several modifications to the standard U-Net architecture. We observed that the Tversky loss improved the segmentation performance as the weight of the false positive and false

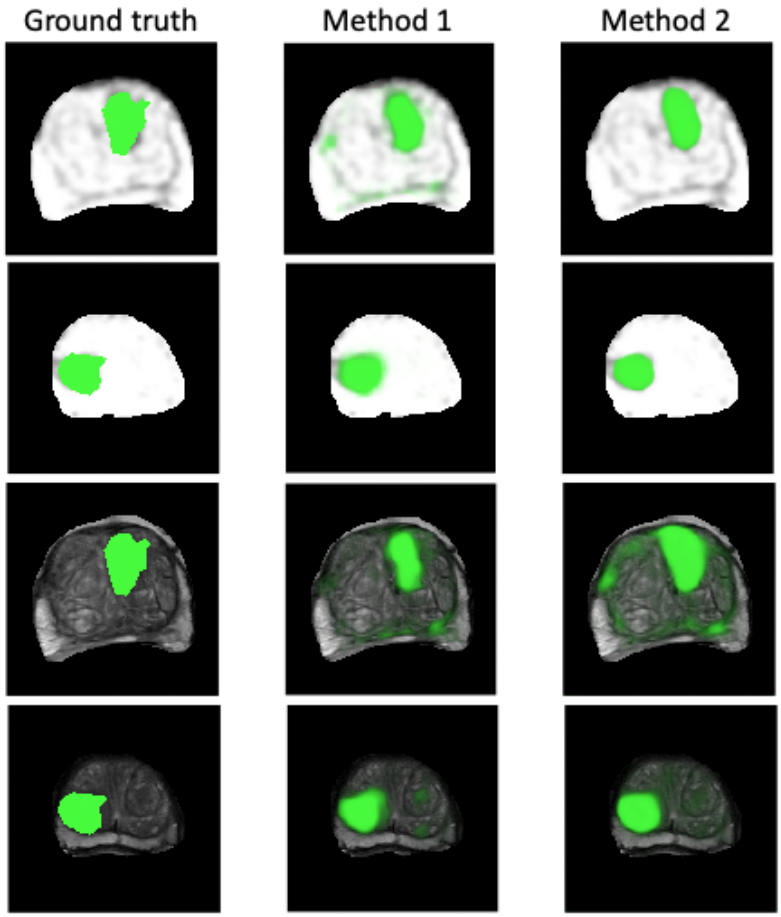


Figure 4.12: Segmentation results using features from both T2 and ADC images.

Table 4.10: Quantitative results for prostate tumor segmentation using T2 and ADC images simultaneously.

	Dice	Recall	Precision
T2	0.49	0.77	0.36
ADC	0.61	0.88	0.46
Method 1	0.40	0.89	0.26
Method 2	0.55	0.80	0.42

negative terms could be controlled unlike in the Dice loss function. Experiments using Focal loss showed that modifying the weights of the classes in the loss function to represent the class imbalance improved the results in ADC images but did not help in T2 images. U-Net with attention gates outperformed the baseline U-Net in Dice score and precision on both T2 and ADC images. The improvement is due to the pruning of feature maps by the attention gates to retain only features relevant for tumor segmentation, thus reducing false positive predictions. Addition of auto-encoder branch helped in ADC images but not in T2 images. We assume this is because ADC images contain less information and the features used for segmentation and reconstruction are similar. On the other hand, T2 images exhibit richer data representation. Reconstruction of T2 images by the auto-encoder probably requires different feature representation than those required by the decoder for segmentation. Therefore it does not help the encoder in learning features helpful for segmentation, but rather deteriorates the segmentation performance as the network has to learn a more diverse set of features. Deep-supervision forces the features to be highly discriminative at every level of the network and therefore increased the segmentation performance. The two approaches of combining information from T2 and ADC images did not produce promising results.

In further studies, other methods of combining information from T2 and ADC images need to be explored. In this chapter, each modification to the U-Net has been analyzed separately. A combination of these modifications can lead to further improvement of the segmentation performance. The results suggest that an attention U-Net with deep supervision trained with Tversky loss could increase the segmentation performance. We used only T2 and ADC images from the multi-parametric MRI suite. DCE-MRI images also contain information pertinent to tumor identification in prostate and the ability of its features in segmenting tumors can be explored in further studies.

Chapter 5

Segmentation: Combining Learning and Tracking-Based Approaches

5.1 Introduction

In the previous chapter, we explored various methods to improve the segmentation performance of U-Net, which performs semantic segmentation. Semantic segmentation approaches cannot separate multiple instances of objects within a category. Instance segmentation networks have been proposed to segment object instances, but they rely on the similarity in appearance of multiple instances of the object in order to learn its representation and segment it. However, in certain applications, such as cell segmentation in electron microscopy images, multiple instances of the object are highly variable and cannot be learnt from a small dataset. Instead, an alternate approach combining information from multiple techniques could aid in separating instances in such settings. In this chapter, we combine a learning based approach - semantic segmentation network - with a tracking based approach - optical flow - to segment instances of cells in electron microscopy images.*

We apply the proposed approach to segment cells and explore convolution and attention-based networks to segment a few key organelles pertinent to cancer development, including nuclei, nucleoli, mitochondria, endosomes and lysosomes from 3D FIB-SEM images [203, 46]. We segment six datasets comprising tissues from breast and prostate cancer and a microspheroid prepared using a breast cancer cell line. Microspheroid refers to cells cultured in a lab. Targeting the structural and functional changes in intra-cellular organelles is emerging as a promising therapeutic approach for

*The work in this chapter is being submitted as a journal article and has been accepted as an extending abstract in A. Machireddy, G. Thibault, K. G. Loftis, K. Stoltz, C. E. Bueno, H. R. Smith, J. L. Riesterer, J. Gray, and X. Song, A Framework to Segment Cellular Ultrastructure from 3D Electron Microscopy Images of Human Biopsies. In Microscopy and Microanalysis 2022.

treatment of cancer [36, 55, 112, 59]. The segmentation of the 3D FIB-SEM image stack of the biopsy sample enables visualization and quantification of ultrastructural inter- and intra-cellular compositions and interactions, and makes it possible to not only identify potential nanoscale sites for targeted therapies, but also construct an ultra-structural atlas of developing cancer from normal to malignant.

5.2 Background

Recent advances in tumor biology have shown that the plethora of interactions between tumor cells and their surrounding environment can significantly influence the behavior of cancer and response to treatment [161, 80, 56]. A deeper understanding of the underlying cellular mechanisms will shed light on how cancer evolves and develops resistance to therapy [165]. Understanding of these dynamic interactions can be used to develop novel approaches to disrupt key inter- and intracellular interactions and facilitate the design and development of efficient therapeutic strategies to fight cancer [12].

Electron microscopy (EM) provides nanometer resolution views of intra and intercellular interactions that are not apparent in images generated using light microscopy [89]. This complete picture of spatial relationships can reveal potential therapeutic targets that can be related back to the macroscale heterogeneity and microenvironment of the tissue. Focused ion beam scanning electron microscopy (FIB-SEM) is especially informative, generating 2D stacks of SEM images that provide 3D information on sub-cellular features in large tissue volumes [62]. FIB-SEM imaging proceeds via serial steps of SEM imaging of a sample surface and FIB removal of a uniform thin layer of the tissue with size comparable to the spatial resolution in x-y plane, thereby revealing a new surface to be imaged. This process is fully automated and ensures that imaged data is equidimensional in all three axes, which significantly improves the accuracy of feature recognition within the dataset. The workflow from FIB-SEM imaging and volume rendering in shown in Figure 5.1.

While 3D FIB-SEM images are being generated with ever increasing rate in ongoing clinical programs, the rate limiting step in their analysis is the delineation of the cellular structures to enable rendering of the images into interpretable forms. Currently, this is done by experts manually annotating images, and while effective, is extremely time consuming, tedious and dependent on the skill of the expert. The development of rapid, robust and automated machine learning methods to segment ultrastructural features is acutely needed for wide spread use of EM in large scale studies [136]. Ultrastructure refers to the structure of the cell and its organelles that is visible only with the high magnification obtainable with electron microscope.

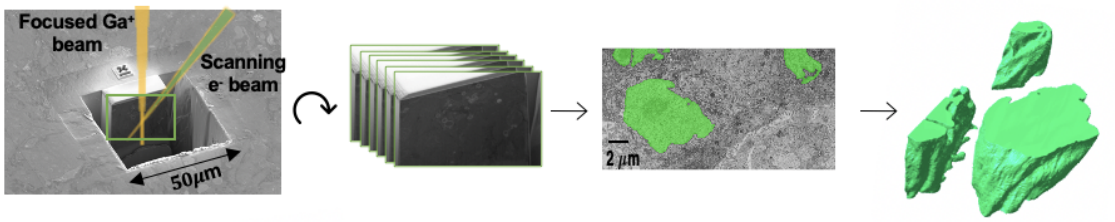


Figure 5.1: FIB-SEM-to-volume rendering workflow. FIB-source sequentially slices a few nanometers from the sample to expose a fresh surface for subsequent imaging using the electron beam. An image stack is acquired and after image alignment and cropping, a small subset of the stack of images was segmented manually to generate a training set for the deep learning model. Once trained, the deep learning model is used to predict segmentation masks for the rest of the images in the stack. These predictions are used to create volume renderings for examination of 3D ultrastructural properties.

5.3 Related Work

Early research on automated EM segmentation was focused on neuronal ultrastructure segmentation of brain tissue. 2D U-Net and its variants using residual connection such as FusionNet [140], fully connected networks (FCN) with skip connections [49] and M2FCN [151] have been proposed for neuronal membrane segmentation in EM images to yield large neurite superpixels which are combined into 3D neuronal object using iterative region agglomeration algorithms. 2D CNNs have the advantage of requiring only 2D ground truth and having a computationally inexpensive training process [85]. The recent automated neurite reconstruction methods use 3D CNNs to yield highly accurate 3D boundary probability maps requiring a simple watershed algorithm for final segmentation [113, 197].

More recently, efforts have been made towards segmentation of organelles in EM images of non-neural tissues. 3D U-Net followed by mutex watershed algorithm was used to segment nuclei and Lifted Multicut-based approach was used for cell segmentation to characterize the morphology of cells in an EM volume of a complete Platynereis worm [181]. Insulin secretory granules and Golgi apparatus segmented using 3D and 2D U-Net respectively helped in generating a comprehensive spatial map of organelle interactions in mouse β cells and facilitated understanding of the supportive role played by secretory granules in insulin secretion [125]. 3D U-Net was used to segment 35 organelles from EM images of HeLa, Jurkat, Macrophage and SUM159 cells [76]. Subcellular structures of liver tissue in mice segmented using 3D U-Net helped demonstrate substantial alterations in hepatic endoplasmic reticulum of lean and obese mice [134]. Though the 3D CNNs exhibit enhanced performance, they necessitate greater number of labelled images for training. In order to reduce the amount of manual labeling required per volume, in this chapter

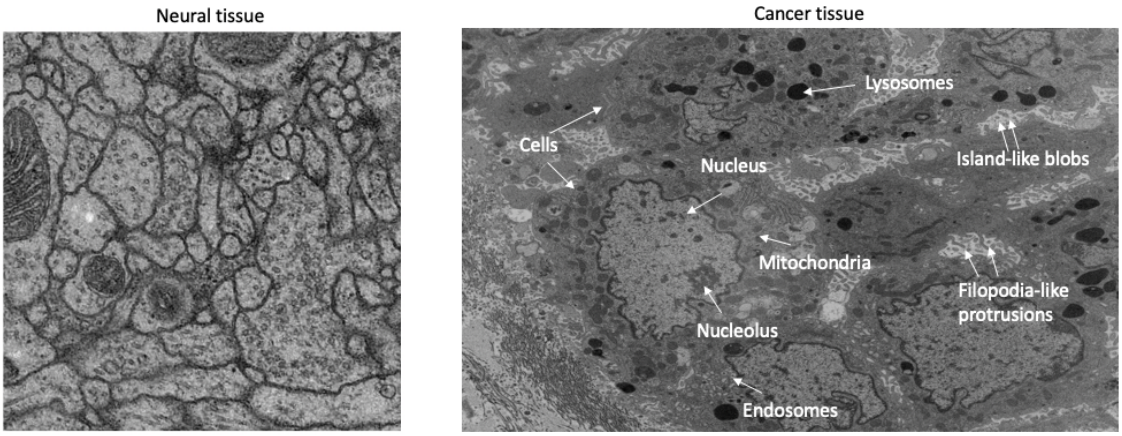


Figure 5.2: Electron microscopy images of a neural tissue and a cancer tissue.

we train models on a small subset of labeled 2D images.

Segmentation of individual cells from EM images is an essential step to perform quantitative cellular analysis. But the ultrastructure of tumor cells and the tumor microenvironment is different from the widely studied neural and other normal cells as seen from Figure 5.2 [11, 203, 41]. The segmentation methodologies designed for the neural cells use cell boundaries as the strongest cue to delineate a cell [151, 113, 181, 125]. Due to the lack of clear boundaries separating cells in cancer tissues, segmentation methodologies designed for neural cells cannot be directly applied for cancer cell segmentation. The convoluted and intertwined nature of cancer cells and the presence of filopodia-like protrusions make it even more challenging [89]. Recently, optical flow has been used to segment cells in EM images of Hepatoblastoma-patient-derived xenograft tissue [45, 52]. In that work, cells were labelled on one image for every 10 images and propogated to the images in-between using optical flow. Propagation-based methods can only propagate the contour of the cells labeled in the ground truth images. But the filopodia-like protrusions present in cancer cells appear as island-like blobs detached from the main cell and a propagation-based method cannot track such islands if they are not present in the ground truth image.

5.4 Proposed approach

We aim to segment the cells and five cell organelles - nuclei, nucleoli, mitochondria, endosomes and lysosomes - from 3D FIB-SEM images. As the organelles mostly appear as distant objects with boundaries they can be segmented using a semantic segmentation network. We use ResUNet

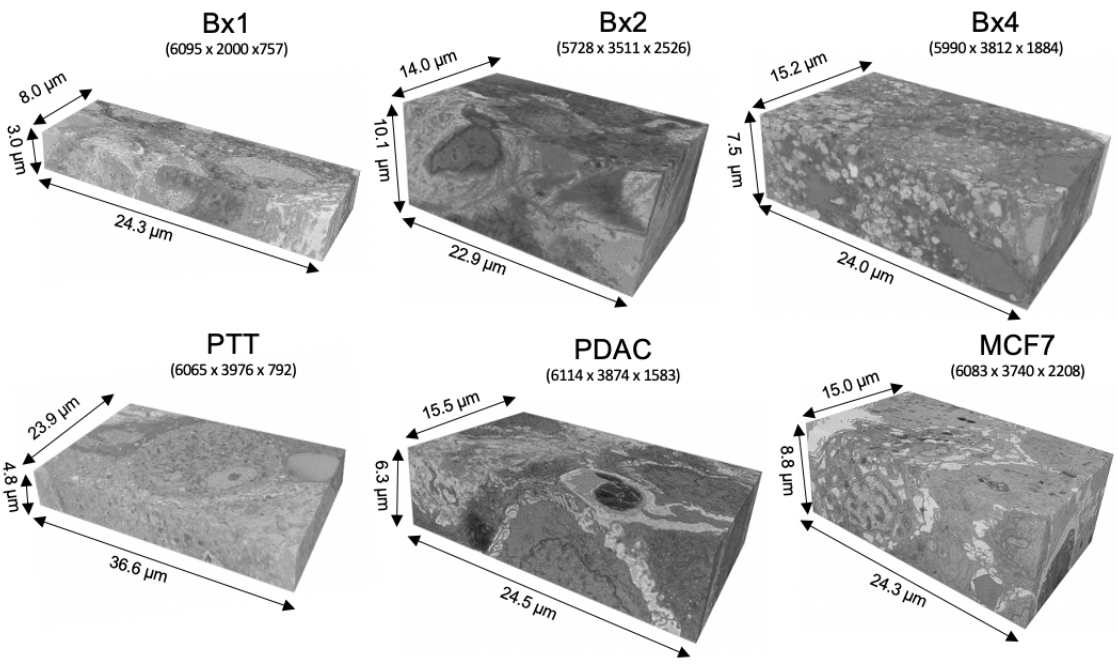


Figure 5.3: The six FIB-SEM datasets and their sizes. The 3D FIB-SEM volumes collected from the biopsy samples Bx1, Bx2 and Bx4 acquired from a patient with metastatic breast ductal carcinoma, two biopsy sample (PTT, PDAC) acquired from two patients with pancreatic ductal adenocarcinoma and a microspheroid prepared using a breast cancer cell line (MCF7)

as the base network for semantic segmentation where we combine residual blocks with the U-Net architecture. Similar to U-Net described in Section 2.2, the proposed network consists of an encoding path that extracts features, and a decoding path that up-samples the extracted features to obtain full-resolution segmentation, but uses residual blocks of convolutional layers as building units (Figure 5.5). A residual block consists of two convolutional layers with a kernel size of 3×3 each preceded by batch normalization and a rectified Linear Unit (ReLU), along with a residual shortcut connection (Figure 5.4).

The encoding path in the proposed network consists of four residual blocks. Instead of using pooling to down-sample the feature maps, a convolution with stride two is performed in the first layer of the residual block, which reduces the computational load on the subsequent layers, as they perform computations on a smaller image. The number of feature maps is doubled along each successive block in the encoding path to enable richer feature extraction. The encoding path is followed by a residual block which acts as a bridge between the encoding and decoding paths. Corresponding to the encoding path, the decoding path also consists of four residual blocks. The decoding path begins with the up-sampling of the feature maps found in the previous level of the

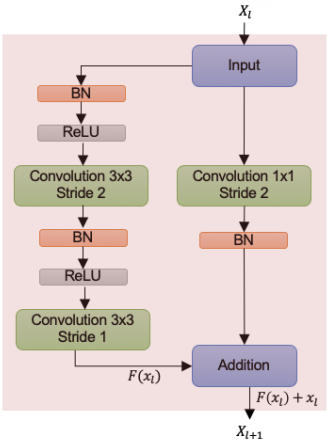


Figure 5.4: Residual block used in ResUNet. Residual block used in ResUNet, BN stands for batch normalization and ReLU stands for rectified linear unit. X_l and X_{l+1} are the input and output features for the residual layer l, and F represents the residual function

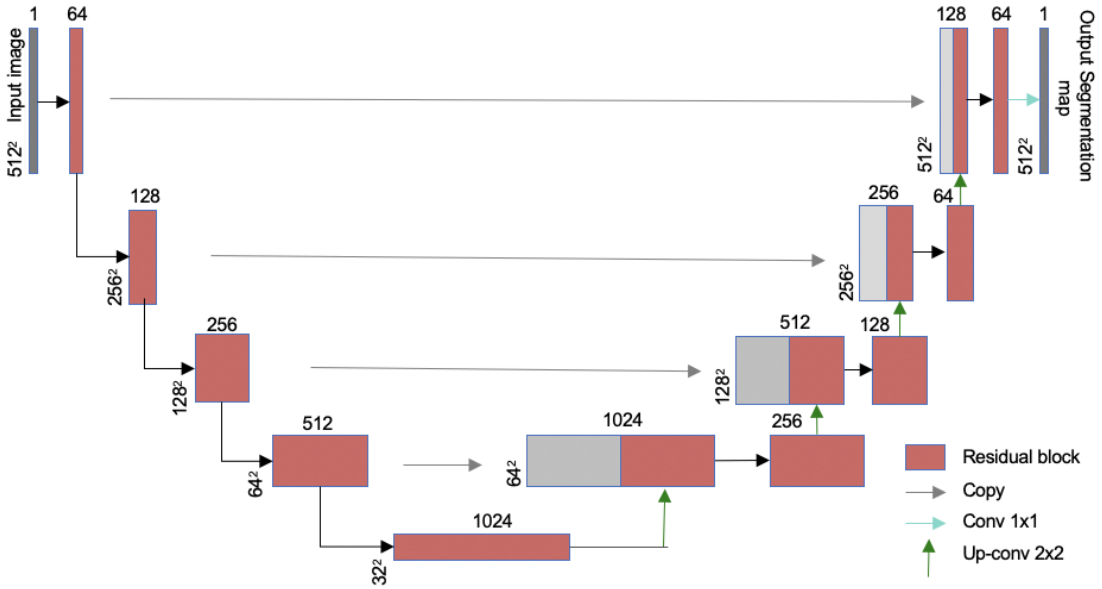


Figure 5.5: ResUNet architecture. Input size is written on the side of each box. The number of feature maps in each residual layer is written on top of each box.

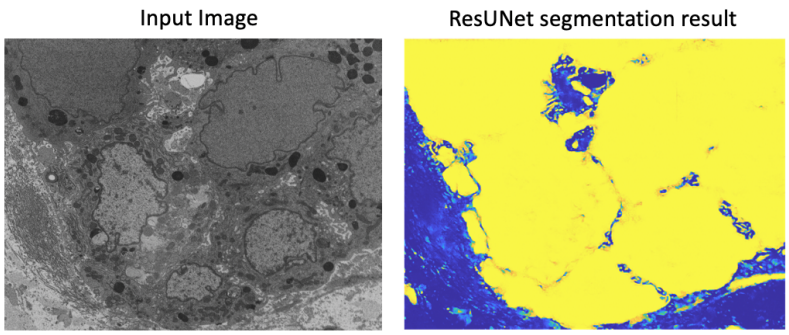


Figure 5.6: Cell segmentation using ResUNet. All cells are grouped into one big blob.

decoding path, followed by a 2x2 convolution to half the number of feature maps. These feature maps are then concatenated with the feature maps at the same level in the encoding path through a skip connection. This concatenation step helps combine the deep, semantic, coarse-grained feature maps from decoder path with high resolution feature maps from encoder path enabling effective recovery of fine-grained details. These concatenated feature maps are then passed through the residual block. The output of the last residual block is passed through a 1x1 convolutional layer followed by sigmoid activation to provide the final segmentation mask. We compare the segmentation results of ResUNet with TransUNet and SETR described in Section 2.2. Leveraging the similarities of structures contained in a 3D EM image stack, the neural network is trained on a small subset of manually labeled 2D images evenly distributed in the 3D EM image stack to efficiently segment the rest of the images in the stack.

As discussed earlier, the cell organelles generally don't appear to be attached and also have a clear boundary around them, therefore they can be segmented with semantic segmentation networks like ResUNet. But cancer cells in electron microscopy images do not have clear boundaries separating them. ResUNet segments all cells together as one big blob as shown in Figure 5.6. Also, cancer cells have filopodia-like protrusions which are thin long finger-like protrusions from the cell membrane that act like antennas to probe the surrounding environment (Figure 5.2). While capturing the FIB-SEM image if the filopodia-like protrusions are cut perpendicular to their length they appear as island-like blobs detached from the main cell as seen in Figure 5.2. In order to capture the main body and the filopodia-like protrusions we propose a multi-pronged approach combining segmentation, propagation and tracking strategies for cell segmentation as shown in Figure 5.7.

Instead of training the ResUNet to segment only the cell interior region, we train the network to segment the cell boundaries along with the cell interior region. To do so we provide two

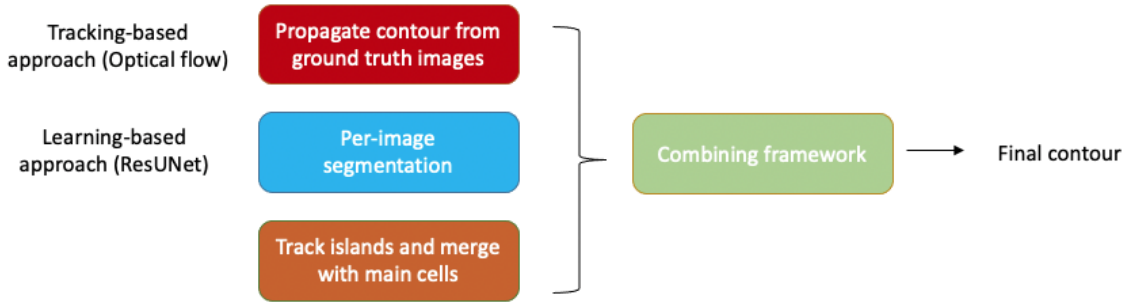


Figure 5.7: Proposed multi-pronged approach for cell segmentation

ground truth maps to the network, one containing the cell masks and another containing the cell boundaries. The cell mask map is an image with all pixels representing cell interior region marked as 1. Thus, using ResUNet we segment the cell interior, cell boundaries and the protrusions that appear as island-like blobs (as they are also marked as cell interior region). However, the boundary segmentation from the neural network alone could not separate adjacent cells with similar intensity and texture variations. In order to obtain a precise separation of adjacent cells with similar appearances, the boundary information from neural network is combined with the boundary propagated using optical flow from the small subset of manually labeled 2D images evenly distributed in the 3D EM image stack that are used to train the ResUNet.

As discussed in Section 2.4.1, optical flow gives the flow vectors representing the apparent motion of individual pixels between two images. We use the Farneback algorithm as it computes a dense optical flow - a flow vector for each pixel [52]. Though it is computationally slower, it produces more accurate results. The Farneback algorithm generates an image pyramid to estimate displacement at multi-scales, starting at a coarser level and refining the estimate on finer levels. The pyramid decomposition enables the algorithm to handle both large and small pixel motions. We experiment with the parameter setting for the Farneback algorithm. The settings that work best for our images are number of pyramid levels = 6, neighborhood size = 5, filter size = 30, pyramid scale = 0.2, number of iterations = 3.

We now have boundary information from optical flow and ResUNet. We experiment with 3 ways of using the boundary information in separating the cells. First we directly overlay the optical flow boundaries on the cell interior mask prediction from ResUNet. Second, we selectively include cell boundaries propagated using optical flow only in regions with overlapping cells. In the analysis so far, propagation of cell boundaries using optical flow was performed as a separate task independent of the estimate from the segmentation model. In order to obtain a better continuity of

the cell boundary, instead of propagating the cell boundaries from the manually labeled image to all the following images, we propagate it only onto the image immediately following the manually labeled image. We then use method two to combine it with boundary from ResUNet output of that image to obtain its cell segmentation. Now optical flow is applied to these newly estimated cell boundaries which have the combined information from ResUNet and optical flow. This is the third method of using the boundary information of separate cells. Once we combine the boundary information using one of these methods, it is then overlaid on the cell-interior mask segmented using ResUNet. Individual cells are then separated from the cell-interior mask by performing watershed (Section 2.2) using centroids of cells in the nearest manually labeled image as seeds.

Now for every image in the 3D FIB-SEM stack we have individual cells and the island like blobs segmented. Each isolated region is given a unique label. The next step is to track each of these regions across images and associate the protrusions to their corresponding cells. We use an overlap-based label propagation technique to obtain the cell associations. Intersection over union (IoU) is a metric that measures the overlap between two regions. Here, the IoU metric given in 5.1 is used to track isolated region across images.

$$\text{IoU} = \frac{\text{Area of overlap}}{\text{Area of union}} \tag{5.1}$$

In the first image in the stack, all isolated regions are given unique tracking labels. For each of the isolated regions in the next image, IoU is calculated against all the labeled regions in the previous image. Each isolated region gets assigned the tracking label associated with the region it has maximum IoU value with in the previous image. If there is no overlap and a new region is detected, a new tracking label is assigned to that region which can then be tracked in subsequent frames. The individual main cell bodies of the cells in EM image are very big. We first run the tracking algorithm on just the individual main cells by using the cell ids in the manually labeled images as the tracking labels. Once we have the main cell body tracked across all the images, we track the island like blobs to these cells. As the blobs are tracked across images, when they meet the main big cell, their label gets modified to that of the big cell and therefore all the regions associated with that blob (now a protrusion from a cell) in the previous slices will also have their label modified to that of the main cell. When a protrusion breaks off from a cell it still retains the label of that cell. As there could be hundreds of blobs in one image, comparing each blob with every other blob in the previous image would take a long time. As the movement of regions between two successive images is small, the corresponding region if present in the next image would be in the vicinity of the region in the previous image and not in some other far off region of the image. Therefore, instead of searching for overlapping regions all across the 6000×4000 image,

we look for a region matching an existing track label in its local neighborhood of 512×512.

Once all cells and the organelles are segmented they can be rendered to visualize them in 3D and quantitative image features can be extracted for further analysis.

5.5 Methods

5.5.1 3D Focused Ion Beam-Scanning Electron Microscopy Dataset Collection

Under an institutional review board approved observational study, three tissue biopsy samples (Bx1, Bx2 and Bx4) were acquired over three time points of cancer treatment from a patient with metastatic ER+ breast ductal carcinoma and two biopsy sample (PTT, PDAC) were acquired from two patients with pancreatic ductal adenocarcinoma at Oregon Health and Science University, Portland. The last sample is a micro-spheroid prepared using a breast cancer cell line (MCF7). Extensive additional information regarding the three biopsies from the patient with metastatic breast cancer are available [89]. The samples were preserved in Karnovsky's fixative (2.0% PFA, 2.5% Gluteraldehyde), post-fixed using an OsO4-TCH-OsO4 staining protocol, and embedded in Epon resin. Post-fixation staining binds heavy metals to lipid-rich membranes to provide contrast in EM imaging. Conductive coating with 8nm-thick carbon was necessary to achieve high-resolution, charge-free, high contrast, and low noise images. A FEI Helios NanoLab 660 DualBeam^{TM} FIB-SEM was used to collect high resolution 3D volumes of the resin-embedded blocks. Targeted volumes were collected by using a Ga+ FIB-source to sequentially slice a few nanometers from the sample to expose a slightly fresh surface for subsequent imaging. The slicing/imaging cycle was automated using the FEI AutoSlice and View^{TM} software extended package, while image collection during 3D data acquisition used the in-column detector. Metastatic breast cancer, primary pancreatic tissues and the micro-spheroid were imaged with an isotropic resolution of 4 nm, 6 nm and 6nm respectively.

5.5.2 Image preprocessing and ground truth generation

After data acquisition, images within the stack are translationally aligned in the XY-plane using an in-house stochastic version of TurboReg affine transformation [166]. The alignment step zero-pads the images in order to maintain a uniform size, which are subsequently cropped to yield the final 3D image volumes. The registration and edge cropping process yields a final resolution of $5634 \times 1912 \times 757$ for the Bx1, $5728 \times 3511 \times 2526$ for Bx2, $5990 \times 3812 \times 1884$ for Bx4, $6065 \times 3976 \times 792$ for the PTT, $6114 \times 3874 \times 1583$ for PDAC and $6083 \times 3740 \times 2208$ for MCF7.

The third dimension refers to the number of slices in each stack. Intermittently, the brightness of a few images in the stack varied, increasing the complexity of the images and making segmentation more challenging. Histogram equalization was applied to ensure consistency across the stack and reduce complexity of the images.

Table 5.1: The availability of ground truth labels over different datasets. ✓✓denotes all images labelled, ✓denotes images sparsely labelled and a blank denotes no images labelled.

Type	Dataset	Cell	Nucleus	Nucleolus	Mitochondria	Lysosomes	Endosomes
Breast cancer	Bx1	✓✓	✓✓	✓✓	✓✓		
	Bx2	✓✓	✓✓	✓✓	✓✓		
	Bx4		✓	✓			
Pancreatic cancer	PTT		✓✓	✓✓			
	PDAC	✓✓	✓	✓	✓	✓	✓
Cultured breast cancer cells	MCF7	✓	✓	✓	✓	✓	✓

The cells and organelles were manually labeled using microscopy image browser [20]. The availability of ground truth labels over different datasets is shown in Figure 5.1. Nucleus and nucleolus are labelled on all slices of PTT, Bx1 and Bx2 dataset and only 19, 16 and 24 slices of Bx4, PDAC and MCF7 datasets respectively. Mitochondria are labelled on all slices of Bx1 and Bx2 datasets and only 13 and 10 slices of PDAC and MCF7 datasets. Only PDAC dataset has all cells labelled in all slices. Bx1, Bx2 and MCF7 datasets have all cells labeled in only on 10, 23 and 22 slices respectively. For Bx1 and Bx2 datasets 6 and 11 cells are selected respectively and labeled completely across all the slices in the datasets. Lysosome and endosomes are labelled on 10-25 images for different dataset. Sparse labeling of nuclei and nucleoli takes 1-2 days, and mitochondria, lysosomes, endosomes and cells about 5-10 days each. Manual labeling of nucleus and nucleolus on all images of PTT, Bx1, and Bx2 datasets required roughly 50-80 hours each. Manual labeling of mitochondria and cells on all images of a dataset took around 8 months each.

5.5.3 Training data

The proposed ResUNet architecture is used to segment each 3D FIB-SEM stack by training on a small subset of manually labeled images and predicting on the remaining images. To estimate the number of manually labeled images required for effective segmentation, the size of the training

set is varied by selecting 7, 10, 15, 25 or 50 images evenly distributed across the dataset. The full resolution of the EM images cannot be analyzed directly by the network due to the memory restrictions of the GPU hardware. Therefore, image crop size of 512×512 is chosen in order to fit a meaningfully large number of images in a batch. We use a batch size of five in all our experiments, meaning that five image tiles are used to estimate the error gradient at each iteration of the network weight update. Using multiple training examples in a batch for error gradient estimation helps make the training process more stable. We choose a batch size of five as it is the maximum number of tiles that could fit in the GPU memory in our case. The nuclei occupied 15-35% of the pixels in each image and random selection of five 512×512 tiles per image ensured there was enough representation of nucleus in a batch. Mitochondria, endosomes, lysosomes and cells also had good representation. However, nucleoli are smaller and sparser, occupying less than 1% of a full image. As a result, if sampled randomly, a large number of tiles would need to be considered in order to form a batch with enough nucleoli representation, which would exceed the GPU memory. To ensure that the model encounters sufficient representation of nucleoli during training, the batch size is still maintained at five but the tiles are selected such that at least 4 out of the 5 randomly selected tiles contain no less than 100 pixels related to nucleolus. Horizontal and vertical flips are randomly applied as data augmentation steps.

5.5.4 Implementational details

The proposed network is implemented using Keras framework [69] with TensorFlow [1] as backend. The network is optimized by adaptive moment estimation (Adam) with 10^{-4} learning rate, exponential decay rates for moment estimates $\beta 1 = 0.9$, $\beta 2 = 0.999$, and epsilon = 10^{-7}. It is trained to minimize the Dice loss function for $5,000$ weight updates. The experiments were performed on a single NVIDIA Tesla P100 GPU.

5.5.5 Inference Methodology

As the size of an EM image is much larger than the 512×512 crop size processed by the network, each image is parsed into multiple overlapping tiles. Each tile is passed through the network to predict a probability for every pixel, and pixels with value greater than or equal to 0.5 are assigned to the foreground class. The resulting overlapping segmentation maps are blended by multiplying each map with a 2-dimensional tapered cosine window and adding the result to reduce the edge artifact at tile borders. If we do not perform any blending or averaging then the resulting overlapping segmentation maps often present some step effect like artifact at the tile borders.

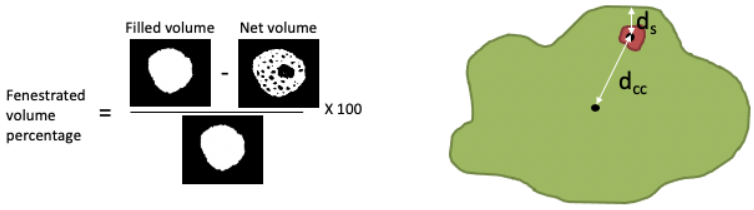


Figure 5.8: Illustration of net volume and filled volume used in the fenestrated volume percentage measure shown on images instead of volumes and distances d_{cc} and d_s used in proximity of nucleolus to the nuclear membrane measure.

5.6 Quantification and Statistical Analysis

5.6.1 Evaluation Metrics

The segmentation is evaluated using three metrics: Dice coefficient, recall and precision as defined in Section 4.4.4.

5.6.2 Morphological and Texture Features

Segmentation of organelles allows us to characterize biologically relevant features such as their morphology and texture. Morphology refers to the structure of the organelle. The morphological measures are designed to capture the size and shape, and include features such as solidity, sphericity, circular variance. The texture features are designed to capture the spatial distribution of intensity patterns, and include features from GLCM, SZM and power spectrum. The mathematical definitions of these features are provided in Section 2.3.

In addition to the standard morphological features, we design two new features to capture properties of nucleolus in cancer cells. The nucleolus in cancer cells are characterized by formation of fenestrations and movement towards the cell membrane. Fenestrations refer to all the cavities inside the nucleolus. We design two measures - percentage of volume fenestrated, and proximity of nucleolus to nuclear membrane - to capture the above mentioned properties.

The percentage of fenestrated volume in the nucleoli is calculated as the ratio of the difference of filled in volume and net volume to the filled in volume. The volume is filled in by performing a fill holes operation as shown in Figure 5.8. This measure is different from solidity as it measures the volume of cavities within the nucleolus, while solidity measures concavities in the outer surface.

$$\text{Fenestrated volume percentage} = \frac{\text{Filled volume} - \text{Net volume}}{\text{Filled volume}} \tag{5.2}$$

If d_{cc} is the distance between centroids of the nucleus and nucleolus and d_s is the distance between the centroid of nucleolus and the nearest point on the surface of the object, the proximity of

nucleolus to nuclear membrane measure is calculated as the following.

$$\text{Proximity to nuclear membrane} = \frac{d_{cc}}{d_{cc} + d_s} \tag{5.3}$$

The distances d_{cc} and d_s are visualized in Figure 5.8. It takes a value of 0 when the nucleolus is at the center of the nucleus and increases as it moves towards the nuclear membrane.

5.7 Results

5.7.1 Model Training Setup Evaluated

We apply the proposed segmentation approach on the six datasets comprising tissues from breast and prostate cancer and a microspheroid prepared using a breast cancer cell line (MCF7). Quantitative evaluations are performed only on features that are manually labeled on all images in the dataset. The availability of ground truth labels over different datasets is shown in Table 5.1.

Overall we perform quantitative evaluation of cell, nuclei and nucleoli segmentation on 3 datasets and mitochondria segmentation on 2 datasets. Few organelles in certain datasets are labeled only on few images which are used for training the segmentation model. In such cases we present only qualitative results. For each dataset, a segmentation model is trained using a small subset of manually labeled images and the trained model is used to generate segmentation masks on the rest of the unlabeled images. We experiment with convolution and attention-based models for semantic segmentation.

5.7.2 Deep Learning Models can Accurately Segment Intra-cellular Organelles

We primarily use the convolution-based ResUNet discussed in 5.4 as the segmentation model. To train the segmentation model, a sparse subset of training images (typically between 7 to 50 slices) was randomly selected from the 3D image stack and used to tune the network weights, by minimizing the difference between the ground truth segmentation and the network generated segmentation on these training slices. Once trained, the network is used to segment the rest of the slices in the stack that were not used for training, and performance in terms of Dice score, precision and recall are calculated. We randomly select the input images ten times, and the mean performance metrics over the ten runs was reported. Figure 5.10B shows the performance metrics for nuclei and nucleoli segmentation in PTT, Bx1 and Bx2 datasets (nuclei results in the top row, and nucleoli results in the bottom row). Depending on the datasets and the training set size, overall Dice scores of 0.95-0.99 and 0.70-0.98 are achieved for nuclei and nucleoli segmentation, respectively.

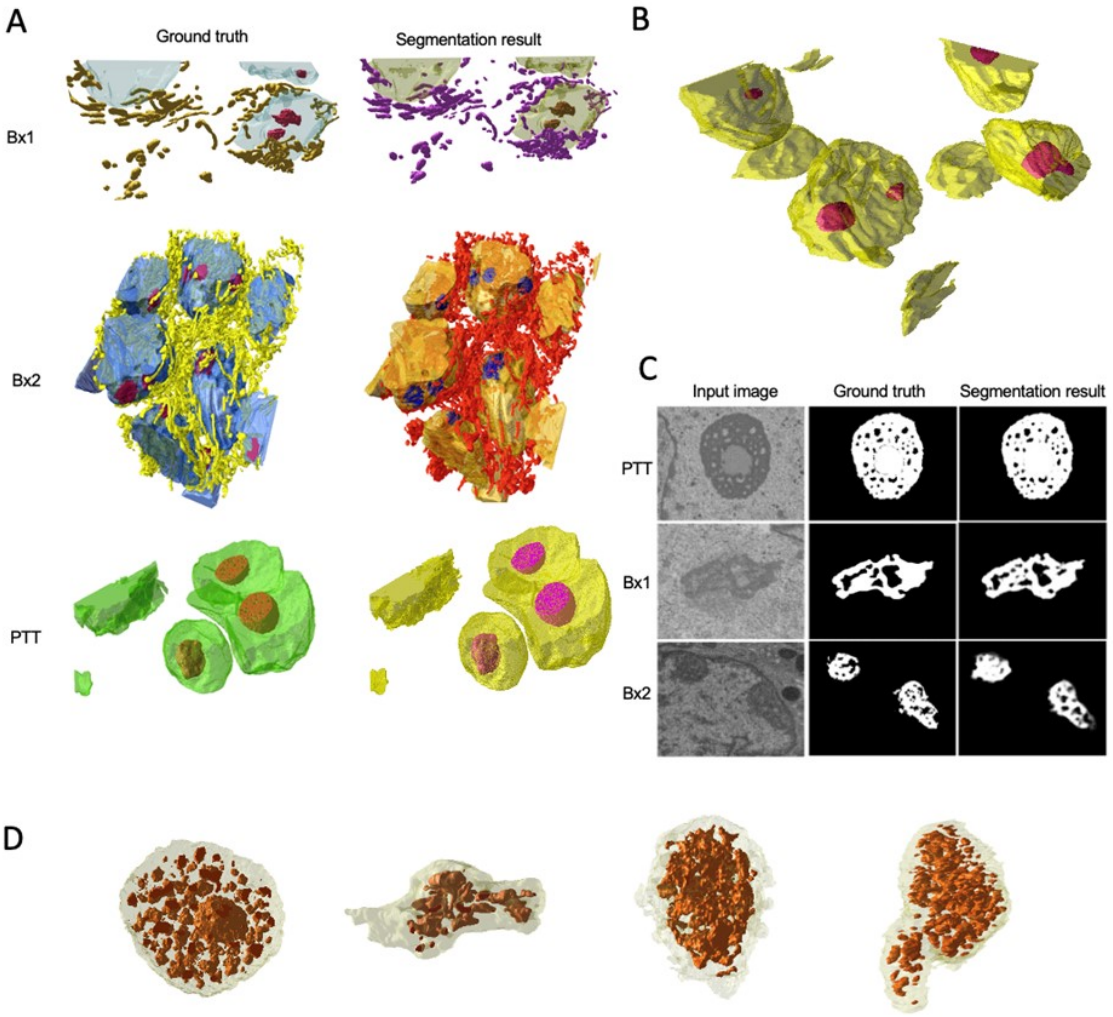


Figure 5.9: Nuclei and nucleoli volume renderings and nucleoli segmentation results. (A) Volume renderings showing the 3D structure of ground truth masks and predicted segmentation masks for PTT, Bx1 and Bx2 datasets (B) Volume rendering showing the 3D structure of predicted segmentation masks for Bx4 dataset. (C) Representative qualitative results showing input images (first column), ground truth (second column) and predicted nucleoli (third column) for PTT (first row), Bx1 (second row) and Bx2 (third row) datasets. (D) Volume renderings of the fenestrations in nucleoli of PTT, Bx1, Bx2 and Bx4 datasets. (E) and (G) Volume renderings showing the FIB-SEM image stack and (F) and (H) the predicted segmentation masks of cells and organelles for the PDAC and MCF7 datasets respectively.

In addition to the nucleoli boundary, the model accurately segments the fenestrations within the nucleoli (Figure 5.9C). Dice score of 0.86 and 0.76 are obtained in segmenting mitochondria from Bx1 and Bx2 datasets respectively. Segmentation performances of endosomes and lysosomes are evaluated qualitatively (Figure 5.12) as they are labeled only on a few slices that are used to train the segmentation model. Figure 5.9A displays the comparison between volumes rendered from manual annotations and those from predicted segmentation, with nucleoli nested inside the nuclei for PTT, Bx1 and Bx2 datasets. Figure 5.9B only displays the volume rendered from predicted segmentation for Bx4 dataset as it does not have ground truth labels for the entire stack. Figure 5.9D displays the volume rendering of the fenestrations inside nucleolus in all six datasets. 2D image slices overlaid with ground truth and predicted segmentation for PTT and Bx4 datasets are shown in Figure 5.11 and for Bx1, Bx2, PDAC and MCF7 datasets are shown in Figure 5.12. They illustrate the accuracy of model segmentation compared to the ground truth annotation.

5.7.3 Larger Context Improves Segmentation Performances

The cancer cells in EM images are relatively large. A 512×512 tile typically includes only a small section of the cell (for example a small part of the nucleus), and therefore, does not contain enough spatial context. We hypothesize that providing better contextual information regarding the surroundings of the organelles could lead to improved segmentation. In order to incorporate more global context while subjected to GPU memory restrictions, we extract tiles of size 2048×2048 and down sample them to 512×512 and train segmentation models on these down sampled tiles. We also experimented with 1024×1024 and 3072×3072 sized tiles, but found that 2048×2048 tiles produced superior segmentation results. We also compare the segmentation performance of fully-convolutional ResUNet with the more recent TransUNet and SETR where the global context is captured by transformers with self-attention mechanism for segmentation of nuclei.

Providing larger context during training improves all three metrics in nuclei (Figure 5.10A, top row) and nucleoli (Figure 5.10A, bottom row) segmentation. The information provided by the larger context seems to outweigh any information loss due to down sampling. Further, using larger tile size enables faster processing of an entire EM image as fewer tiles would be required to reconstruct the final result. Performance of TransUNet and SETR are on par or slightly lower than ResUNet and did not provide any added advantage in segmentation as shown in Figure 5.13. As experimented in Chapter 4, we tried using attention mechanism, deep-supervision, Tversky, focal and also dual-Dice loss functions but did not observe considerable improvement in segmentation performance.

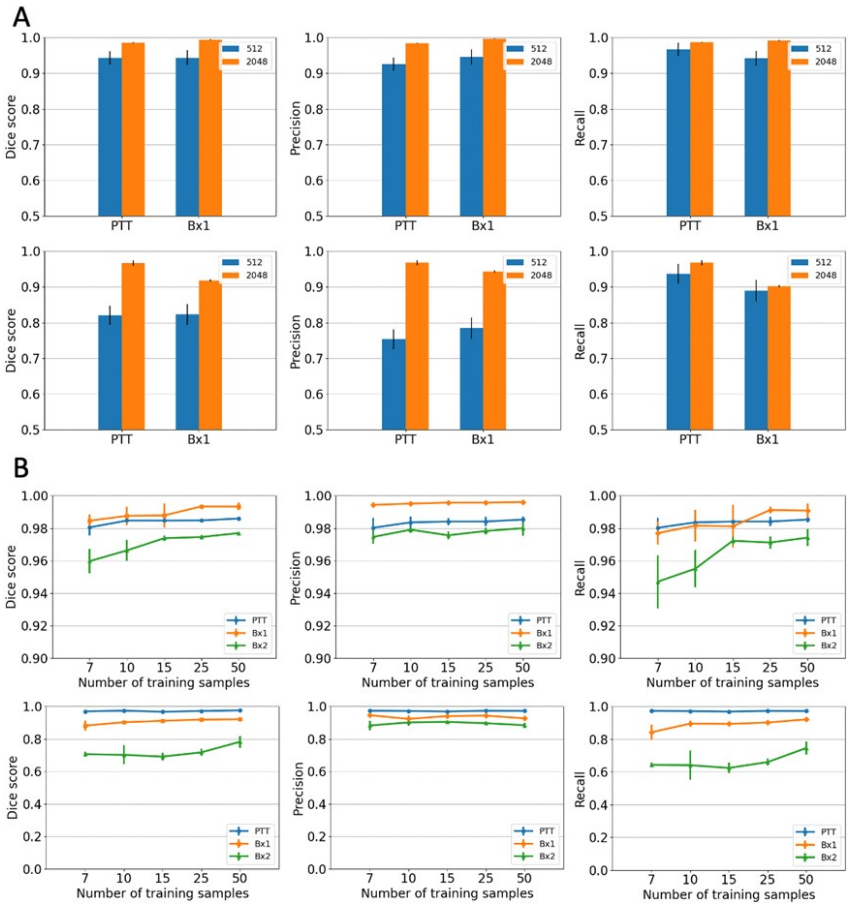


Figure 5.10: Nuclei and nucleoli segmentation performance. (A) Effect of image tile size (context window). Segmentation performances for nuclei (top row) and nucleoli (bottom row) using different input tile sizes measured by Dice score (first column), precision (second column) and recall (third column) on the Bx1 and PTT datasets. The blue bar represents the results of training the network directly on the 512x512 sized tiles from the FIB-SEM images. The orange bar represents the results of training the network on image tiles of size 2048x2048 down-sampled to 512x512, which provide larger contextual information. Each error bar represents 10 separate experiments in which a new model was trained from scratch using the specified number of training images. (B) Effect of training set size. Segmentation performances for nuclei (top row) and nucleoli (bottom row) using different training set sizes, measured by Dice score (first column), precision (second column) and recall (third column) on PTT (blue), Bx1 (orange) and Bx2 (green) datasets. For each dataset, performance was evaluated over training set sizes of [7,10,15,25,50] in order to find the minimum number of images required to generate accurate segmentation. Each error bar represents the mean and standard deviation of results obtained from 10 separate experiments in which a new model was trained from scratch using the specified number of training images.

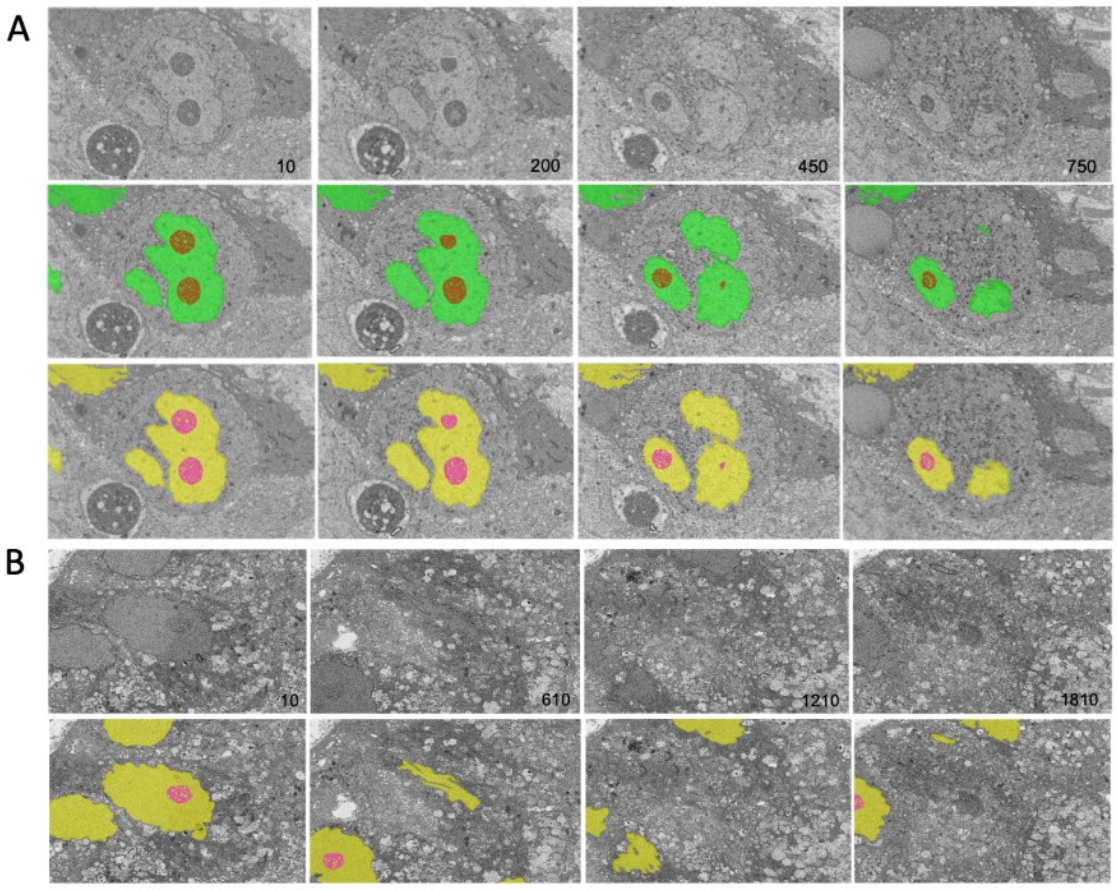


Figure 5.11: Nuclei and nucleoli segmentation results on PTT and Bx4 datasets. Representative qualitative results showing input images (top row) overlaid with nuclei (green) and nucleoli (red) ground truth masks (middle row) and predicted nuclei (yellow) and nucleoli (pink) segmentation masks (bottom row) for (A) PTT and (B) Bx1 datasets. Numbers in the lower right-hand corner of images indicate the slice position of the image in the full image stack.

5.7.4 Small Training Set is Adequate and Larger Training Set Brings Further Improvement

We experiment with varying the number of manually labeled full images used for training by selecting 7, 10, 15, 25 or 50 training images evenly distributed across the image stack. The models are trained on 2048×2048 sized tiles down sampled to 512×512 pixels. The model performs well with just 7 training images, with an overall dice score of 0.95-0.99 for nuclei (Figure 5.10B, top row), and a dice score of 0.70-0.98 for nucleoli (Figure 5.10B, bottom row), depending on the data sets. The performance continues to improve with more training images for all three metrics for both nuclei and nucleoli, reaching a dice score range of 0.97-0.99 for nuclei and 0.80-0.99 for

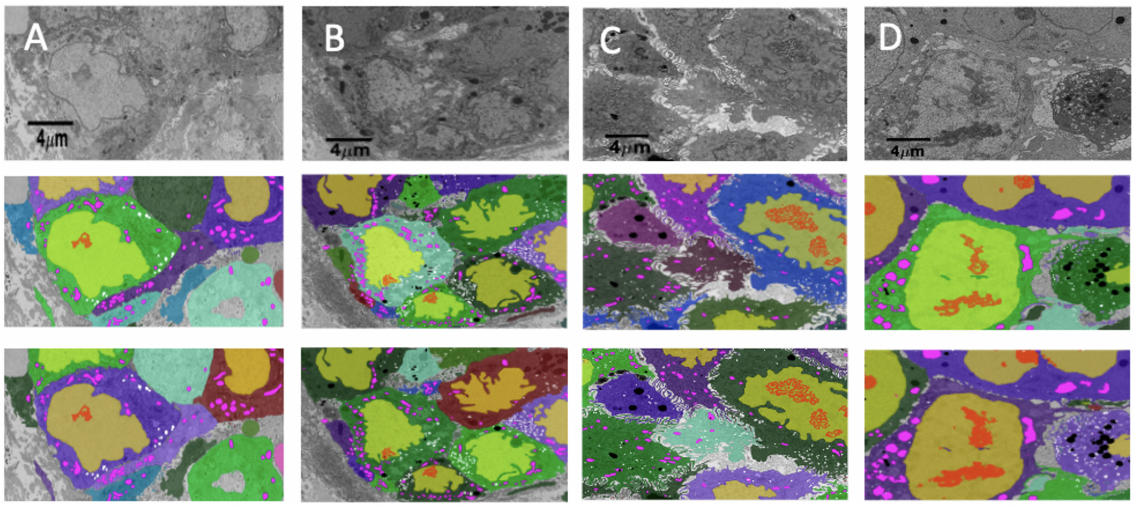


Figure 5.12: Cell and organelle segmentation results on Bx1, Bx2, PDAC and MCF7 datasets. Qualitative results showing input images (first row) overlaid with nuclei (yellow), nucleoli (red), mitochondria (pink), endosomes (white), lysosomes (black) and cell segmentation (random colors) on ground truth masks (second row) and predicted segmentation masks (third row) for (A) Bx1 (B) Bx2, (C) PDAC, and (D) MCF7 datasets.

nucleoli with 50 training images. Even with 50 training images, it is still a very small training set, representing only 2-5% of all the images in the stack. This illustrates that it is possible to train a reliable model with sparse manual labeling for 3D EM segmentation. The performances for the Bx2 dataset are lower than the other two datasets as it is a larger image stack exhibiting significant variability among the nuclei and nucleoli within the dataset.

5.7.5 Generalizability of the Segmentation Model

In the current setting, we train a separate model for each organelle in each dataset. We experiment training a model to segment all organelles of a dataset at once. Qualitatively the results appeared very similar to the results from individually trained networks as seen in Figure 5.14. We compare the segmentations of nuclei and mitochondria quantitatively in Bx1 and Bx2 datasets (Table 5.2). In Bx1 dataset, the segmentation performance is similar to single organelle segmentation networks. But in Bx2 dataset, the recall remains the same while the precision decreases. This could be due to the larger intra-organelle variability in the Bx2 dataset which could make it harder for the segmentation model to eliminate the patterns similar to the target organelle simultaneously for all organelles and therefore introduces few false positives. Nucleoli cannot be segmented by a network trained for all organelles together as they occupy a very small portion of the image and a random sample of crops does not provide enough representative examples. Therefore, nucleolus

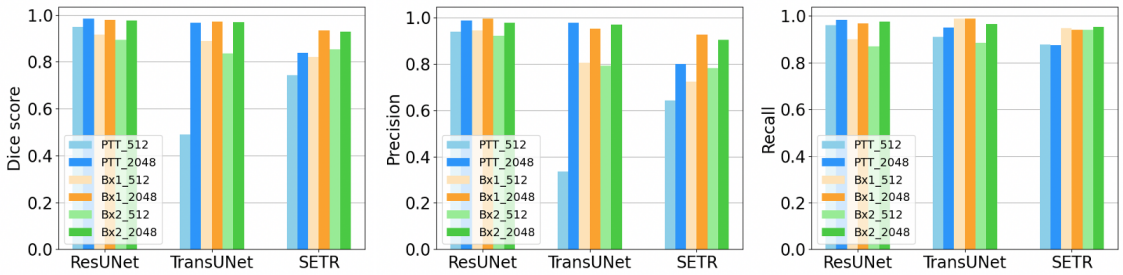


Figure 5.13: Nuclei segmentation performance using ResUNet, TransUNet and SETR. Segmentation performances for nuclei using different input tile sizes measured by Dice score (first column), precision (second column) and recall (third column) on the PTT, Bx1 and Bx2 datasets. The lighter bar represents the results of training the network directly on the 512x512 sized tiles from the FIB-SEM images. The darker bar represents the results of training the network on image tiles of size 2048x2048 down-sampled to 512x512, which provide larger contextual information.

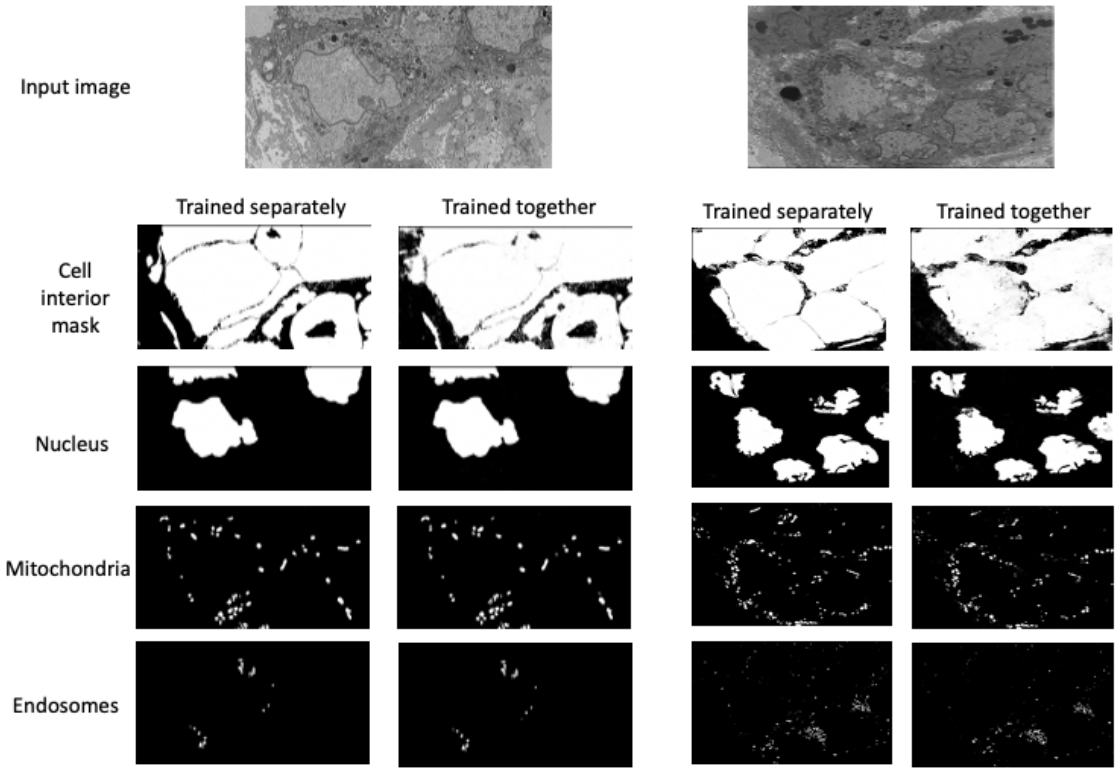


Figure 5.14: Segmentation results on multiple organelles in Bx1 and Bx2 datasets by models trained on each organelle separately and by a model trained all organelles together.

Table 5.2: Segmentation performance for nuclei and mitochondria in Bx1 and Bx2 datasets using models trained individually on each organelle vs model trained on multiple organelles together.

Dataset	Organelle	Models trained individually			Models trained all together		
		Dice	Precision	Recall	Dice	Precision	Recall
Bx1	Nucleus	0.98	0.96	0.99	0.98	0.96	0.99
	Mitochondria	0.86	0.85	0.86	0.85	0.84	0.87
Bx2	Nucleus	0.98	0.98	0.98	0.94	0.90	0.98
	Mitochondria	0.76	0.78	0.75	0.70	0.64	0.78

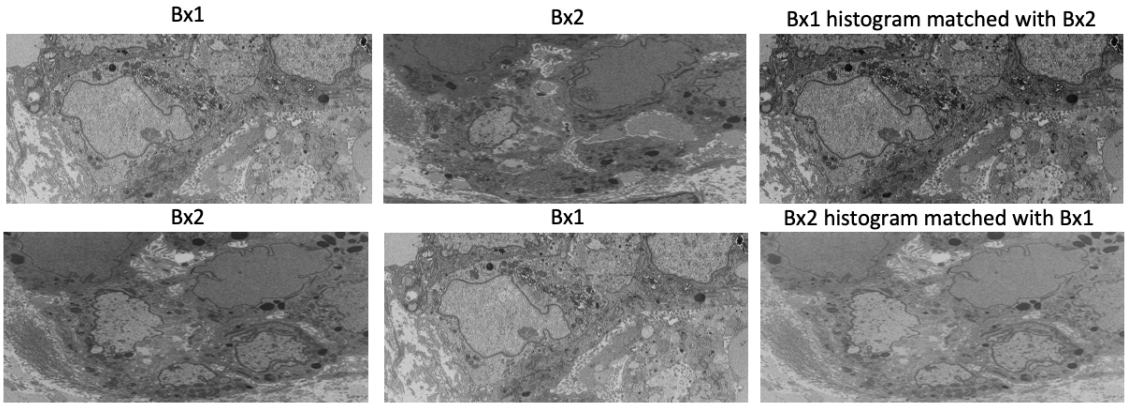


Figure 5.15: Results of Histogram matching. The first column represents the target image, second row represents the source image (source of the histogram) and the third row represents the histogram matched target image

segmentation requires training of individual segmentation models while other organelles could be trained together in a single model.

Due to intensity variations between datasets we observe that a model trained on one dataset does not generalize well on other datasets. We try three methods to adapt models trained on one dataset to produce good results on other datasets. First, we perform simple histogram matching by adjusting the histogram of the images of a dataset to match those of the images the model was trained on. Second, we try to match the style of the images using CycleGAN [202]. The CycleGAN is an approach of training a neural network to translate an image from a source domain to a target domain in the absence of paired examples. Third, we finetune the model with few labeled images. We evaluate the performance metrics for these methods on nuclei segmentation in Bx1 and Bx2 datasets. Overall, we observe that a model trained on Bx2 could segment Bx1 images with the

Table 5.3: Nuclei segmentation performance in Bx1 and Bx2 datasets using model trained on one dataset to predict on images from the other dataset

	Measure	Train from scratch	Predict on unseen dataset	Histogram adjustment	Style transfer using cycleGAN	Finetune	
						1 image	3 images
Trained on Bx2, Predicted on Bx1	Dice	0.98	0.48	0.80	0.67	0.96	0.97
	Precision	0.96	0.94	0.68	0.58	0.96	0.95
	Recall	0.99	0.34	0.97	0.80	0.97	0.98
Trained on Bx1, Predicted on Bx2	Dice	0.98	0.13	0.37	0.45	0.51	0.95
	Precision	0.98	0.56	0.29	0.72	0.95	0.96
	Recall	0.98	0.09	0.53	0.35	0.36	0.94

help of the above mentioned methods better than the Bx1 model on Bx2 images (Table 5.3). This is because Bx2 dataset has a larger variability among the nuclei and contained representative examples of different presentations of the nuclei and therefore could generalize better. As Bx1 dataset has only three nuclei with less variability its training examples are not rich enough to capture the variability of nuclei in Bx2 datasets and therefore could not segment some nuclei no matter which adaptation method is employed. Simple histogram adjustment of Bx1 images to Bx2 images results in a recall of 0.97 while predicting nucleus in Bx1 dataset with a model trained on Bx2 dataset. Histogram matched images for Bx1 and Bx2 datasets are shown in Figure 5.15 and segmentation results on these images in shown in Figure 5.16. Style transfer using CycleGAN modifies the content of the image while trying to match the styles as seen in Figure 5.17, which results in poorer segmentation performance as seen in Figures 5.18 and 5.19. Finetuning Bx2 model with one image from Bx1 dataset produces results comparative to a Bx1 model trained with 7 images. Though finetuning Bx1 model with one Bx2 image increased the Dice score when compared to histogram adjustment and CycleGAN style transfer it is still low as it is not representative of the variability of the dataset. Finetuning with 3 images results in good segmentation performance for both datasets. But we observe that a segmentation model trained from scratch with 3 images for each dataset produces results similar to those of models pretrained with images from one dataset and finetuned with 3 images from other dataset. We therefore feel that if we can build a model with data containing representive examples of the variability of an organelle it can generalize to new datasets with simple techniques like histogram adjustment. Further if we can build a model from

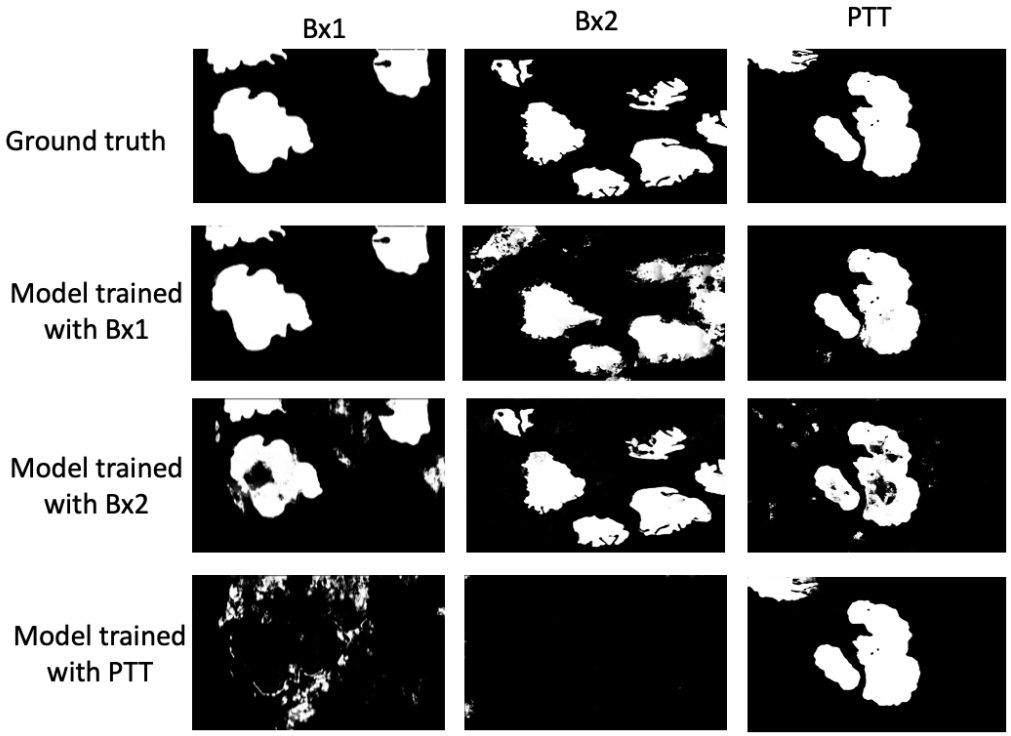


Figure 5.16: Representative segmentation results on histogram adjusted images. The first row represents ground truth images and second third and fourth row represent segmentation results using models trained on Bx1, Bx2 and PTT datasets respectively. The first, second and third columns represent the Bx1, Bx2 and PTT images. If an image is not used to train the model it is histogram adjusted to the images of that model.

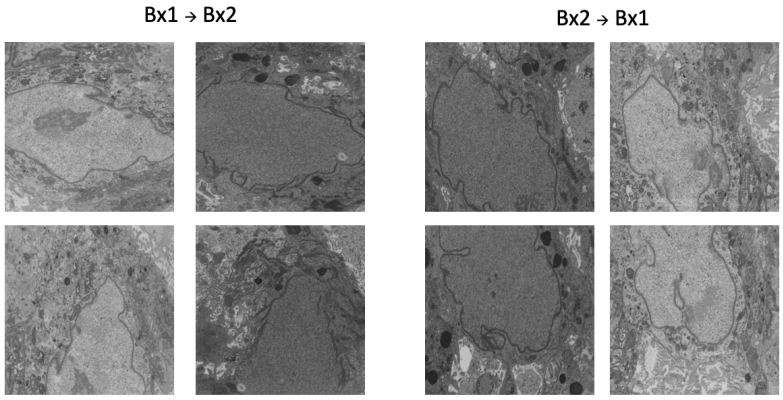


Figure 5.17: Results of CycleGAN style transfer. The first column represents images from Bx1 dataset. The second column is the style transferred version of the first column image. It is transferred to the style of Bx2 dataset. Similarly, the third column represents images from Bx2 dataset and the fourth column is the style transferred version of the third column image. It is transferred to the style of Bx1 dataset.

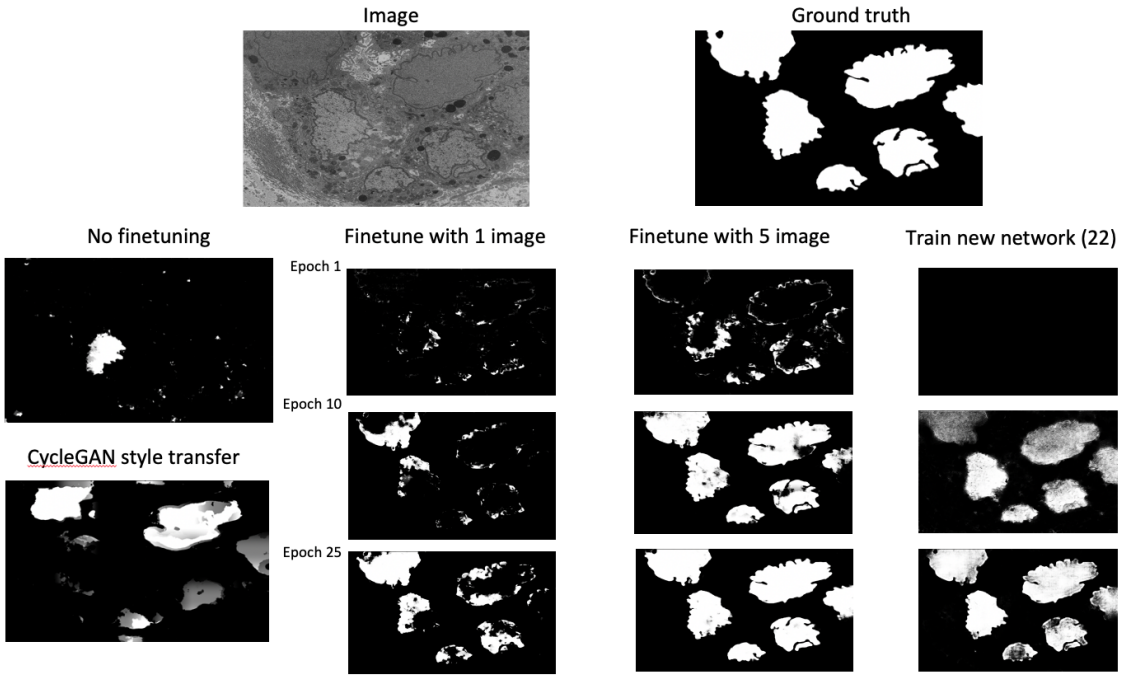


Figure 5.18: Representative segmentation results on Bx2 dataset using a model trained on Bx1 dataset using style transfer and fine tuning approaches

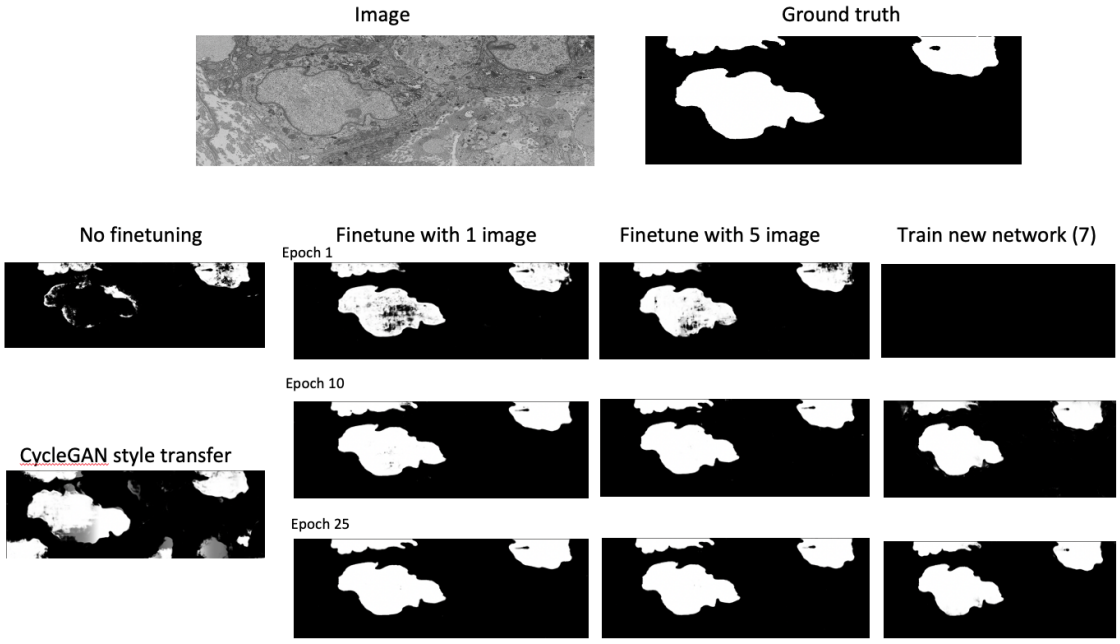


Figure 5.19: Representative segmentation results on Bx1 dataset using a model trained on Bx2 dataset using style transfer and fine tuning approaches

multiple datasets with different acquisition parameters resulting in different intensity distributions the model will be robust and could probably directly predict on new datasets with good accuracy.

5.7.6 Cell Segmentation

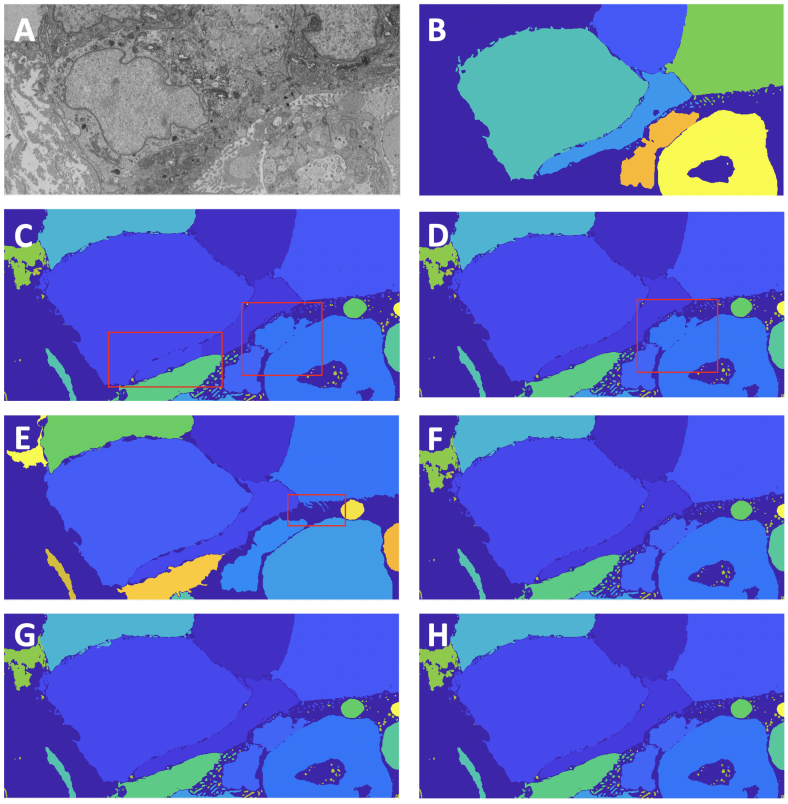


Figure 5.20: Cell segmentation results on Bx1 dataset. Representative qualitative results showing (A) input image, (B) ground truth segmentation (few cells are not labeled in the ground truth), (C) segmentation of cells using cell-interior mask alone, (D) segmentation of cells using cell-interior mask and boundary predictions from segmentation model, (E) segmentation of cells using optical flow alone, (F) segmentation of cells by overlaying boundaries propagated using optical flow on cell-interior mask obtained from segmentation model to separate cells, (G) segmentation of cells by selectively combining optical flow boundary estimate with the boundary estimated from segmentation model and (H) segmentation of cells by propagating boundaries estimated in the previous frame using optical flow and combining with the boundary estimated from segmentation model for Bx1 dataset

A multi-pronged approach combining propagation, segmentation and tracking strategies is used for cell segmentation, as described in Section 5.4. Optical flow is the propagation-based method used to propagate the cell boundaries from the nearest labeled ground truth image. We observe that as we moved farther from the ground truth image the accuracy of the propagated boundary

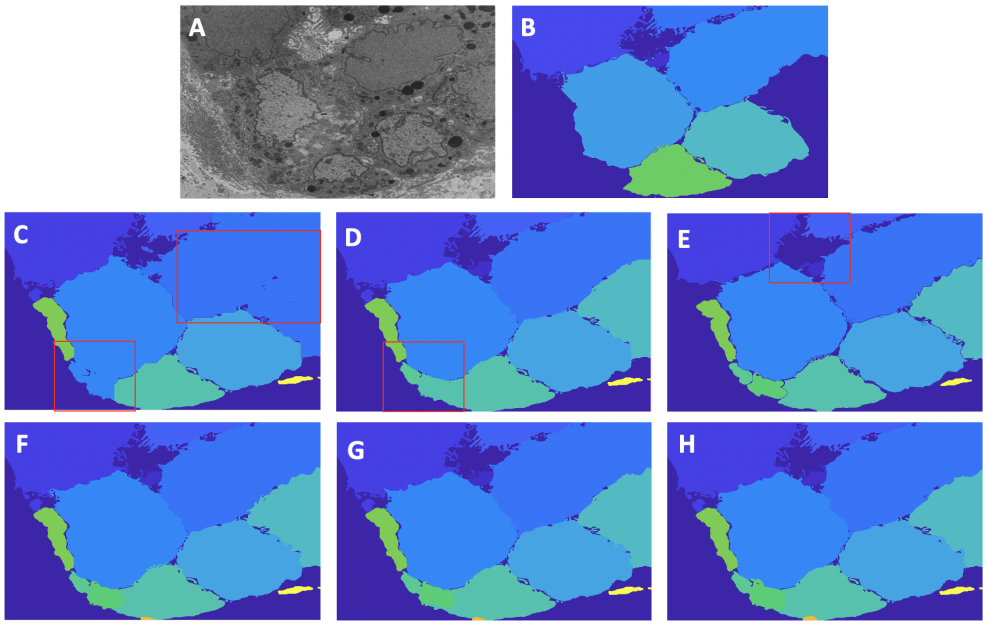


Figure 5.21: Cell segmentation results on Bx2 dataset. Representative qualitative results showing (A) input image, (B) ground truth segmentation (few cells are not labeled in the ground truth), (C) segmentation of cells using cell-interior mask alone, (D) segmentation of cells using cell-interior mask and boundary predictions from segmentation model, (E) segmentation of cells using optical flow alone, (F) segmentation of cells by overlaying boundaries propagated using optical flow on cell-interior mask obtained from segmentation model to separate cells, (G) segmentation of cells by selectively combining optical flow boundary estimate with the boundary estimated from segmentation model and (H) segmentation of cells by propagating boundaries estimated in the previous frame using optical flow and combining with the boundary estimated from segmentation model for Bx2 dataset

decreased. Further, the filopodia-like protrusions of cancer cells that appear as island-like blobs detached from the cell body, begin and end within a few images in the stack. Therefore, optical flow cannot track such islands if they are not present in the ground truth image and also cannot detect the start and end of such islands (Figure 5.20E and 5.21E). Therefore, in addition to boundary propagation, methods to detect these islands in each image and track them to the corresponding main cell are required. To do so, we train a ResUNet to segment the cell-interior mask and boundaries of all cells in the EM images. ResUNet provided precise cell boundaries even on images farther from ground truth images. However, the boundary segmentation from ResUNet could not separate adjacent cells that looked alike (Figure 5.20D and 5.21D). We combine the strengths of ResUNet and optical flow by taking the boundary segmentation from ResUNet and adding in the boundary propagated by optical flow in regions where ResUNet could not predict the boundary (in the presence of similar adjacent cells). By doing so, we retain the precise boundary

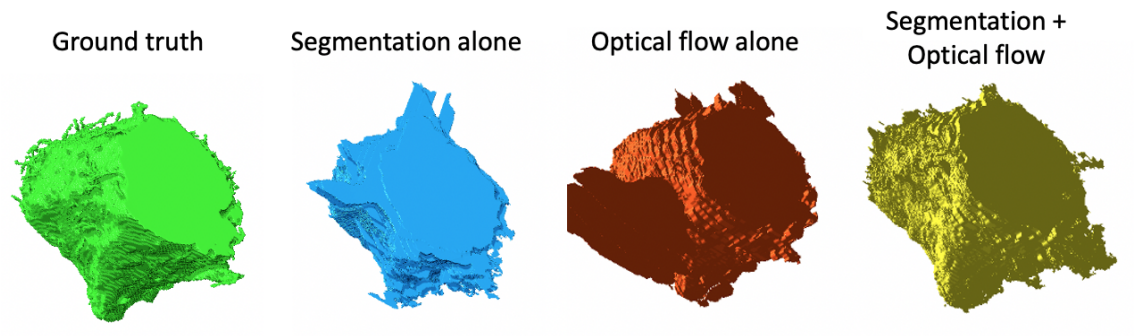


Figure 5.22: Volume rendering of a cell from Bx2 dataset. Volume rendering of the (a) ground truth, (b) cell segmented using ResUNet result and watershed, (c) cell segmented using boundary obtained from ground truth using optical flow, (d) cell segmented by combining boundary information from both ResUNet and optical flow.

segmentation from ResUNet while benefiting from the cell separating ability of the optical flow estimate. The combined boundary information is then overlaid on the cell-interior mask segmented using ResUNet. Individual cells are then separated from the cell-interior mask by performing watershed using centroids of cells in the nearest ground truth image as seeds. We experiment with different approaches of combining the segmentation and optical flow boundary estimates. First, we directly overlay the optical flow boundaries on the cell interior mask prediction and separated cells using the watershed algorithm (Figure 5.20E and 5.21E). We observe that as optical flow does not move the boundary accurately there was a step effect in the final 3D reconstruction of segmentation as seen in Figure 5.22. Therefore, we selectively include cell boundaries propagated using optical flow only in regions with overlapping cells. This helps retain the accurate cell boundaries from the segmentation model while separating overlapping cells (Figure 5.20F and 5.21F). In order to obtain a better continuity of the cell boundary, we propagate the combined cell boundary estimate of the previous image instead of the optical flow estimate alone (Figure 5.20G and 5.21G). All three methods of combining the segmentation and propagation boundary results improved the overall cell segmentation performance when compared to using the segmentation or propagation methods alone as shown in Table 5.4 and Figure 5.22. In PDAC dataset the cells mostly had extra-cellular matrix between them and therefore just using segmentation could result in good separation of cells as seen from Table 5.4. But still in cases were filopodia-like protrusions touched each other the combined method with propagation of boundary using optical flow based on the boundary estimate of the previous image produced better separation of the filopodia-like protrusions. Finally, for the third task, we track the islands to the main cells by calculating the intersection over union measure for all regions detected in consecutive images and associating regions with maximum overlap. As

Table 5.4: Cell segmentation performance measured by Dice score in Bx1, Bx2 and PDAC datasets.

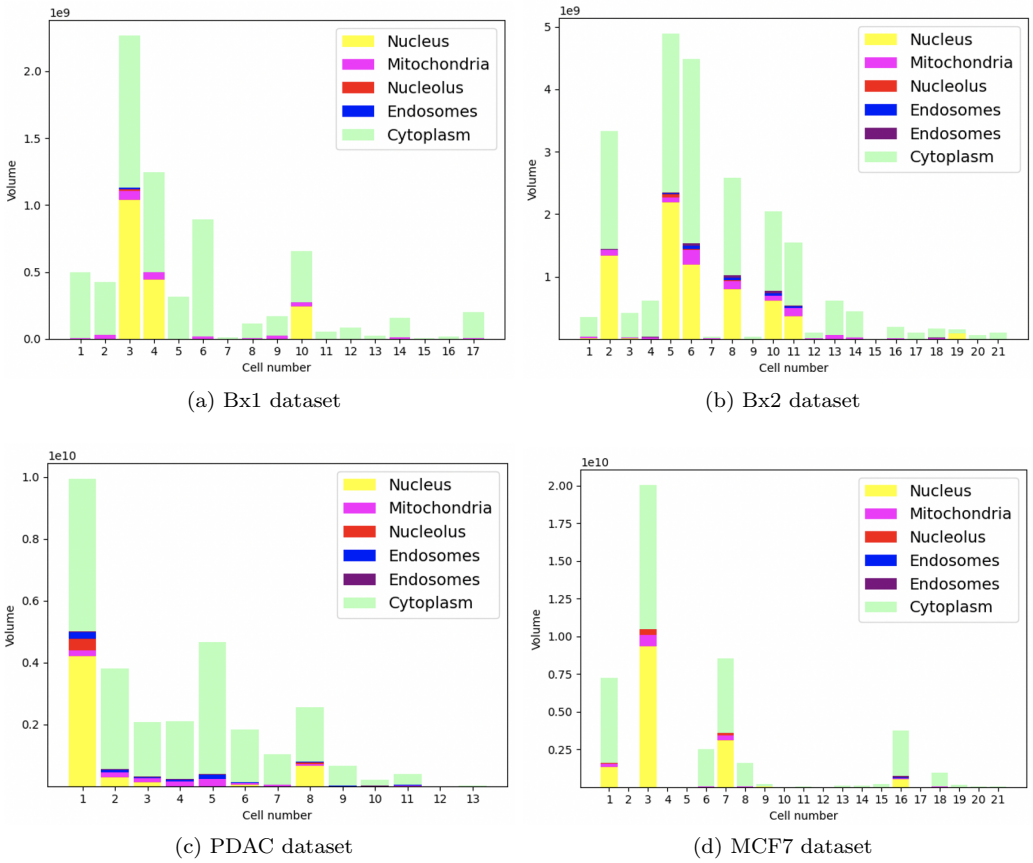
		Bx1	Bx2	PDAC
Segmentation alone	Mask	0.900	0.704	0.889
	Mask + Border	0.927	0.841	0.888
Propagation alone	Optical flow	0.936	0.881	0.888
Segmentation + propagation	Mask + optical flow	0.952	0.902	0.888
	Mask + Border + OF (selective)	0.950	0.900	0.888
	Mask + Border + OF (propagate through frames)	0.952	0.894	0.888

movement of regions between frames is minute, an overlap-based measure works well in tracking protrusions and cells in EM images. Volume renderings showing the FIB-SEM volume and the predicted cell and organelle segmentations for PDAC and MCF7 dataset are shown in Figure 5.24. Segmentation of cells enables quantitative cellular analysis, where we can localize organelles to a cell and estimate their variations in density, size and number among different cells. Figure 5.23 shows the volume occupied by the predicted organelles for each cell in each dataset.

The training of the segmentation model takes about 4 hours and prediction and splitting of the cells about 2-3 hours. Tracking of protrusions to main cells takes a longer time as each image has to be processed serially and the time also varies depending on the number of islands present in each image. It approximately takes around a minute to process each image in tracking of protrusions step. Complete cell segmentation of Bx1, PDAC, Bx2 and MCF7 take approximately 1, 1.5, 2 and 2 days respectively. This is a huge reduction from months it takes to manually segment cells in these datasets.

5.7.7 Quantitative Characterization of Nuclei and Nucleoli Morphology and Texture

Segmentation of nuclei and nucleoli allows us to characterize biologically relevant features such as their morphology and texture. Figure 5.25 shows some of the morphological features we extract for nuclei (Figure 5.25A) and nucleoli (Figure 5.25B and 5.25C). The solidity feature, measuring the concavity of a surface, captures the nuclear envelope invaginations. The higher level of invaginations in Bx2 and Bx4 when compared to PTT, Bx1, PDAC and MCF7 (Figure 5.9A) is



(a) Bx1 dataset

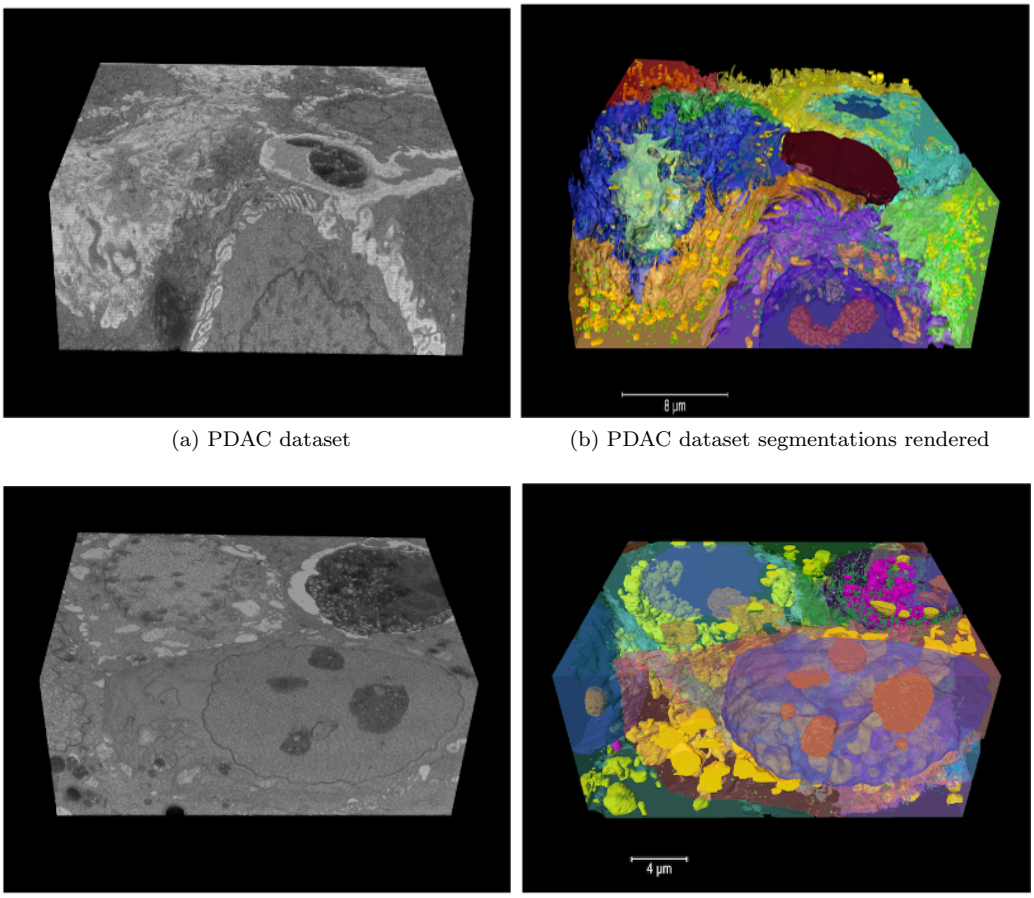
(b) Bx2 dataset

(c) PDAC dataset

(d) MCF7 dataset

Figure 5.23: Volume occupied by the predicted organelles for each cell in each dataset.

reflected by their lower solidity value (Figure 5.25A). Similarly, the relatively smooth envelope of nucleoli in PTT is reflected by its higher solidity value (Figure 5.25B). The solidity of the nucleolus is calculated by filling the holes in the volume to exclude the effect of the volume fenestrated by pores and quantify only the overall change in shape. The sphericity and circular variance features measure roundness of an object, and can be used to capture shape irregularities of nuclei and nucleoli, common characteristics of cancer cell. Additionally, for the nucleolus, we calculate the percentage of volume fenestrated by pores. These values are high for the PTT and Bx4 datasets as a result of the complex structure of pores within the nucleoli. Finally, varying levels of proximity of nucleoli to the nucleus membrane are also observed. This observation is consistent with published studies which suggest that nucleoli in cancer cells often move towards the nuclear membrane and form intranuclear canalicular systems between nuclear membrane and nucleolus [11]. Accordingly, the proximity of nucleolus to nuclear membrane feature captures the more centered positioning of nucleoli in the PPT dataset and the close proximity of nucleoli to nuclear membrane in the Bx2



(a) PDAC dataset (b) PDAC dataset segmentations rendered

(c) MCF7 dataset (d) MCF7 dataset segmentations rendered

Figure 5.24: Volume renderings showing the FIB-SEM volume and the predicted cell and organelle segmentations.

dataset. While most of the nuclei contain one nucleolus, few nuclei contain 2-6 nucleoli (figure 5.25C).

Figure 5.26 shows the texture features we extract for nuclei (Figure 5.26A) and nucleoli (Figure 5.26B). The texture features capture the spatial distribution of intensity patterns associated with chromatin and other internal structures. Texture characterization is well studied in the field of image processing. We use three classic methods, namely the grey level co-occurrence matrix (GLCM), the pattern spectrum, and the size zone matrix (SZM), to derive a set of texture features listed in Figure 5.26. GLCM captures the distribution of co-occurring gray-level intensities [15]. From GLCM, we extracted features such as homogeneity, correlation, variance and contrast for the four datasets (Figure 5.26A). The pattern spectrum feature characterizes the distribution of the sizes of various objects in an image [170]. The SZM quantifies the size of homogenous grey

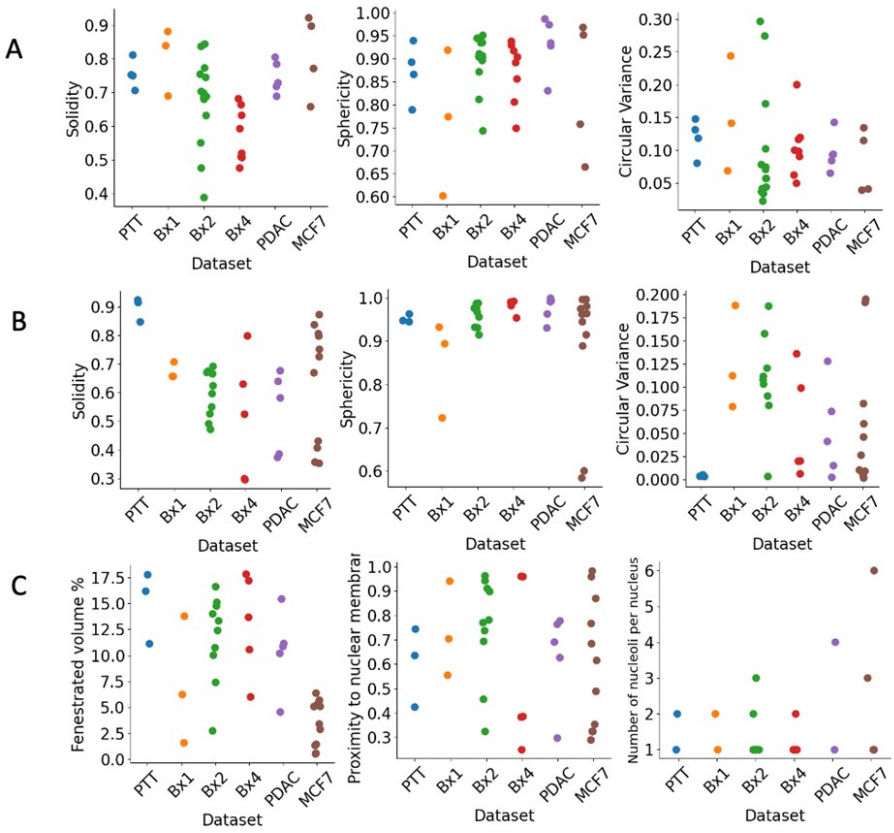


Figure 5.25: Morphological features extracted from nuclei and nucleoli. Solidity, sphericity and circular variance measures for (A) nuclei and (B) nucleoli in PTT, Bx1, Bx2 and Bx4 datasets. (C) Percentage of fenestrated volume in nucleoli and proximity of nucleoli to nuclear membrane for all datasets. Each dot represents the value of the feature for a nucleus or nucleolus.

level zones in an image [168], from which we extracted three features - zone sizes centroid, zone percentage and small zone high grey level emphasis. These features capture the sizes of groups of voxels of similar gray level intensities and can characterize the granularity of the chromatin structure in nuclei. These texture features are capable of capturing the differences between the tissue samples, and can be potentially used for downstream analysis.

5.8 Conclusion

We present here a framework for segmentation of cellular ultrastructure of cancer tissues from 3D FIB-SEM images, enabling rendering of the ultrastructure into interpretable forms and extraction of quantitive features. We used a ResUNet architecture to segment cells, nuclei, nucleoli, mitochondria, lysosomes and endosomes and evaluated the performance of the model on five human cancer

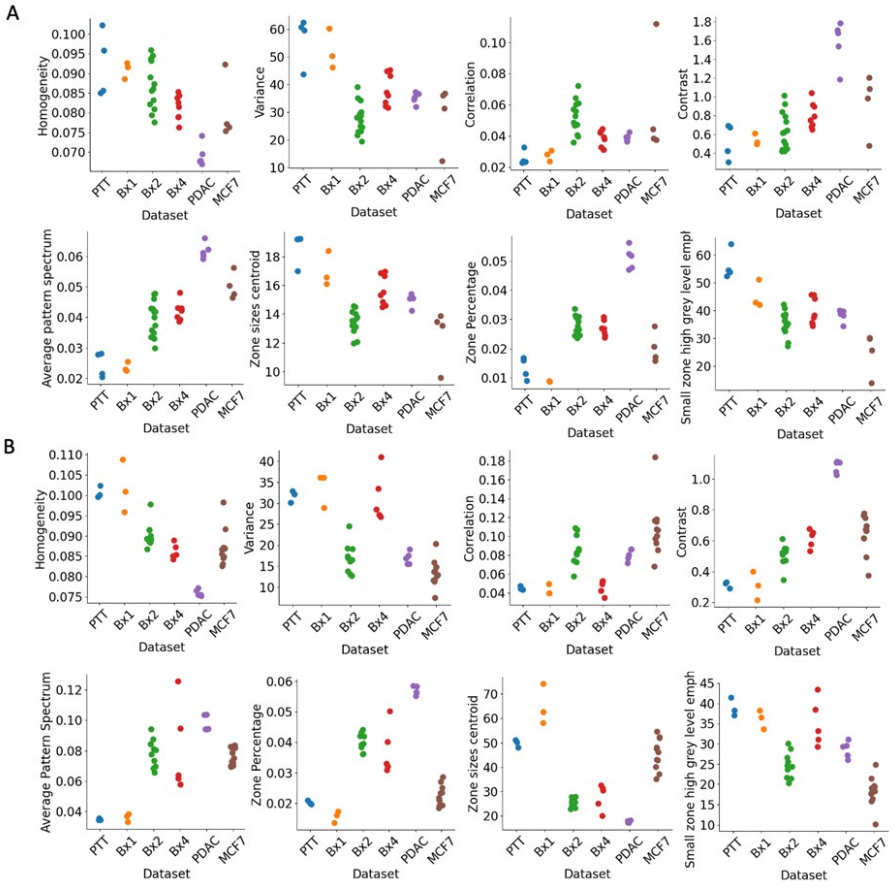


Figure 5.26: Texture features extracted from nuclei and nucleoli. GLCM features (top row), pattern spectrum and SZM features (bottom row) for (A) nuclei and (B) nucleoli in PTT, Bx1, Bx2 and Bx4 datasets. Each dot represents the value of the feature for a nucleus or nucleolus.

biopsy samples from a clinical trial at OHSU supported by HTAN and a microspheroid prepared using a breast cancer cell line. The ResUNet architecture was trained with different sizes of training datasets to evaluate the number of manually labeled images required for accurate segmentation. It was observed that the number of manually labeled images required greatly depended on the variability of the structure across the dataset. When the structure was uniform across the dataset, roughly 1% of the dataset labeled was enough to train an efficient model. Even in the presence of large variations in the structure, 2% of dataset labeled seemed to be sufficient for good segmentation results. Also, using image tiles of size 2048×2048 down sampled to 512×512 improved segmentation results when compared to using 512×512 crops directly, as larger crops provided greater context which appeared to increase the segmentation performance. We demonstrated that ResUNet provided segmentation of nuclei with dice score 0.99, 0.99 and 0.98 and nucleoli with dice

score 0.98, 0.93 and 0.80 for PTT, Bx1 and Bx2 datasets respectively and mitochondria with 0.86 and 0.76 for Bx1 and Bx2 datasets respectively. We presented a multi-pronged approach combining segmentation, propagation and tracking strategies to segment cells in EM images. Combining propagation and segmentation methods helped to accurately segment cells even in the presence of adjacent cells with similar intensity and texture variations. In PDAC dataset the cells mostly had extra-cellular matrix between them and therefore just using segmentation could result in good separation of cells. But still in cases were filopodia-like protrusions touched each other the combined method with propagation of boundary using optical flow based on the boundary estimate of the previous image produced better separation of the filopodia-like protrusions. The two extreme cases are when all cells in the dataset are either separate or all are closely packed touching each other. When all are separate with exra-cellular matrix between them, a segmentation method alone can result in good cell segmentation and when all cells are closely packed and adjacent cells look alike only propagation based methods can separate them. But in most of the cancer datasets we observe that only parts of the cells lie adjacent to another cell and therefore an approach combining the merits of segmentation and propagation methods is beneficial. The filopodia-like protrusions in cancer cells are an important feature in understanding the interactions of cancer cells and therefore it is necessary to segment them accurately. We demonstrated a tracking mechanism that can track the ends of filopodia-like protrusions that appear as island-like blobs detached from the main cell in few images and correctly associate them with their respective main cells.

Structures in 3D data collected via FIB-SEM exhibit high variability due to several factors, including the sample quality, tissue type, sample preparation techniques, microscope settings, and the imaging pixel resolution [129]. A small subset of the whole dataset contains enough information to capture most of the variability of a given structure with respect to the dataset, and we demonstrate that a ResUNet trained with sparse labels is able to generate segmentation masks for the rest of the images in the stack. Due to intensity variations caused by acquisition parameters and the variability of features among datasets the generalizability of the segmentation models is currently limited. In order to have an automatic segmentation framework which does not necessitate sparse manual labeling for every new dataset it is necessary to have a model that has enough representative training data. Our initial trials to segment individual 3D volumes by sparsely labeling each volume is a step towards building a dataset large enough to capture the variability among different organelles and the variability caused by different image acquisition settings. Segmentation of cells and organelles allowed us to extract quantitative features from the datasets. We also demonstrate the feasibility of morphology and texture quantification in nuclei and nucleoli. These quantitative features can be extracted efficiently, robustly, and reproducibly.

While it is beyond the scope of our work, we anticipate linking them to clinically relevant variables such as patient drug response in the future. This method can be extended to other cellular structures, enabling deeper analysis of inter- and intracellular state and interactions. The proposed segmentation of EM images fills the gap that has prevented modern EM imaging from being used routinely for research and clinical practices. It will enable interpretative rendering and provide quantitative image features to be associated with the observed therapeutic responses.

Chapter 6

Characterization: Multi-Resolution Fractal Analysis

6.1 Introduction

The next step in the image processing pipeline after segmenting a region of interest is to characterize the region. Characterization provides informative numerical features facilitating downstream analysis. The two spectral texture features defined in Section 2.3.4, fractal analysis - measure of structure's geometrical complexity, and wavelet analysis - multi-resolution representation, have been used separately as features in multiple applications. In this chapter, we combine the two spectral texture analysis methods to obtain a new method, multi-resolution fractal analysis, wherein we perform fractal analysis at multiple resolutions to obtain a richer representation of the heterogeneity of the texture. We analyze the texture of DCE-MRI parametric maps with the new method to evaluate its potential in early prediction of breast cancer pathologic response to chemotherapy.*

*The work in this chapter has been published in

- A. Machireddy, G. Thibault, A. Tudorica, A. Afzal, M. Mishal, K. Kemmer, A. Naik, M. Troxell, E. Goranson, K. Oh, N. Roy, N. Jafarian , X. Song, 2019, Early Prediction of Breast Cancer Therapy Response using Multiresolution Fractal Analysis of DCE-MRI Parametric Maps. Tomography, 2019 March, 5(1):90-98.
- A. Machireddy A, G. Thibault, W. Huang, and X. Song, 2018, Analysis of DCE-MRI for Early Prediction of Breast Cancer Therapy Response, in 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 682-685).
- A. Machireddy, G. Thibault, A. Tudorica, A. Afzal, M. Mishal, K. Kemmer, A. Naik, M. Troxell, E. Goranson, K. Oh, N. Roy, N. Jafarian , X. Song, 2019, Integration of DCE-MRI Texture Features with Clinical Data for Improved Early Prediction of Breast Cancer Therapy Response, in Annual meeting of International Society for Magnetic Resonance in Medicine.

6.2 Background and Related Work

6.2.1 Breast Cancer Response to Chemotherapy

Breast cancer is the second leading cause of death amongst all cancers occurring in American women [155]. The survival rate and prognosis of a breast cancer patient is dependent on the stage of cancer at diagnosis. Locally advanced breast cancers (generally with tumor size > 2 cm) are often treated with neoadjuvant chemotherapy (NACT) before surgery to reduce the tumor size for breast conserving surgery [33, 92]. The NACT regimen typically consists of 4 cycles of doxorubicin–cyclophosphamide administration every 2 weeks followed by 4 cycles of a taxane every 2 weeks, or 6 cycles of the combination of all three drugs every 3 weeks [173, 170]. The targeted agent, trastuzumab, is added to the regimen for tumors with positive HER2 (human epidermal growth factor receptor 2) receptor status. A full NACT course therefore would normally last $4-5$ months.

Pathological analysis of the post-NACT surgical specimens is performed to determine the response to NACT. The values of the cross-sectional size of the tumor in 2D, tumor cell density, number of lymph nodes involved, and the greatest dimension in the largest involved node are measured and used in the equation given by Symmans et al. to compute the residual cancer burden [160]. A pathological complete response (pCR) is defined as the absence of residual invasive tumor, indicated by residual cancer burden $= 0$. Non-pCR includes all cases with residual cancer burden > 0.

A pathological complete response to NACT is considered a surrogate marker for overall and long-term disease free survival [142]. A surrogate marker is a measure that correlates with a real clinical endpoint and hence is used as its substitute in many studies. However, the pCR rate is only $6\% - 45\%$ depending on breast cancer subtypes and treatment regimen [26, 183, 196]. It is therefore important to identify the non-responders at an early stage of NACT so that their treatment regimen can be modified, sparing them the long- and short-term toxicities from ineffective chemotherapies. In the current standard of care, the response to NACT is evaluated based on the histological examination of a surgical specimen taken after the completion of NACT. Noninvasive or minimally invasive methods that can predict therapy response at the early stages of NACT can potentially play an important role in the emerging era of precision medicine to help guide regimen de-escalation/alteration in NACT treatment of breast cancer [173].

6.2.2 Indicators of Response

A significant change in the microenvironment of the tumor, such as perfusion and metabolism, usually precedes a reduction in tumor size as a response to chemotherapy [101, 191, 138, 107]. As a noninvasive imaging method for assessment of microvascular perfusion, dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) is increasingly used in research and early phase clinical trial settings to predict and evaluate cancer response to treatment [173, 101].

Several studies [173, 109, 84, 108, 162, 106, 86, 4, 190, 133] have shown that changes in quantitative parameters estimated from pharmacokinetic modeling of DCR-MRI data can be useful markers for early prediction of breast cancer response to NACT. When compared to normal tissue vasculature (arrangement of blood vessels), tumor vasculature exhibits greater spatio-temporal heterogeneity. The heterogeneity of the tumor vasculature reflects the tumor stage and the disease progression [50]. Therefore, texture features capable of capturing the change of heterogeneity in tumor microvasculature can provide early prediction of breast cancer response to NACT.

6.2.3 Prediction of Response

The aforementioned DCE-MRI studies [173, 109, 84, 108, 162, 106, 86, 4, 190, 133] generally reported changes in mean parameter values of the entire breast tumor, masking the potential changes in spatial heterogeneity of the microvasculature in response to NACT. Image texture features that can capture the heterogeneity of tumor vasculature from DCE-MRI images or parametric maps could be highly useful in assessing tumor response to NACT.

Analysis directly on images: We first look at methods that analyze the texture from the DCE-MRI images directly. Several texture analysis methods such as grey-level co-occurrence matrix (GLCM) and gray-level run length matrix (RLM) (Section 2.3.3) have been frequently used in DCE-MRI analysis [164, 66]. They were initially used on DCE-MRI images directly. Teruel et al. analyzed T1-weighted DCE-MRI images using GLCM features to predict breast cancer response to NACT [164]. They extracted sixteen textural features at each time point of a DCE-MRI acquisition and the most significant feature yielded an area under the receiver operating curve (ROC AUC) of 0.77 for prediction of pCR vs. stable disease. Similarly, Golden et al. used GLCM features from pre- and post-NACT DCE-MRI images to evaluate NACT response [66]. The pre-NACT features were able to predict pCR with an AUC of 0.68. Though the post-NACT features showed favorable performances in predicting pCR, these were obtained after the completion of NACT and are not useful for early prediction of pCR.

Analysis on parametric maps: In contrast to analyzing the DCE-MRI images, several studies have performed the same texture analysis on parameteric maps. A paramatric map is estimated from pharmacokinetic modeling of DCE-MRI data. The DCE-MRI time-course data from the voxels within the tumor ROI is fitted with a two-compartment-three-parameter shutter speed model [173, 192], using a population averaged arterial input function from the axillary artery [173, 192]. This pharmacokinetic analysis yields the following four parametric maps: K^{trans} (volume transfer rate constant), v_e (volume fraction of extravascular and extracellular space), k_{ep} ($= K^{trans}/v_e$, efflux rate constant), and τ_i (mean intracellular water lifetime). A flowchart of DCE-MRI data acquisition and analysis is shown in Figure 6.6.

Banerjee et al. extracted a combination of intensity, texture, shape and edge-based features from 2D maps of pharmacokinetic parameters before and after NACT to assess treatment response [13]. Their best model obtained an AUC of 0.83, using a concatenation of Riesz and first-order statistical features. However, the use of pre- and post-NACT data limits the utility of this model for early prediction of NACT response. Thibault et al. extracted multiple statistical texture features from 3D pharmacokinetic parametric maps before and after one cycle of NACT, and correlated them with residual cancer burden values using a regression model [170]. They found that 3D GLCM features were most effective for early prediction of NACT response.

Fractal analysis: In all the analysis methods described above, texture has been studied on a statistical level, by analyzing the spatial distribution of the grey-level values. Textures can also be characterized by fractals, which describe irregular structures that show self-similarity at various scales, as elaborated in Section 2.3.4. Fractal based texture analysis correlates texture heterogeneity to fractal dimension (FD), which is a mathematical descriptor of a structure's geometrical complexity based on the concept of spatial pattern self-similarity. Rose et al. showed that fractal analysis could be used to quantify spatial heterogeneity in DCE-MRI parametric maps and differentiate between low and high-grade tumors [146]. Several other studies have used fractal analysis of breast DCE-MRI images to classify benign vs. malignant tumors [156, 130].

Multi-resolution analysis: Another important aspect while considering textures is the resolution at which it is analyzed. Due to the highly heterogeneous nature of the tumor vasculature, analyzing images at a single resolution may not be able to capture the entire complexity of the tumor vasculature. A multi-resolution approach can decompose an image into different levels of resolution, giving an opportunity to extract informative features at each level. Lower resolution

levels best represent large structures or high contrast objects, while higher resolution levels represent small structures or low contrast objects [67]. Multi-resolution analysis gives the advantage of analyzing both small and large object characteristics in a single image at several resolutions and therefore may be better suited to describe the highly heterogeneous tumor vasculature structure.

Multi-resolution methods, such as the wavelet analysis (Section 2.3.4), transform images into a representation containing both frequency and spatial information [148]. The mean and entropy values extracted from the sub-images resulting from wavelet decomposition of DCE-MRI images have been used to classify malignant and benign breast tumors [174, 175]. Braman et al. used Gabor wavelet to generate 1980 features from DCE-MRI images to predict breast cancer response to NACT [28]. A feature selection step was carried out to select the top 10 features for final classification.

6.2.4 Combining Multi-Resolution and Fractal Analysis

Al-Kadi et al. combined wavelet analysis with fractal analysis to characterize tissue in ultrasound images of liver tumors and showed that multi-scale analysis of tumor-heterogeneity improved prediction of response to therapy and disease characterization [6]. Inspired by that work, in this chapter, we design a multi-resolution fractal analysis framework for analysis of tumor heterogeneity in DCE-MRI parametric maps.

6.3 Proposed Approach

In this chapter, we combine the two spectral texture analysis methods, fractal and wavelet analysis, to obtain a new method, multi-resolution fractal analysis, wherein we perform fractal analysis at multiple resolutions to obtain a richer representation of the heterogeneity of the texture. As currently many image analysis tasks are seeing great results from general purpose convolutional neural network over methods specifically developed for a task, we analyze the potential of convolutional neural network in extracting reliable features for early prediction of NACT response.

6.3.1 Multi-Resolution Fractal Analysis

Wavelet Analysis for Multi-Resolution Decomposition

As seen in Section 2.3.4 wavelet analysis decomposes an image into a set of frequency sub-bands based on small basis functions of varying frequency and limited time duration called wavelets, enabling the characterization of texture at appropriate frequency levels. We decompose each of

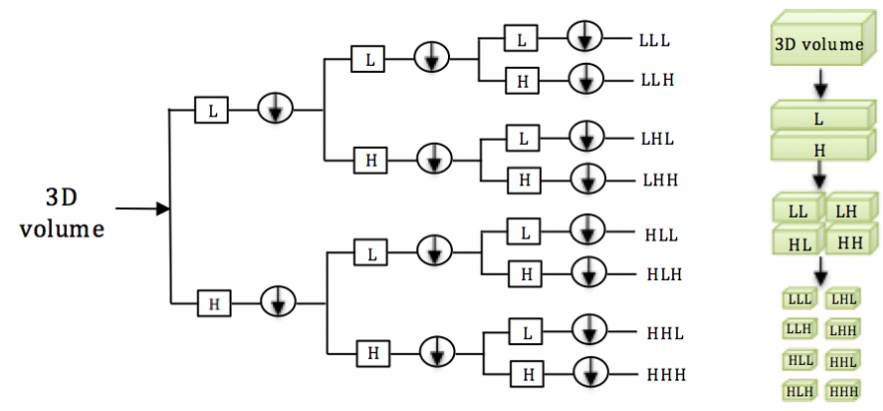


Figure 6.1: One level of wavelet transform for a 3D volume. The down arrow in the circle represents downsampling, L represents low-pass filter and H represents high-pass filter.

the four DCE-MRI parametric maps (described in Section 6.4.2) into a multi-resolution representation using wavelet analysis. One level of the wavelet decomposition for a 3D volume is shown in Figure 6.1. The 3D wavelets can be constructed as separable products of 1-D wavelets in the three spatial directions x, y and z. The process of decomposing a parametric map with a 1-D wavelet can be viewed as passing it through a low-pass and high-pass filter and downsampling. The 3D volume is first filtered along the x-direction resulting in a low-pass filtered sub-volume L and a high-pass filtered sub-volume H. These resulting sub-volumes are further filtered along y-direction, resulting in four decomposed sub-volumes: LL, LH, HL, and HH. Then each of the four sub-volumes are filtered along the z-direction, resulting in eight sub-volumes: LLL, LLH, LHL, LHH, HLL, HLH, HHL and HHH. We use Daubechies wavelets, as they account for signal discontinuities and self-similarity, which make them the most suitable wavelet for describing signals exhibiting fractal patterns [44]. Unlike Haar wavelets, they use overlapping windows that help capture changes in high frequency, and they also demonstrate better recognition of fine characteristic structures [119]. One level of decomposition results in eight sub-volumes. Fractal dimension is calculated for each of these sub-volumes.

Multi-Resolution Fractal Analysis

We calculate the fractal dimension based on the power spectrum analysis of the 3D Fourier transformation of the sub-volumes as defined in Section 2.3.4. In standard wavelet analysis, the energy of the sub-volume is used to guide further decomposition, but this value is highly dependent on the intensity values of the sub-volume [185]. Instead of using energy, we select the sub-volume with the highest fractal dimension for further decomposition as it represents the roughness of the

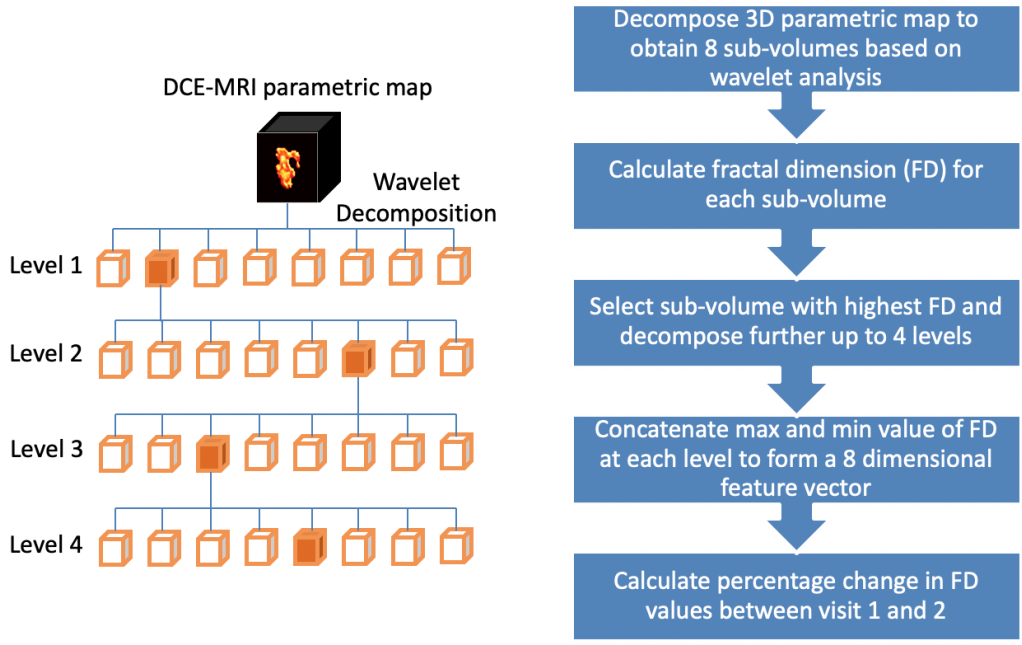


Figure 6.2: Schematic of decomposition of parametric map and a flowchart of multi-resolution fractal analysis of parametric maps of DCE-MRI to evaluate cancer response to NACT.

texture surface and does not depend on the local intensity variations.

We decompose each parametric map down to four levels using fractal dimension to guide the sub-band tree structure. We experimented with multiple combinations of features from the four levels. For the results presented in this chapter, we concatenate the highest and the lowest fractal dimension at each level of decomposition to form a feature vector. Therefore, for each parametric map we generate an 8-dimensional feature vector from multi-resolution fractal analysis. The schematic of decomposition of parametric map and a flowchart of multi-resolution fractal analysis of parametric maps of DCE-MRI to evaluate cancer response to NACT is shown in Figure 6.2.

6.3.2 Convolutional Neural Network as Feature Extractor

We also analyze the potential of 2D convolutional neural networks in extracting reliable features for early prediction of NACT response. As the size of the dataset is small, we design a small network with five layers consisting of two convolutional layers and three fully connected layers. A pooling layer performing a 2 x 2 max pooling follows each convolutional layer. The proposed convolutional network is shown in Figure 6.3. All parametric maps are resized to a fixed resolution of 128 x 128 before feeding into the CNN. The first convolutional layer uses filters of size 5 x 5, whereas the second layer uses filters of size 3 x 3. In order to avoid over-fitting, dropout and data augmentation

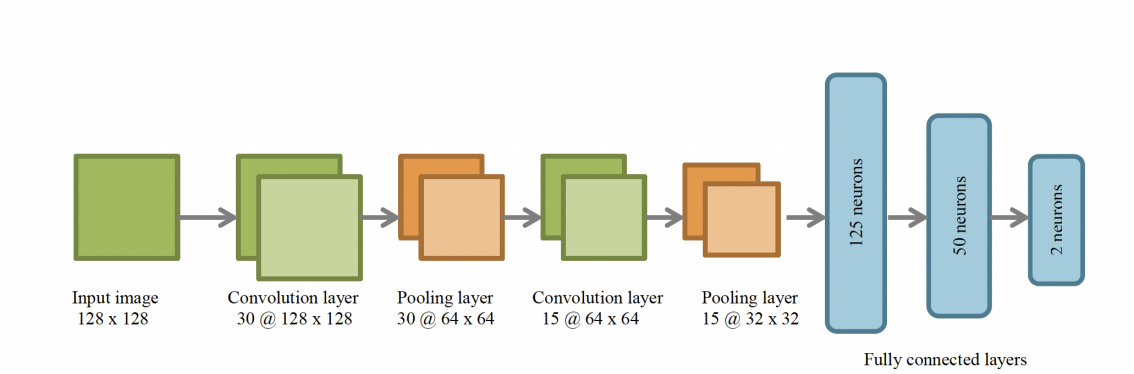


Figure 6.3: The architecture of the convolutional neural network used to analyze the parametric map

methods are applied. The dropout method randomly sets a fraction of input units to 0 at each update during training time, thus reducing over-fitting. In order to further reduce over-fitting the network on the small dataset, we augmented the data using rotations and horizontal reflections.

6.4 Methods

6.4.1 Patient Cohort and Study Schema

Fifty-five patients diagnosed with locally advanced breast cancer received standard of care NACT. They were consented to participate in a longitudinal research DCE-MRI study approved by the local IRB. The DCE-MRI images were collected at Oregon Health and Science University, Portland.

A total of four DCE-MRI exams are performed before, during, and after the NACT course: pre-NACT (visit-1), after the first NACT cycle (visit-2), at NACT midpoint (visit-3; usually after 3 or 4 cycles of NACT, or before the change of NACT agents), and after the completion of NACT but before surgery (visit-4). The DCE-MRI schedule is shown in Figure 6.4. Except for the visit-1 exam, all exams are performed at least a week after administering the latest cycle of NACT agents to allow time for the drugs to take effect.

In this chapter, only data from the visit-1 and visit-2 DCE-MRI studies are used as we want to predict the response to NACT at an early stage of the NACT cycle.

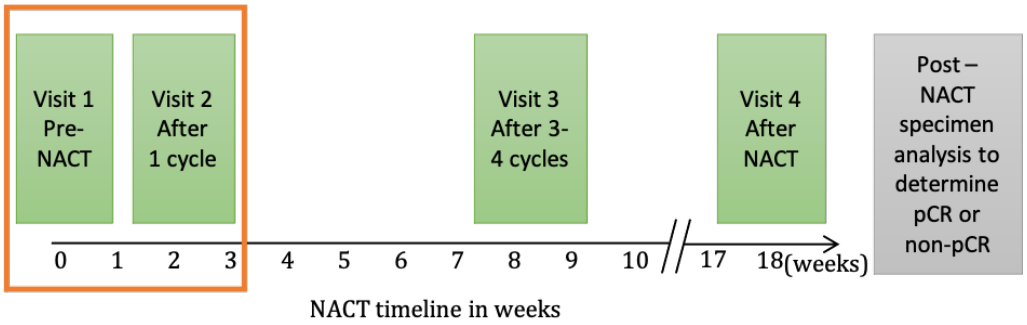


Figure 6.4: The schedule of DCE-MRI scans during the longitudinal study. We only use data from visit-1 and visit-2 to assess the capability for early prediction of breast cancer response to NACT.

6.4.2 DCE-MRI Data Acquisition and Analysis

Data Acquisition

DCE-MRI data acquisition is performed using a Siemens 3T system with the body coil as the transmitter and a 4-channel bilateral phased-array breast coil as the receiver. During each MRI session, following pilot scans and pre-contrast axial T1- and T2-weighted MRI acquisitions, axial bilateral DCE-MRI images with full breast coverage are acquired using a three-dimensional gradient echo-based time-resolved angiography with stochastic trajectories sequence [173, 192]. DCE-MRI acquisition parameters include 10° flip angle, 2.9/6.2 millisecond echo time/repetition time, a parallel imaging acceleration factor of two, 30 to 34 cm field of view, 320×320 in-plane matrix size, and 1.4-mm slice thickness. About $32 - 34$ image volume sets of $104 - 128$ slices each are acquired over a period of about 10 min with a temporal resolution of $14 - 20$ seconds. The contrast agent Gd (HP-DO3A) is injected intravenously (0.1 mmol/kg at 2 ml/s) using a programmable power injector after acquisition of two baseline image volumes, followed by a 20-ml saline flush at the same injection rate.

Delineation of the Tumor Region

Three experienced breast radiologists manually delineated the tumor region of interest (ROI) on post-contrast ($90 - 120$ seconds after the injection of the contrast agent) DCE-MRI image slices that contain the contrast-enhanced tumor. To minimize inter-observer variability in tumor ROI drawing for the same patient, one radiologist drew ROIs for the entire longitudinal study of a single patient. With only two patients having multi-focal disease, ROIs are drawn for the primary breast tumors only. Figure 6.5 shows an example of post-contrast DCE images from a pCR patient, drawn tumor ROIs, and ROI mean signal intensity ratio time-courses at visit-1 and visit-2. For

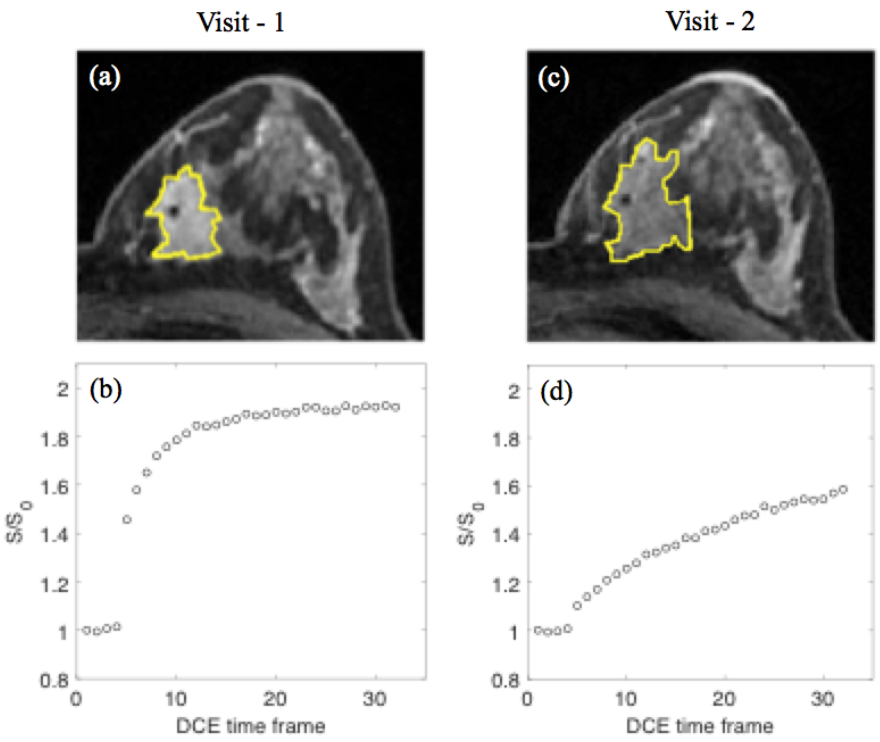


Figure 6.5: The visit-1 and visit-2 post-contrast DCE-MRI image slice (a and c, respectively) through the center of the primary breast tumor of a pCR patient [35 years, grade 2 invasive ductal carcinoma, 2.9 cm in the longest diameter at visit-1, ER (estrogen receptor) -, PR (progesterone receptor) +, HER2 + receptor status]. The tumor ROI boundries are shown in yellow. The time courses of mean signal intensity ratio, S/S_0, in the tumor ROI are shown in b and d for visit-1 and visit-2, respectively. S_0: signal intensity at baseline prior to contrast injection.

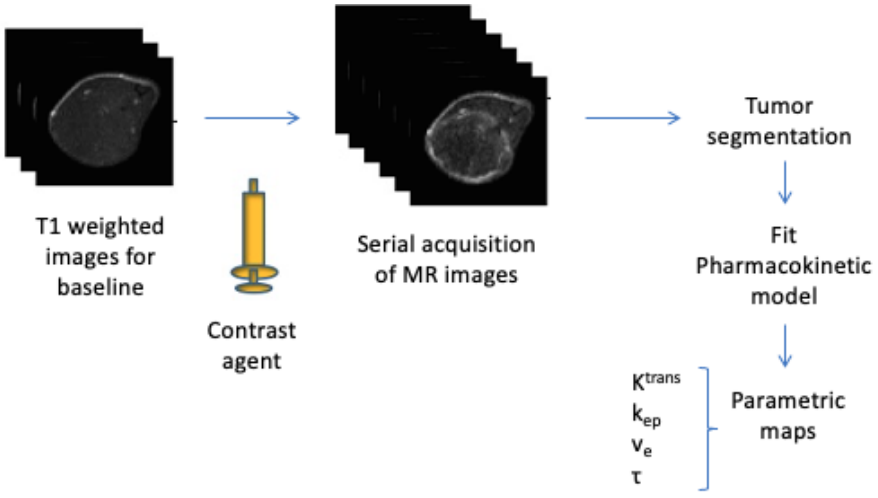
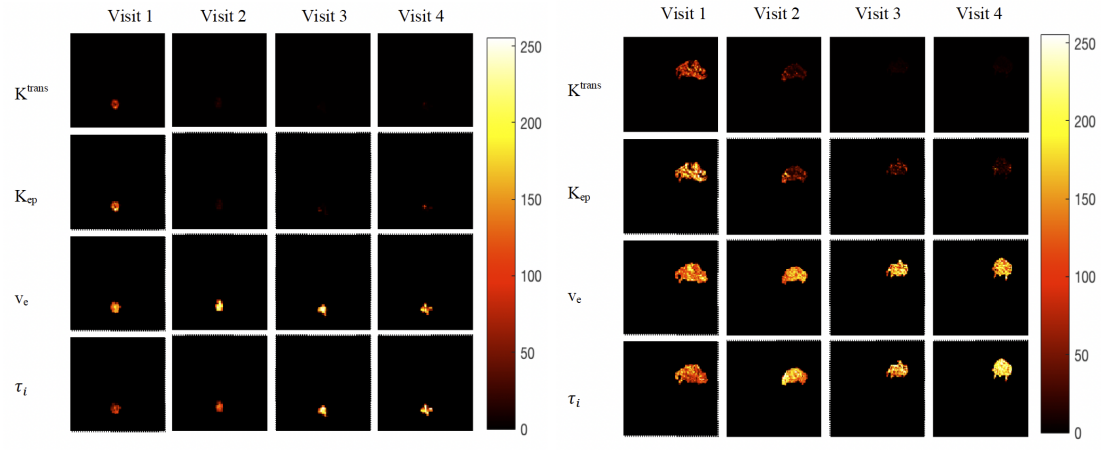


Figure 6.6: Steps in DCE-MRI Data Acquisition and Analysis.



(a) pCR subject (b) non-pCR subject

Figure 6.7: The central slices of the parametric maps K^{trans}, k_{ep}, v_e and τ_i of the tumor from the patient with (a) pCR and (b) non-pCR

pharmacokinetic analysis, pre-contrast tissue T1 value, T10, is determined using a proton density method [173, 192] by acquiring proton density images just before DCE-MRI that are spatially co-registered with the DCE images.

Generation of Parametric Maps

As mentioned earlier pharmacokinetic analysis on DCE-MRI yields the following four parametric maps: K^{trans}, v_e, k_{ep}, and τ_i. The variation of the parametric maps during the four DCE-MRI acquisitions for a patient achieving pCR and non-pCR are shown in Figure 6.7. Previous studies have observed that there is a significant decrease in the values of K^{trans} and k_{ep} in patients achieving pCR, whereas patients achieving non-pCR have a significant increase in v_e during the initial stages of NACT [139]. Similar trends are observed in the data as seen in Figure 6.7. Figure 6.8 shows examples of voxel based parametric maps of the four parameters for a pCR (6.8a) and a non-pCR (6.8b) tumor at visit-1 and visit-2. The parametric maps from the visit-1 and visit-2 studies of all patients are subjected to multi-resolution fractal analysis.

6.4.3 Conventional Texture Feature Analysis

We compare the performance of our multi-resolution fractal analysis features with that of GLCM, RLM and single resolution fractal analysis described in Sections 2.3.3 and 2.3.4. GLCM is a second order statistical method, which estimates the joint probability $P(i, j|d, \theta)$, where two voxels with intensity i and j are separated by distance d and direction θ. A GLCM matrix is constructed by

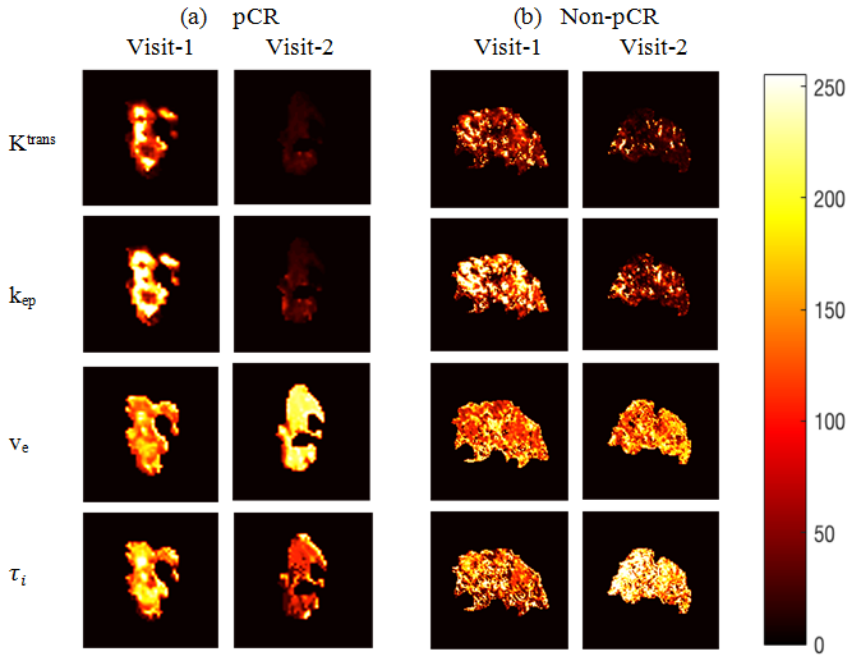


Figure 6.8: The visit-1 and visit-2 parametric maps of K^{trans}, k_{ep}, v_e, and τ_i of the tumor ROI on an image slice through the center of the tumor: (a) a 27 year-old pCR patient with a grade 3 invasive ductal carcinima (5.0 cm in the longest diameter at visit-1) and ER -, PR -, HER2 + receptor status; (b) a 45 year-old non-pCR patient with a grade 2 invasive mammary carcinoma (11.9 cm in the longest diameter at visit-1) and ER +, PR +, HER2 - receptor status.

averaging the matrices obtained over 13 directional offsets at distance $d = 1$ [146]. Twelve Harlick features are derived from this GLCM matrix [73]. RLM $P(i, r|\theta)$ is defined as the number of pixels with gray-level i and run-length r, for a given direction θ. RLM is computed by adding all possible run lengths in the 13 directions of the 3D space and thirteen statistical features are derived from this matrix [57]. Fractal analysis describes the roughness or smoothness of the texture through the fractal dimension measure. Here, single-resolution fractal analysis refers to the estimation of fractal dimension of the tumor ROI from 3D parametric maps directly [192].

6.4.4 Evaluation of Predictive Performance for NACT Response

For each of the features obtained from the GLCM, RLM, multi- and single-resolution fractal analysis, the percentage change in the feature values is calculated between the visit-1 and visit-2 DCE-MRI studies. These percentage changes are given as input to support vector machine (SVM) [42], a robust classifier, to generate a predictive model for classification of pCR vs. non-pCR. The performances of the models are evaluated using accuracy, sensitivity, specificity and ROC AUC

analysis.

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{Total positive} + \text{Total negative}} \tag{6.1}$$

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{Total positive}} \tag{6.2}$$

$$\text{Specificity} = \frac{\text{True negative}}{\text{Total negative}} \tag{6.3}$$

Sensitivity here refers to the proportion of pCRs correctly identified as pCRs, while specificity refers to the proportion of non-pCRs correctly identified as non-pCRs. ROC AUC represents the model's capability of distinguishing between pCRs and non-pCRs. The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings. The area under the ROC curve provides an aggregate measure of performance across all possible classification thresholds. The ROC curve visualizes all possible classification thresholds, whereas accuracy, specificity and sensitivity represent results for a single threshold.

Among the fifty-five patients in the study cohort, fourteen achieved pCR to NACT, while the other 41 patients are non-pCRs based on pathological analysis of the surgical specimens. Table 6.1 shows the clinicopathologic characteristics of the pCR and non-pCR groups. The SVM classification performance is evaluated by calculating the average over ten random partitions of the data for training and testing. Each partition is formed by randomly selecting 9 pCRs/31 non-pCRs and 5 pCRs/10 non-pCRs to form the training and testing sets, respectively. The mean and standard deviation of accuracy, sensitivity, specificity and ROC AUC values obtained over the ten partitions of training and testing datasets are reported. The predictive performance is assessed for the features extracted from each of the four parametric maps as well as those constructed by concatenating the texture features from all four parametric maps of K^{trans}, k_{ep}, v_e and τ_i, designated as 'All'. The ROC AUC values of the multi-resolution fractal features are compared with those of the conventional features by calculating the critical ratio according to the Hanley and McNeil's formula [72]. The statistical significance is set at $P < 0.05$.

6.5 Results

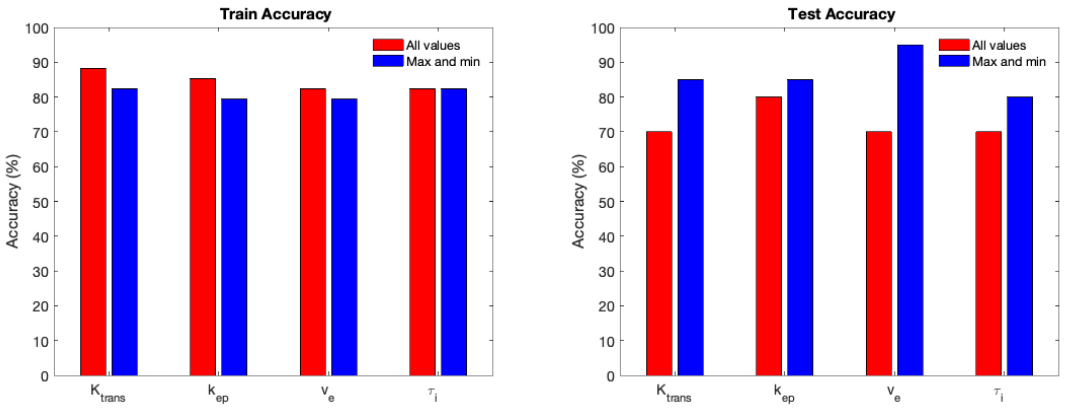
In Section 6.5.1, we compare the performance of multi-resolution fractal analysis with the conventional methods of GLCM, RLM and single-resolution fractal analysis. In Section 6.5.2 we evaluate the potential of 2D convolutional neural networks in extracting reliable features for early prediction of NACT response. In Section 6.5.3 we evaluate the potential of integrating DCE-MRI features with clinical data in further improving the response prediction.

Table 6.1: Clinicopathologic characteristics of pCR and non-pCR groups. (IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; IMC, invasive mammary carcinoma.)

	pCR (n=14)	non-pCR (n=41)
Age at diagnosis (years)	27-63	27-79
Tumor type	14 - IDC	34 - IDC 3 - ILC 4 - IMC
Tumor grade		
1	1	4
2	7	16
3	6	21
Tumor size in longest diameter (cm)	1.0 - 6.9	1.2 - 12.8
ER		
Positive	2	24
Negative	12	17
PR		
Positive	3	26
Negative	11	15
HER-2		
Positive	12	25
Negative	2	16

6.5.1 Comparision with Classical Methods

Design of feature vector: We first experimented with the design of the feature vector from multi-resolution fractal analysis. Each level of decomposition results in eight sub-volumes and fractal dimension is calculated for each of these sub-volumes. We experiment with two designs for the feature vector. First, we formed the feature vector by concatenating all fractal dimensions at each level. Next, we chose only the maximum and minimum valued fractal dimension at each level. Classification accuracy using both choices is shown in Figure 6.9. We observe that though using all features from each level results in better training accuracy, its test accuracy is lower when compared to choosing only maximum and minimum values at each level. This could be due to the high dimentionality of the feature vector in comparison to the size of the dataset leading to over-fitting. Further, the fractal dimensions of all sub-volumes may not have the same discrimination power, and inclusion of weak sub-bands may have a negative impact on the performance of the classifier. Therefore, we concatenate the maximum and minimum valued fractal dimension at each level of decomposition to form a feature vector.



(a) (b)

Figure 6.9: The accuracy values for classification of pCR from non-pCR patients in the training (a) and testing (b) datasets using the feature vector with all values (red) and feature vector with maximum and minimum values at each level of decomposition (blue) for K^{trans}, k_{ep}, v_e, and τ_i parametric maps.

Comparison with classical methods: Figure 6.10 shows the ROC AUC values for classification of pCR vs. non-pCR using the GLCM, RLM, single-resolution fractal, and multi-resolution fractal features from parametric maps of different DCE-MRI parameters considered individually and the concatenated features from all four parametric maps. GLCM and RLM features over-fit on the training data as they show high training AUCs but low testing AUCs. For example, the AUC values are 0.76 and 0.75 from the training K^{trans} maps, and 0.47 and 0.40 from the testing K^{trans} maps for the GLCM and RLM methods, respectively.

Overall, the multi-resolution fractal features from each DCE-MRI parametric map and the concatenated features perform the best in prediction of pCR vs. non-pCR in both the training and testing datasets with AUC = 0.85, 0.86, 0.87, 0.86, 0.91 (for K^{trans}, k_{ep}, v_e, τ_i, and All, respectively), and 0.80, 0.63, 0.74, 0.70, 0.78 for the training and testing datasets, respectively. The only exception is from the k_{ep} maps in the testing datasets, where the single-resolution fractal analysis provides the highest AUC of 0.71 among the four feature analysis methods. Within the testing or training sets, the predictive performances of multi-resolution fractal features are significantly better than the GLCM features from the K^{trans} map (AUC = 0.47, P = 0.022) in the testing set, RLM features from the τ_i map (AUC = 0.71, P = 0.012) in the training set, RLM features from the K^{trans} (AUC = 0.40, P = 0.013) and 'All' (AUC = 0.47, P = 0.049) maps in the testing set, and single-resolution fractal features from the v_e (AUC = 0.71, P = 0.012) and τ_i (AUC = 0.67, P = 0.037) maps in the training set (indicated by * in Figure 6.10).

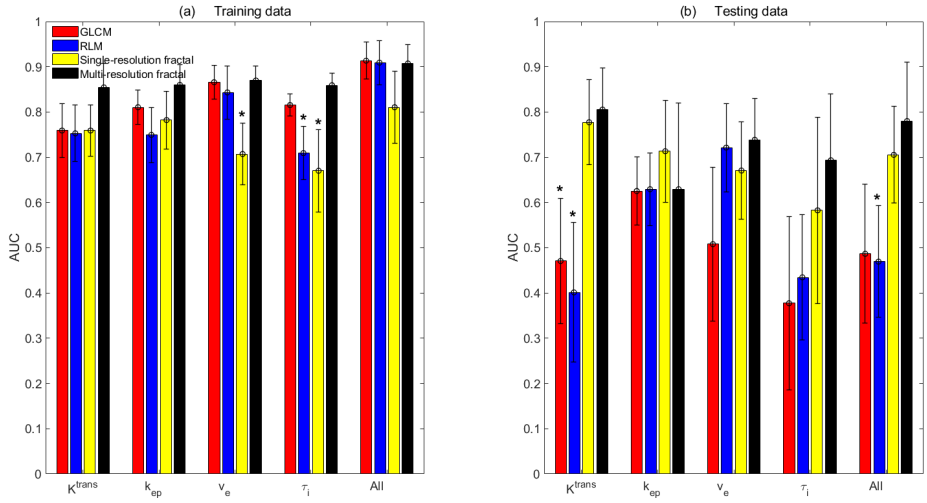
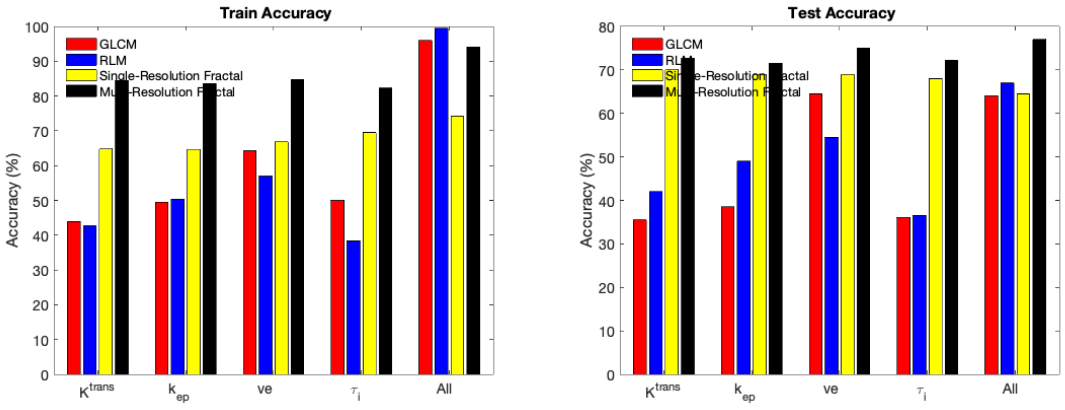


Figure 6.10: The ROC AUC values for classification of pCR from non-pCR patients in the training (a) and testing (b) datasets using the GLCM (red), RLM (blue), single-resolution fractal (yellow) and multi-resolution fractal (black) features extracted from the K^{trans}, k_{ep}, v_e, and τ_i parametric maps. The final column 'All' represents the concatenated features from all four parametric maps. The error bars represent the standard deviation obtained over the 10 different partitions of train and test data. *: significant ($P < 0.05$) difference in AUC compared to that of multi-resolution fractal features.



(a) (b)

Figure 6.11: The accuracy values for classification of pCR from non-pCR patients in the training (a) and testing (b) datasets using the GLCM (red), RLM (blue), single-resolution fractal (yellow) and multi-resolution fractal (black) features extracted from the K^{trans}, k_{ep}, v_e, and τ_i parametric maps. The final column 'All' represents the concatenated features from all four parametric maps.

Table 6.2: Specificity values [mean (standard deviation)] in the testing data set for the GLCM, RLM, single-resolution fractal and multi-resolution fractal methods with sensitivity set at 60% and 80%.

	GLCM	**RLM**	**Single-resolution fractal**	**Multi-resolution fractal**
Sensitivity = 60				
K^{trans}	49.3 (22.3)	34.7 (27.5)	84.0 (16.1)	89.3 (11.4)
$Kkep$	73.3 (9.40)	67.3 (17.9)	84.0 (7.2)	70.7 (23.3)
v_e	64.0 (31.6)	82.0 (14.4)	75.3 (7.1)	80.7 (12.4)
τ_i	40.7 (28.4)	44.0 (21.6)	57.3 (27.6)	68.7 (25.0)
All	49.3 (21.6)	42.0 (18.1)	82.0 (15.7)	82.7 (17.0)
Sensitivity = 80				
K^{trans}	19.3 (16.2)	25.3 (14.7)	63.3 (25.4)	68.7 (13.7)
k_{ep}	49.3 (19.9)	45.3 (19.1)	55.3 (29.3)	49.3 (27.8)
v_e	22.0 (30.0)	67.3 (19.7)	56.0 (20.7)	62.0 (17.2)
τ_i	18.0 (23.1)	30.0 (24.6)	47.3 (24.2)	62.0 (37.7)
All	32.0 (21.5)	28.7 (16.3)	62.0 (26.9)	62.0 (17.8)

We evaluate the specificities of the classification models at two levels of sensitivities for the testing datasets: 60% (3 out of 5 pCRs were classified correctly) and at 80% (4 out of 5 pCRs were classified correctly), as shown in Table 6.2. At both sensitivity levels, with a few exceptions, fractal features presents higher specificities than the GLCM and RLM features with the multi-resolution method generally outperforming the single-resolution method (except when they were applied to the k_{ep} map). A similar observation can be made from the accuracy measure shown in Figure 6.11, the multi-resolution fractal analysis outperforms the conventional texture analysis measures in classifying pCR from non-pCR patients.

Effect of each level of decomposition: We decompose each parametric map down to 4 levels using wavelet analysis. We analyze the ability of features from each level of decomposition in predicting response to NACT. Figure 6.12 shows the accuracy values (for the training and testing

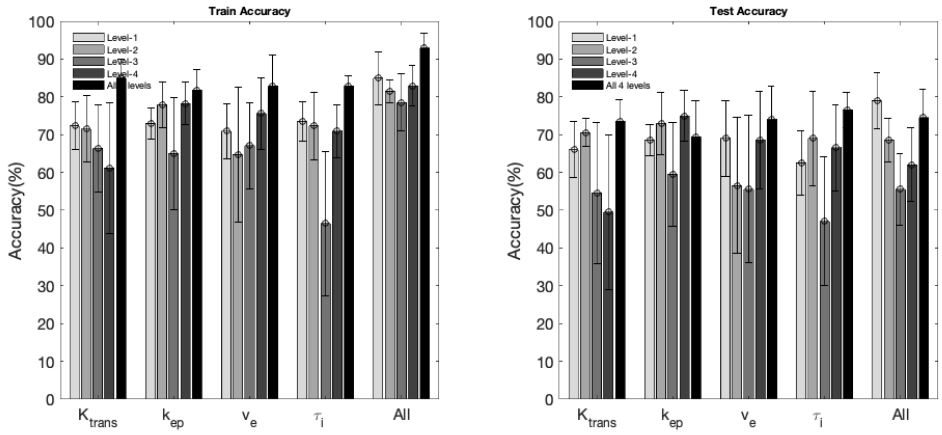


Figure 6.12: The accuracy values for classification of pCR from non-pCR patients in the training (a) and testing (b) datasets using feature vectors obtained from the K^{trans}, k_{ep}, v_e, and τ_i parametric maps. The final column 'All' represents the concatenated features from all four parametric maps. The error bars represent the standard deviation obtained over the 10 different partitions of train and test data. The increasing saturation of grey levels represent the first, second, third, and fourth level of decompositions, while the black column corresponds to the combination of features from all four levels.

datasets) for each level of decomposition of K^{trans}, k_{ep}, v_e, and τ_i parametric maps and the concatenated feature vector from all four parametric maps. The combination of features from all four levels is represented by the black bar. In the training dataset the combination of features from all four levels always outperforms features from individual levels, while in the testing dataset, in k_{ep} and 'All', level-3 and level-1 features outperform the combination of all features. Features from level-3 consistently seem to perform poorly.

Comparison of single and multi-resolution fractal analysis: The mean accuracy, sensitivity, specificity and ROC curves obtained by 10-fold cross-validation for single and multi-resolution fractal analysis are shown in Figure 6.13. Multi-resolution FD demonstrates better predictive performance than single-resolution FD in both the training and testing data sets. The multi-resolution FD estimated from voxel-based DCE-MRI parametric maps provides good early prediction of breast cancer response to NACT (ROC AUC > 0.8), and has better predictive ability than single-resolution FD, though the difference in ROC AUC was not statistically significant ($P = 0.15$) in this small data sets. The advantage of multi-resolution FD analysis may be due to its capability of filtering out irrelevant features and noise at different resolutions and simultaneously rendering more emphasis on distinct features.

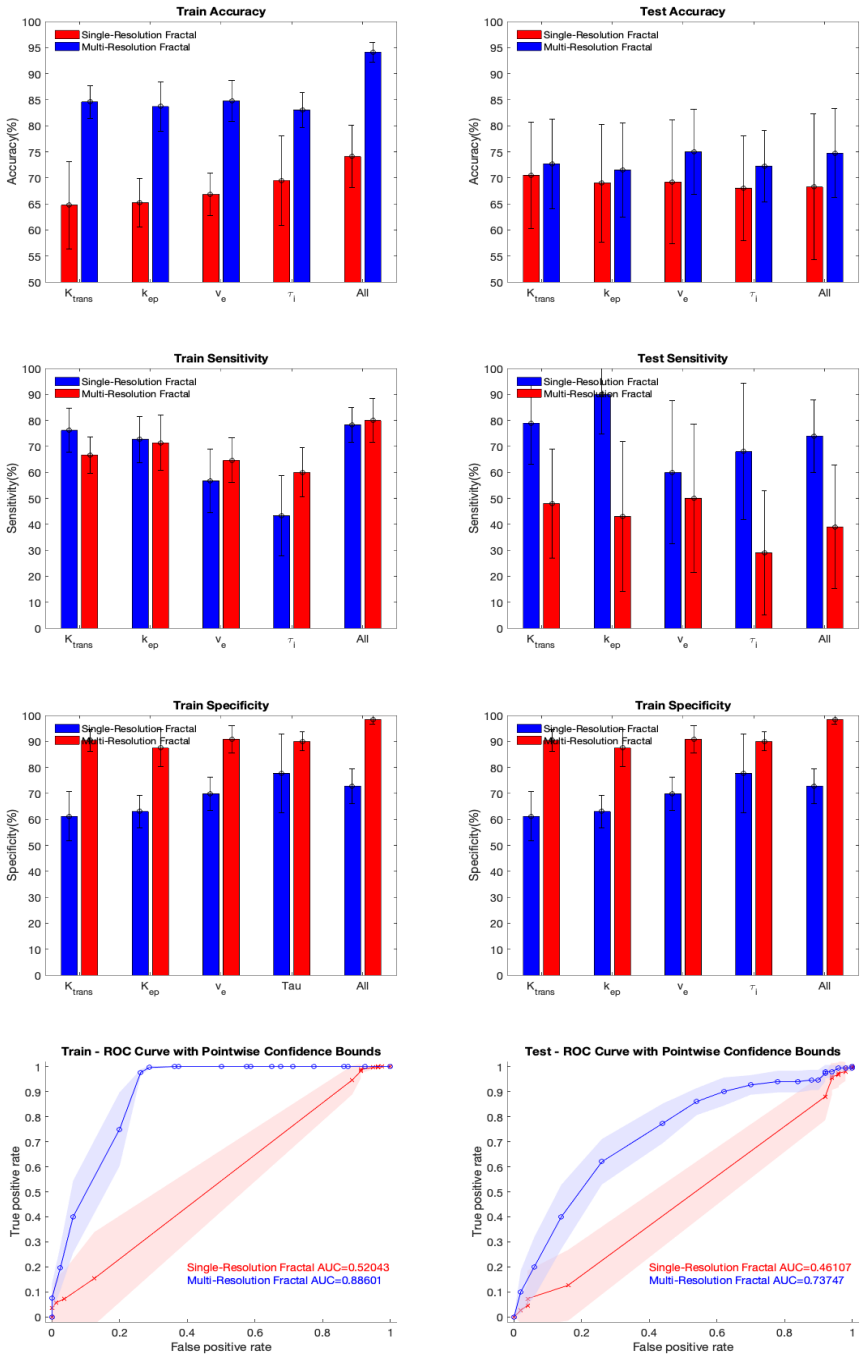


Figure 6.13: Performance comparison for Fractal and multi-resolution Fractal analysis. First row: accuracy, second row: sensitivity, third row: specificity and fourth row: ROC AUC. Left and right columns represent train and test results respectively.

Table 6.3: Classification accuracy of PCR vs. non-PCR scores using different parameters

Parameter	Visit 1 (%)	Visit 2 (%)
K^{trans}	75	66
k_{ep}	75	66
v_e	72	64
τ_i	70	68

6.5.2 Analysis using Convolutional Neural Networks

We examined the capability of a convolutional neural network as a combined feature extractor and classifier to classify pCR vs. non-pCR patients. We consider data from visit-1 and visit-2 separately. As we use a 2D CNN we classify individual slice from the parametric map instead of the whole 3D volume. In this analysis each slice from a parametric map of a patient is considered as a separate sample. For each visit, a CNN is built for each of the four parametric maps to classify slices as pCR or non-pCR. The classification accuracies obtained in classification of pCR vs. non-pCR using the four parametric maps from DCE-MRI scans are shown in Table 6.3. The features from the parametric maps K^{trans} and k_{ep} from visit-1 result in the best classification accuracy of 75%. We experimentally determine a dropout rate of 0.2 to work well for extracting features in this dataset.

The feature from the parametric maps from visit-1 result in better classification accuracies than those from visit-2. This may be attributed to the fact that, in response to the treatment the heterogeneity of the tumor is reduced, and therefore there is lesser information to gather from scans during visit-2. Considering the change in features between visit-1 and visit-2 as we did in the previous section may improve the classification performance rather than considering each visit separately. This cannot be done in 2-dimensional analysis, as the DCE-MRI scans from the two visits cannot be matched at a slice level. In further studies, 3D convolutional networks may be explored for this purpose.

Further, the non-pCR patients could be divided into three classes based on the residual cancer burden score. We perform four-class classification based on residual cancer burden scores, as accurate assessment of residual cancer burden after NACT can help in making an informed decision in considering breast conservation surgery versus mastectomy. But we observe that the accuracy for classification based on residual cancer burden is low. This may be due to the low sample size

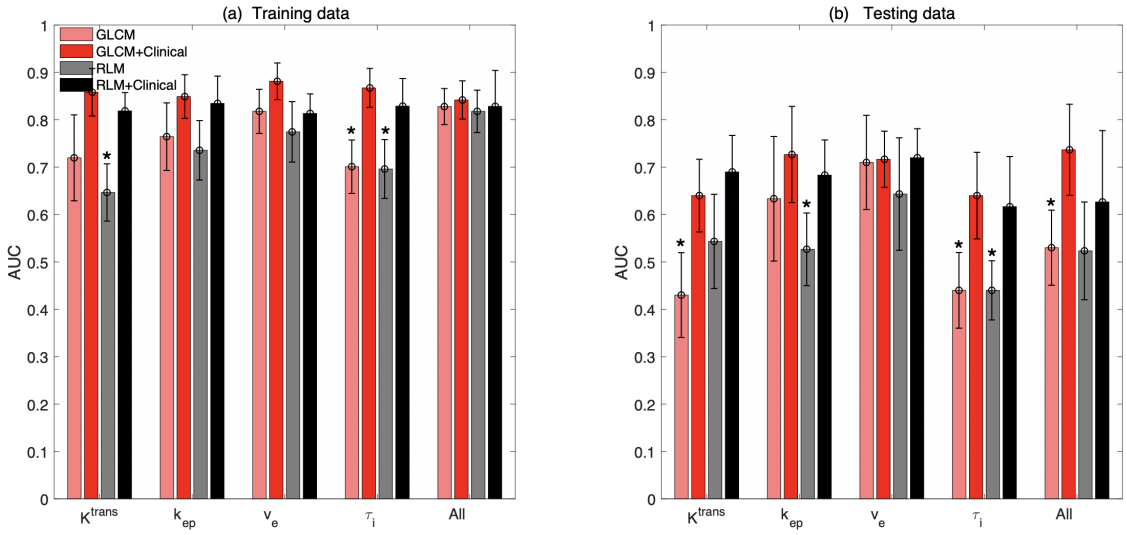


Figure 6.14: The mean ROC AUC values for classification of pCR vs. non-pCR in the training (a) and testing (b) datasets using the GLCM (light red) and RLM (grey) features extracted from PK parametric maps alone "All" represents the concatenated features from all four parametric maps) and in combination with clinical features: GLCM+Clinical (dark red) and RLM+Clinical (black). The error bars represent the standard deviation from 10 different partitions of the training and testing datasets. *: significant ($P < 0.05$) difference in AUC compared to combined features with clinical data.

per class, which probably over-fit the network. Smaller networks or different architectures need to be explored to obtain better performance.

6.5.3 Integration of DCE-MRI Texture Features with Clinical Data

It has been reported that molecular markers of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) are independently associated with pCR rate [82]. In this section, we evaluate the potential of integrating DCE-MRI texture features with clinical data (including the molecular markers) for early prediction of breast cancer pathologic response to NACT. To analyze the predictive performance of integrated DCE-MRI texture features and clinical features, the two feature vectors are concatenated and submitted to SVM. The clinical features consist of age, tumor type, tumor grade, ER, PR, HER2 status, and TNM (tumor, node, and metastasis) stage. The 12-dimensional GLCM features and 13-dimensional RLM features are each concatenated with the 9-dimensional clinical features to form new feature vectors of 21 and 22 dimensions, respectively. Critical ratio calculated using the Hanley and McNeil's formula is used to compare the AUC values, with statistical significance set at $P < 0.05$.

The mean ROC AUC values for classifying pCR vs. non-pCR using GLCM and RLM image

Table 6.4: ROC AUC values for prediction of pCR vs. non-pCR. In each row, the increase in AUC from "imaging features alone" to "imaging + clinical feature" is statistically significant ($P < 0.05$) based on analysis using the Hanley and McNeil method.(PK represents pharmacokinetic)

Dataset	Texture feature map and PK map	Imaging features alone	Imaging + Clinical features
Training	RLM, K^{trans}	0.647	0.819
	GLCM, τ_i	0.701	0.867
	RLM, τ_i	0.696	0.829
Testing	GLCM, K^{trans}	0.430	0.640
	RLM, k_{ep}	0.527	0.683
	GLCM, τ_i	0.440	0.640
	RLM, τ_i	0.440	0.617
	GLCM, All	0.530	0.737

features alone and in combination with the clinical data are shown in Figure 6.14. Addition of clinical features to image features improves predictive performances in both the training and testing datasets. Table 6.4 shows the AUC values for the pairs of texture feature and pharmacokinetic parameter for which the integration with clinical features increased the AUC values significantly ($P < 0.05$). Our results suggest that integration of DCE-MRI texture features with clinical data provide even better prediction of NACT response, with the improvement in predictive performance statistically significant for several imaging features.

6.6 Discussion

This preliminary study shows that multi-resolution fractal analysis has the potential to better capture the heterogeneity in the breast tumor vasculature as measured by DCE-MRI and that the extracted features from voxel-based DCE-MRI parametric maps are good early predictors of breast cancer response to NACT. In general, the concatenated features extracted from parametric maps of all the DCE-MRI parameters provide the best predictive performance. Multi-resolution analysis filters out irrelevant features and noise at different resolutions, rendering more emphasis on distinct features, and fractal analysis at each level appears to be able to capture these distinct

features. The GLCM and RLM features reflect the overall correlation between adjacent voxels in terms of second-order and higher-order statistical features, respectively [26]. For the small dataset used in this study, the generally higher AUC values from the multi-resolution fractal analysis when compared to GLCM and RLM methods, suggest that decomposing the texture may give further insights into the heterogeneity of the tumor microvasculature shown on DCE-MRI parametric maps and help capture the subtle variations in the texture which cannot be assessed by the single-resolution approach. However, this observation needs to be validated with a larger patient cohort. Consistent with the studies reporting mean parameter changes [8, 13 – 21], the results from this study provide further proof that changes in vascular perfusion represented by DCE-MRI imaging biomarkers are important features in identifying responders and non-responders at the early stage of NACT.

The AUC values from Figure 6.10, show that the single-resolution fractal features perform consistently well in prediction of response for both the training and testing sets, though not as well as the multi-resolution approach. The higher AUCs for fractal based features suggest that they provide a richer representation of the heterogeneity in the tumor when compared to GLCM and RLM methods. The low dimensionality ($d = 1$) of single-resolution fractal feature is less likely to cause over-fitting for the small sample size of our dataset and therefore could lead to a good discriminative model. This could be one of the reasons that contributed to its effectiveness. On the other hand, in spite of increased dimensionality ($d = 32$), multi-resolution fractal features exhibit better predictive performance, suggesting that analyzing heterogeneity at multiple resolutions provides a more comprehensive measure of the texture and thus increases the discriminative power of the feature. At each level of decomposition, the approximation coefficient (W_ϕ from Equation 2.7) represents the low-frequency component, which characterizes the coarse structure of the data, and the detail coefficients (W_ψ^i from Equation 2.8) represent the high-frequency components, which capture the discontinuities and singularities in the data. Therefore, combination of features from different scales and frequencies gives a richer representation of the overall underlying texture. The advantages of multi-resolution fractals can be expected to be even more significant when the dataset is large enough to offset their high dimensionality.

The tumor heterogeneity appears to be captured well at the first two levels of decomposition as shown by Figure 6.12. Each decomposition level analyzes the signal at a particular band of frequencies. Higher decomposition levels have better frequency resolution. The first level of decomposition encompasses the entire frequency band of the input data in its sub-volumes. Thereafter, we select the sub-volume with the highest fractal dimension and perform multi-resolution fractal analysis on that sub-band alone. By doing this we are effectively looking at finer frequency resolutions of the

selected sub-band frequencies alone. Few features from finer frequency resolutions exhibit lower discriminative power when considered in isolation, but when combined with features from other levels appears to enrich the representation and provide incremental improvement.

As shown in Table 6.2 for the test dataset, at fixed sensitivity, the higher AUC values from the multi-resolution fractal features generally result in higher specificity values compared to those from other features. It is important to have high sensitivities so that most pCR patients will be correctly identified and continue with the original or de-escalated NACT regimen. At 80% sensitivity, the $> 60\%$ specificity (except for the k_{ep} features) of the multi-resolution fractal features implies that were this method used in clinical care, more than half of the non-pCRs would be correctly classified after the first NACT cycle, potentially enabling alteration of treatment plans for these non-responders at the early stage of NACT to receive more personalized care.

The features extracted by convolutional neural networks from parametric maps from visit-1 demonstrate better classification accuracies than those from visit-2 as shown in 6.3. The features from k_{ep} and K^{trans} parametric maps from visit-1 could classify pCR vs. non-pCR patients with 75% accuracy. Rather than the current setting of using a slice of a parametric map from a single visit, using combination of parametric maps and comparing features from multiple visits (like we did in multi-resolution fractal analysis) can potentially better represent the variations in the classes and help improve the classification accuracy.

Addition of clinical features to image texture features increases the predictive capability in discriminating pCR vs. non-pCR compared to using imaging features alone as shown in Figure 6.14. In the emerging era of precision medicine, the "big data" approach of integrating imaging, proteomic, and genomic data is the way forward in management of cancer patients. In addition to validating our encouraging preliminary results with a larger patient cohort, other relevant data such as genetic test score and NACT regimen can be included in the predictive model for NACT response in future investigations.

6.7 Conclusion

In this chapter, we demonstrated that multi-resolution fractal analysis of voxel-based DCE-MRI parametric maps could be a promising tool for early prediction of breast cancer response to NACT. The multi-resolution fractal features generally have better predictive performances than those extracted with the more conventional methods of GLCM, RLM and single-resolution fractal analysis. Furthermore, compared to features extracted from individual DCE-MRI parametric maps, the use

of concatenated features from all DCE-MRI parameters further improved prediction of NACT response. Convolutional neural networks could extract reliable features to predict response to NACT using scans from just one visit, but a measure comparing parametric maps from two visits would be a stronger indicator than using maps from a single visit. Finally, we showed that addition of clinical features to image texture features further increases the predictive capability.

This study has several limitations. The first being the small size of the dataset used. The preliminary results obtained need to be evaluated on a larger patient cohort. Also due to the small size of the dataset, dimensionality increase in feature vectors impedes the performance of the classifier. Larger dataset can enable the choice of a richer feature vector from different levels in the multi-resolution fractal decomposition, which might consistently outperform the other features. Finally, the DCE-MRI parametric maps used for feature analysis were obtained with the shutter-speed model, which is not commonly used in pharmacokinetic analysis of DCE-MRI data. In future studies, parametric maps obtained with the widely used standard Tofts model [172, 171], which generates only the K^{trans}, v_e, and k_{ep} parameters and thus results in reduced dimensionality of the feature vector, can be used for feature extractions.

Chapter 7

Tracking: Vision and Sensor Fusion

7.1 Introduction

A sequence of images is a richer source of information than a still image as it captures the motion in the scene. Tracking the motion of objects can provide objective measurements characterizing their movement. In this chapter, we delve into the fourth component of the image analysis pipeline dealt with in this dissertation - tracking. Though a regular video camera can capture a scene with high spatial resolution, its temporal resolution might not be sufficient to capture subtle movements. Certain applications, such as tracking fidgety movements in infants, require analysis of subtle movements. Fidgety movements are defined as small circular motions of moderate speed with variable acceleration [51] and their absence in infants has shown to be a strong early marker for cerebral palsy [3]. Wearable sensors like accelerometer and gyroscope have high temporal resolution capable of tracking subtle movements, but lack the spatial information [120]. In this chapter, we develop a hybrid system that combines measurements from a basic camera and wearable sensors thus benefiting from the superior spatial and temporal resolution capabilities of each modality. Specifically, we use inertial measurement units, containing accelerometer, magnetometer and gyroscope unlike other motion analysis studies related to cerebral palsy where only one of these sensors was used [120]. The true 3D motion is estimated using an extended Kalman filter, which combines measurements from video and inertial measurement unit. Features extracted from the estimated motion are used to classify fidgety movements from non-fidgety movements using SVM.*

*The work in this chapter has been published in A. Machireddy, J. Van Santen, J. L. Wilson, J. Myers, M. Hadders-Algra, X. and Song, A video/IMU hybrid system for movement estimation in infants, in 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 730-733).

7.2 Background and Related Work

7.2.1 Diagnosis of Cerebral Palsy

It is estimated that around 15 million babies are born preterm (before 37 weeks of gestation) worldwide every year [132]. Preterm infants are at increased risk of developing neurological and motor impairments [204]. The most common motor impairment affecting preterm infants is cerebral palsy (CP) and is found in 4-20% of preterm infants based on their gestational age [7]. Cerebral palsy is a non-progressive disorder causing functional impairment of the brain. The neurological damage leads to complex dysfunctions affecting movement, motor skill and muscle co-ordination. Although the primary damage cannot be repaired, identifying CP at an early stage enables early intervention, which can help minimize resultant impairments [122]. Currently, the diagnosis of CP in infants is based on the assessment of their spontaneous "general movements", i.e., movements made while lying in supine position on a flat surface, not surrounded by any objects that obstruct movement or attract attention. The development of movements in early infancy is a good predictor of movement and cognitive performance later on [53]. Prechtl and Einspieler [51] studied the development of spontaneous movements in infants, and came up with Prechtl's general movement assessment, which can predict CP with sensitivity of 98% [27]. Notably, the absence of fidgety movements in infants between 2 to 4 months has shown to be a strong marker in identifying infants who will develop CP [3].

7.2.2 Automated Methods for Analyzing Motion

Although Prechtl's general movement assessment can predict CP with high accuracy, it is not widely adopted clinically as it requires clinicians to be trained specifically in the method and the judgments are subjective [3]. This led to the development of automated methods for analyzing motion in infants. An infant's limb movements can be captured by using a video camera or wearable sensors such as accelerometers, and analyzed to identify different movements [120]. Video based movement analysis gives high spatial resolution and is easy to interpret, but can be hampered by occlusion, and generally has low temporal resolution unless expensive special-purpose high frame rate cameras are used. To detect fidgety movements it is necessary to capture subtle changes in movements of the limbs, which may not be possible using video based methods. In contrast, accelerometer based methods have high temporal resolution, which can be helpful to track subtle changes, but they lack spatial information [120]. The automated methods currently used for analysis of motion in infants can be broadly classified into video-based, sensor-based and hybrid (video and sensor -based) methods.

7.2.3 Related Work

Video-Based Assessment

We first discuss methods used to analyze motion from videos. Meinecke et al. performed 3D motion analysis by placing reflective markers on infant and capturing motion using seven infrared cameras [122]. They extracted 53 quantitative parameters among which they found an optimal combination of eight parameters by cluster analysis. These optimal eight parameters were used as features for quadratic discriminant analysis to obtain an overall CP detection rate of 73%. But such 3D motion capture systems are costly, hard to set-up, and have high computational complexity, thus limiting their clinical use. Kanumera et al. also used reflective markers, but used a digital video camera to capture motion [90]. They examined different features extracted from the 2D-positional data from the video. They found that the movements of infants who develop CP were jerkier than those of normal infants. As only the 2D-positional data was considered, movement perpendicular to the camera was completely overlooked.

Adde et al. developed a general movements toolbox to detect fidgety movements in video recordings [2]. They used the 2D representation of movement over time, called the motiongram, to extract eight quantitative features. They achieved a sensitivity of 81.5% in classifying fidgety movements. Stahl et al. extracted motion information by applying optical flow [157]. Features were extracted using wavelet analysis and classified using SVM to achieve 93% accuracy in identifying fidgety movements.

But these studies used marker-less tracking, which could be less accurate as subtle differences in moving patterns cannot be detected due to low temporal resolution of a camera. The advantage of using standard video cameras is the reduced cost of the system and the set-up effort, enabling applications beyond the research setting, to regular clinics as well as allowing for continuous analysis even at home.

Sensor-Based Assessment

Next we discuss methods used to analyze motion using sensors. Gravem et al. used accelerometers to monitor movements in ten preterm infants to identify cramped synchronous general movements [68]. Their statistical models were able to predict cramped synchronous general movements with 70-90% accuracy. No attempts were made to predict CP and the age of the infants studied (30-43 weeks) was too old for predicting CP from general movements [71].

Heinze et al. extracted 32 features based on velocity and acceleration extracted from the accelerometer data for 19 healthy and 4 affected subjects [77]. They used decision trees to obtain a

detection rate between 88-92%. Philippi et al. used a magnetic tracking system to record movements [137, 91]. They extracted three kinematic features out of which repetitive movement in the upper limbs proved to be a good predictor of CP. But the accelerometer and magnetic tracking system used by the above two studies were wired and large in size, posing significant practical problems.

Hybrid Systems

Berge et al. proposed a software tool called ENIGMA (enhanced interactive general movement assessment) which helped visualize movement patterns in a video [23]. In their system, the motion was captured using a video camera and a set of six wired sensors. They proposed a periodicity feature for detecting fidgety movement, but did not provide any quantitative analysis on how well this feature performed.

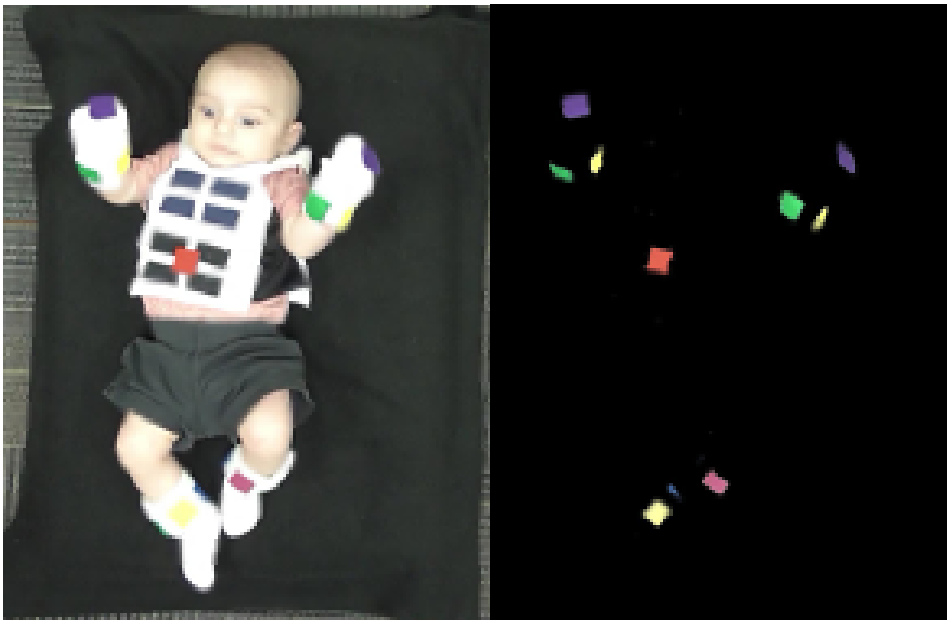
7.3 Proposed approach

To estimate the true 3D motion, we designed an extended Kalman filter based on the framework described in Section 2.4.2 by fusing the positional estimates from the camera with the inertial measurement unit signals.

7.3.1 Position Estimation from Video

We decided to use a fiducial-based motion tracking approach to track the movement of infant limbs. We sewed color patches on the wrist, ankle and chest bands holding the sensors as shown in Figure 7.1. Doing so would ensure that the patches would be stable with respect to the sensors. The locations of these patches on the image plane can be detected accurately since each patch has a unique and predetermined color. We placed five patches of different colors, one each on four sides of the wrist and one on top of the mitten so that we have at least one patch visible at all times, irrespective of the orientation of the hand. Since the relative positions of the color patches to one another are known, the position of the color patch on the sensor can be estimated, even when it is occluded. The sensor on the chest had only had only one colored patch on it.

We determined the accurate position of the limb by segmenting the colored patches by setting thresholds for each color. We used the HSV (hue, saturation, value) color space to segment the patches. The main advantage of HSV over RGB color space is that it separates color from intensity information [38]. We set thresholds, using the exact range of hue, saturation and value measures representing each color patch to segment them. We estimated the true 2D positions



(a) (b)

Figure 7.1: (a) Baby with the Shimmer sensors and color patches (b) detected color patches.

of the patches by back projecting the point on the image plane through the camera calibration matrix. The camera calibration matrix maps the pixel coordinates on the image to their true 3D world coordinates. It is calculated using the camera calibrator application in Matlab which uses the algorithm developed by Zhang [199]. We also use the color patches to obtain an estimate of depth (z dimension). As the true size of the color patch (w) and the focal length of the camera (f) are known, we use the triangular similarity to get its distance from the camera using the formula $D = (w * f)/p$, where p is the width of the patch on the image in pixels [158]. Therefore, this provides the 3D spatial position of the colored patch.

7.3.2 Synchronizing Data from the Two Modalities

Each sensor records its measurements locally and the camera captures video separately. In order to combine data from all sensors and the video camera, we need to synchronize them. The sampling rate for the video signal is 60 frames per second, while for the Shimmer sensor it was 256 Hz. To synchronize these two modalities, a buzzer connected to an additional Shimmer is used. It simultaneously produces a 4175 Hz audio signal and a voltage spike in the Shimmer signal. As all Shimmers are synchronized, the timing of voltage spike is looked up on the other Shimmers. The Goertzel algorithm [93] is used to detect the 4175Hz audio signal from the video, which is then

aligned with all the Shimmers.

7.3.3 Combining Measurements using Extended Kalman Filter

Segmenting the colored patches from the video gives the positional information regarding the limbs, while accelerometer, gyroscope and magnetometer give the acceleration, angular velocity and the magnetic field respectively. To track the movements of the limbs by combining information from the video and the sensors we use the EKF framework. As shown in Section 2.4.2, the EKF propagates an estimate of the system state x (e.g. position, orientation), given a sequence of observations z (e.g. Shimmer output, image features) using the equations $x_k = f(x_{k-1}) + \eta_k$ and $z_k = h(x_k) + \epsilon_k$, where f is the system model, h the observation model, and η and ϵ are the system and observation noise, respectively. Formulation of EKF requires definition of the state vector (x) and information of the model given by system model (f), observation model (h) and covariance matrices of the process (Q_k) and observation noise (R_k), each of which is described below.

State vector In order to track the limb we need its position and orientation at every time instant. We define the state vector as $x = [p\ v\ a\ q\ w]^T$, where p, v and a are the 3D position, velocity and acceleration, q is the quaternion describing rotation from world frame to sensor frame, and w is the angular velocity. A quaternion is a four-dimensional vector describing rotation. Here, world frame refers to the real-world coordinate system and the sensor frame refers to the coordinate system of the sensor. The quaternion q describes the rotation of the sensor with respect to the world.

System model The system model and process noise covariance matrix (Q_k) are given by Equations 7.1 and 7.2, where I is a 3×3 identity matrix, and Z, Z_1 and Z_2 are 3×3, 3×4 and 4×3 zero matrices respectively (Equations 7.3 and 7.4 are used in Equations 7.1 and 7.2). Similar to previous work on modeling human motion, we assume constant acceleration and angular velocity [96]. The actual variation is modeled by zero-mean white Gaussian noise with covariance matrices $\Sigma_a = \sigma_a^2 I$ and $\Sigma_w = \sigma_w^2 I$, respectively.

Observation model The measurements we obtain are the position of the limbs from the camera, acceleration from the accelorometer, angular velocity from the gyroscope and the magnetic field from the magnetometer. Using these measurements we define the observation vector as $y = [p_o\ a_o\ w_o\ m_o]^T$, where p_o is the 3D position estimated from video, and a_o, w_o and m_o are the measurements from accelerometer, gyroscope and magnetometer, respectively. The observation equations and the observation noise covariance matrix (R_k) are given by Equations 7.5 - 7.9. The

reason to adopt the EKF approach, instead of the Kalman filter, is the nonlinear nature of the measurement equations (due to the presence of rotation matrix R). It can be seen from Equation 7.6 that the observed acceleration is the sum of the true body acceleration and the gravitational acceleration (g) as perceived in the sensor frame. The rotation matrix R defined by Equation 7.10 projects the measurement from the real-world coordinates to the sensor coordinates. As the magnetic field in a room can be distorted due to the presence of other metal structures, we use a distortion compensation method proposed by Madgwick et al., wherein the earth's magnetic field is represented as $b = \left[\sqrt{h_x^2 + h_y^2} \;\; 0 \;\; h_z\right]$, where $h = Rm_o$, so that it is has the same inclination as the measurement [118].

State equation:

$$\begin{bmatrix} p_{k+1} \\ v_{k+1} \\ a_{k+1} \\ q_{k+1} \\ w_{k+1} \end{bmatrix} = \begin{bmatrix} I & dtI & \frac{1}{2}dt^2 I & Z_1 & Z \\ Z & I & dtI & Z_1 & Z \\ Z & Z & I & Z_1 & Z \\ Z_2 & Z_2 & Z_2 & I_4 + \frac{1}{2}S_w dt I_4 & Z_2 \\ Z & Z & Z & Z_1 & Z \end{bmatrix} \begin{bmatrix} p_k \\ v_k \\ a_k \\ q_k \\ w_k \end{bmatrix} + \eta_k \tag{7.1}$$

State noise covariance matrix:

$$Q_k = \begin{bmatrix} \frac{dt^5}{20}\Sigma_a & \frac{dt^4}{8}\Sigma_a & \frac{dt^3}{6}\Sigma_a & Z_1 & Z \\ \frac{dt^4}{8}\Sigma_a & \frac{dt^3}{3}\Sigma_a & \frac{dt^2}{2}\Sigma_a & Z_1 & Z \\ \frac{dt^3}{6}\Sigma_a & \frac{dt^2}{2}\Sigma_a & dt\Sigma_a & Z_1 & Z \\ Z_2 & Z_2 & Z_2 & \frac{dt^2}{4}S_q\Sigma_w S_q^T & Z_2 \\ Z & Z & Z & Z_1 & \Sigma_w \end{bmatrix} \tag{7.2}$$

$$S_q = \begin{bmatrix} -q_1 & -q_2 & -q_3 \\ q_0 & -q_3 & q_2 \\ q_3 & q_0 & -q_2 \\ -q_2 & q_1 & q_0 \end{bmatrix} \tag{7.3}$$

$$S_w = \begin{bmatrix} 0 & -w_x & -w_y & -w_z \\ w_x & 0 & -w_z & -w_y \\ w_y & w_z & 0 & -w_x \\ w_z & -w_y & w_x & 0 \end{bmatrix} \tag{7.4}$$

Observation equations:

$$p_o = p_k + \epsilon_p \tag{7.5}$$

$$a_o = R(a_k + g) + \epsilon_a \tag{7.6}$$

$$w_o = w_k + \epsilon_w \tag{7.7}$$

$$m_o = R(b) + \epsilon_m \tag{7.8}$$

Observation noise covariance matrix:

$$R_k = \begin{bmatrix} \sigma^2_{p_o} I & Z & Z & Z \\ Z & \sigma^2_{p_o} I & Z & Z \\ Z & Z & \sigma^2_{p_o} I & Z \\ Z & Z & Z & \sigma^2_{p_o} I \end{bmatrix} \tag{7.9}$$

Rotation matrix:

$$R = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1 q_2 + q_0 q_3) & 2(q_1 q_3 - q_0 q_2) \\ 2(q_1 q_2 - q_0 q_3 & q_0^2 1 q_1^2 + q_2^2 - q_3^2 & 2(q_2 q_3 + q_0 q_1) \\ 2(q_1 q_3 - q_0 q_2) & 2(q_2 q_3 - q_0 q_1 & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix} \tag{7.10}$$

Processing schedule As the two modalities are sampled at different rates, a multi-rate filtering strategy is employed. The exact time instances of the visual and shimmer measurements are known, therefore the time difference between successive measurements (dt) is known. The state equations are updated for every time instant that has a measurement from either modality, and the rows of the observation vector for the modality missing the measurement at that instant are set to zero [110].

The proposed hybrid system provides an estimate of the 3D position, velocity, acceleration, orientation and angular velocity of the infant's limbs with high spatial and temporal resolution. As fidgety movements are characterized by their velocity and acceleration, we use the estimated 3D velocity, acceleration, angular velocity and their magnitudes to predict fidgety movement.

7.4 Methods

7.4.1 Participants

We collected the data at our lab in Oregon Health and Science University in Portland. Twenty infants between the ages of 2-4 months were recruited for this study. A research assistant obtained a written consent from all parents. We placed the infants in the supine position on a mattress with sensors attached, and placed a stationary video camera above, which captured movement in

the horizontal plane. We recorded a video for 30 minutes when the infant was in an active wakeful state.

7.4.2 Sensor Recording

We use five *Shimmer3* sensors from Shimmer Inc. [30], and attach them to the left leg, left hand, right leg, right hand and the chest using soft wrist and ankle bands, and a vest as shown in Figure 7.1. *Shimmer3* sensors have an accelerometer, a gyroscope and a magnetometer. A sampling rate of 256Hz is used. To calibrate the sensor, the constant bias in the accelerometer and gyroscope are determined beforehand from static recordings and subtracted from the measurements.

7.4.3 Camera Recording

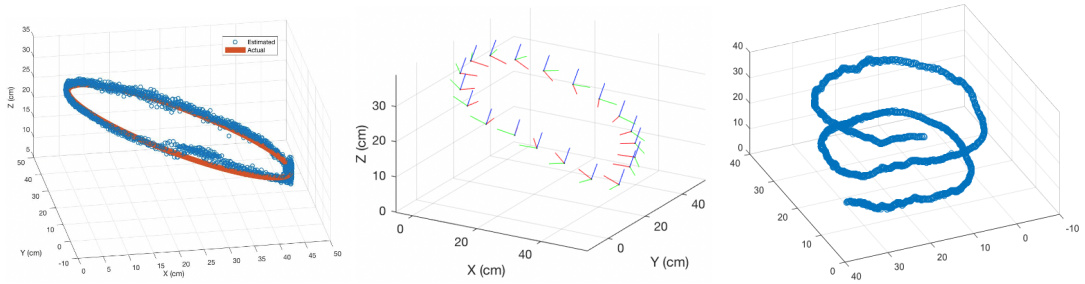
We use a down-pointing video camera (Panasonic HC-V550) and calibrate it using the algorithm developed by Zhang [199] to obtain intrinsic parameters (focal length, scaling and skewing factors along the horizontal and vertical axis of the image plane) and extrinsic parameters (rotation and translation of the camera coordinates relative to the world coordinates). To avoid the difficulty of calibrating camera arrays and reducing distraction to the infants, we use only one camera, which captures only the 2D projection of true 3D motion.

7.5 Results

7.5.1 3D Motion Estimation using Simulated Data

We first test our model using simulated data so that the exact ground truth motion is known. To evaluate the accuracy of the EKF model in predicting the position and orientation and in particular the prediction of depth, we made a model of human arm using plywood, and simulated circular motion by rotating it using a drill. Knowing the length of the arm, exact position and angle of the drill, the ground truth movement is determined. A *Shimmer3* sensor is placed on the plywood-arm with a color marker and the motion is recorded using a camera. This motion is estimated using EKF as explained above.

From Figure 7.2a and 7.2b, it can be seen that the EKF framework is able to correctly predict the position and orientation of the plywood-arm during the simulated circular motion. We simulate other movements such as the 3D spiral motion and the model is able to estimate the positions accurately as seen from Figure 7.2c.



(a) (b) (c)

Figure 7.2: (a) The ground truth (red) and estimated (blue) position of the sensor on plywood arm during circular motion; (b) estimated orientation at different time points during circular motion. (c) 3D Spiral motion.

7.5.2 3D Motion Estimation in Infants to Detect Fidgety Movements

The videos of two at-risk infants and three not at-risk infants are analyzed by an expert to mark time intervals with fidgety movements and non-fidgety movements. The motions of different limbs are estimated using the hybrid system. The estimated velocity, acceleration, angular velocity and their magnitude values are chosen to form a 12 dimensional feature at each time instant. Equal lengths of fidgety and non-fidgety movement segments are chosen to form a balanced dataset. The dataset consists of 100 segments of fidgety and non-fidgety movements each. A 10-fold classification using SVM on data from all limbs results in 84% accuracy in classifying fidgety from non-fidgety movements. A 10-fold classification on data from a single limb (i.e. train and test data from same limb) results in an accuracy of 90%, while using train and test data from different limbs results in 70% classification accuracy.

7.6 Conclusion

A novel method combining video and IMU sensor inputs to estimate 3D infant body movements was presented. A new approach to estimate depth by using markers captured by a video was demonstrated. Accelerometer, gyroscope and magnetometer features are combined with the positional information from video using EKF to derive the 3D position and orientation of the limbs of infants. The proposed method was shown to be able to accurately estimate 3D motion.

Based on the estimated motion, fidgety movements were classified with over 84% accuracy. The results in detecting fidgety movements are promising on the small dataset used, but need to be further assessed on a larger population. Currently we assume constant acceleration and angular velocity in the system model of EKF. A dynamic model characteristic of infant's limb movements

can give better features and deeper insights into the behavior of the fidgety movement. Addition of dynamic features will further help capture the temporal behavior of fidgety movement and help increase the classification accuracy. The applicability of fidgety movements in predicting cerebral palsy needs to be evaluated in further studies by recruiting high-risk population and carrying out a long-term neurological study.

Chapter 8

Summary and Future Work

The primary goal of this dissertation was to develop machine learning methods to aid in automatic processing of medical images. To this end we developed multiple methods addressing different aspects of the image processing pipeline while tacking specific medical problems. In this chapter we summarize the proposed methods and provide future directions for each of the problems.

8.1 Point Set Registration

In Chapter 3, we proposed a new point set registration algorithm that preserved both the global and local structure of the point set. We proposed a probability density estimation framework to align the point sets by incorporating the local structural relationships among neighboring points. We defined a new local structural measure that represents distances between points as probability distributions and calculated the Kullback-Leibler distance between the distributions to define a new measure of local structural similarity. This measure defined the membership probabilities for the GMMs in the density estimation framework. This process resulted in different membership probabilities of the points based on their local structure unlike previous approach where fixed probability values were assigned [116]. This method could be used for dimensions greater than two, since the measure depends only on the distance between the points. The experimental results demonstrated that our approach outperformed other point set registration algorithms, even when the point sets are degraded by various levels of noise, outliers, deformation, occlusion, and rotation.

We observed that the proposed local structural measure reduced the computational time and error in pairwise point set registration. The next step is to apply this measure in group-wise point set registration setting to see if it results in comparable outcome. The present group-wise registration algorithms take a few hours to obtain results on small datasets [141]. Reducing computational time in this setting would be a prominent contribution. Further in group-wise registration of sequence of images (e.g. Heart MRI sequence), there is temporal information that

can be modeled into the registration algorithm, to give better registration results.

Noise and outliers are presently accounted by a uniform distribution. This can probably be improved by modeling noise and outliers using better fitting models. In addition to improving registration performance, a local structural measure can also help in modeling noise and outliers. When we see an image containing an object along with outliers and noise, we may be able to roughly estimate which of the points belong to the object and which among them are outliers. We can identify two different distributions of points. If we come up with a measure to model these two distributions mathematically, then the error in the registration process can be further reduced. The regularization in the proposed algorithm only makes sure that the motion is coherent. A regularization method conserving local structure along with motion coherence can help achieve better registration, preserving the overall structure. Further, the proposed algorithm is still slow when applied to large datasets. There is a need to come up with a new framework that can speed up the registration process to greater extent.

8.2 Prostate Tumor Segmentation

In Chapter 4, we experimented with various modification to the U-Net architecture to improve its performance in segmenting tumors in MRI images of human prostates. As the tumor occupied a small portion of the image, there was imbalance in the tumor and non-tumor classes. We showed that the Tversky loss, which was specially designed for datasets with imbalanced classes, improved the segmentation performance as the weight of the false positive and false negative terms could be controlled unlike the Dice loss function. We further modified the U-Net architecture by adding attention gates, using auto-encoder style regularization and using deep supervision. The attention gates pruned the feature maps to retain only features relevant for tumor detection, thus reduced false positive predictions. The auto-encoder based regularization helped when the information in the image was sparse but degraded the segmentation performance when the image contained rich diverse information as the features required for reconstruction of rich data could be different from those required for segmentation. Deep-supervision forced the features to be highly discriminative and therefore increased the segmentation performance and also lead to faster convergence.

We analyzed the performance of each modification separately. However, a combination of these modifications can lead to further improvement of the segmentation performance. We experimented two ways of combining data from T2 and ADC images, but did not improve the performance over using ADC images alone. New methods of combining data from multiple scans can be explored. DCE-MRI images also contain information pertinent to tumor identification in prostate and the

ability of its features in segmenting tumors can be explored in further studies.

8.3 Cell Segmentation by Combining Learning and Tracking Based Approaches

In Chapter 5, we presented a framework for segmentation of cells and sub-cellular ultrastructure in electron microscopy images of tumor biopsies. The intracellular organelles were well distinguished from the surrounding ultrastructures enabling accurate segmentation by training deep learning models on sparse manual labels. Cell segmentation on the other hand was a complex task due to the lack of clear boundaries separating cells in the cancer tissue. We used a multi-pronged approach combining segmentation and tracking strategies for cell segmentation. Our proposed method cut down the time taken for cell segmentation from 8 months to 1-2 days. The proposed segmentation of electron microscopy images enabled interpretative rendering and quantitative characterization of biologically relevant features. These features were capable of capturing the differences between the tissue samples, and could be potentially linked to biologically or clinically relevant variables such as patient drug response in downstream analysis.

We trained the segmentation network with only the images containing the ground truth labels and did not utilize the unlabeled images in the 3D stack. The unlabeled images could be a rich source of information and could be used for self-supervised pre-training or labeled and unlabeled images could be used together for semi-supervised training. This could further improve the segmentation performance. Due to intensity variations caused by acquition parameters and the variability of features among datasets the generalizability of the segmentation models is currently limited. In order to have an automatic segmentation framework that does not necessitate sparse manual labeling for every new dataset, it is necessary to have a model that has enough representative training data. Our initial trials to segment individual 3D volumes by sparsely labeling each volume is a step towards building a dataset large enough to capture the variability among different organelles and the variability caused by different image acquisition settings. A model trained on a larger dataset could generalize to a new dataset without requiring any manual labeling. Style transfer techniques could be explored to transfer the style of the new dataset to be similar to the images the network was trained on. Further, associating objects with transformers style approach [189] could be adopted as a single step solution for cell segmentation.

8.4 Multi-Resolution Fractal Analysis of DCE-MRI Parametric Maps for Early Prediction of NACT Response

In Chapter 6, we characterized the texture of DCE-MRI parametric maps with a new spectral texture analysis method called multi-resolution fractal analysis, to evaluate its potential in early prediction of breast cancer pathologic response to chemotherapy. Multi-resolution analysis filtered out irrelevant features and noise at different resolutions, rendering more emphasis on distinct features, and fractal analysis at each level captured these distinct features. The multi-resolution fractal features generally had better predictive performances than those extracted with the more conventional methods of GLCM, RLM and single-resolution fractal analysis. Furthermore, compared to features extracted from individual DCE-MRI parametric maps, the use of concatenated features from all DCE-MRI parameters generally further improved prediction of NACT response. Addition of clinical features to image texture features, GLCM and RLM, increased predictive capability in discriminating pCR vs. non-pCR compared to using imaging features alone.

The preliminary results obtained need to be evaluated on a larger patient cohort. Also due to the small size of the dataset, dimensionality increase in feature vectors impedes the performance of the classifier. Larger dataset can enable the choice of a richer feature vector from different levels in the multi-resolution fractal decomposition, which might consistently outperform the other features. Further, the DCE-MRI parametric maps used for feature analysis were obtained with the Shutter-Speed model, which is not commonly used in pharmacokinetic analysis of DCE-MRI data. In future studies, parametric maps obtained with the widely used standard Tofts model [172, 171], which generates only the K^{trans}, v_e, and k_{ep} parameters and thus results in reduced dimensionality of the feature vector, can be used for feature extractions.

8.5 Tracking Infant Limb Movements

In Chapter 7, we developed a hybrid tracking system, that combined measurements from a basic camera and wearable sensors, thus benefiting from the superior spatial and temporal resolution capabilities of either modality, to accurately track motion of infant limbs. Features extracted from the estimated motion classified fidgety movements from non-fidgety movements using Support Vector Machine. The proposed approach could not only accurately estimate trajectories, including complex ones, but could also – and with good accuracy given the dearth of training data – detect fidgety movements.

The results in detecting fidgety movements are promising on the small dataset used, but need

to be further assessed on a larger population. A dynamic model characteristic of infant's limb movements can give better features and deeper insights into the behavior of the fidgety movement. Addition of dynamic features will further help capture the temporal behavior of fidgety movement and aid in increasing the classification accuracy. The applicability of fidgety movements in predicting cerebral palsy needs to be evaluated in further studies by recruiting high-risk population and carrying out a long-term neurological study.

Bibliography

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. {TensorFlow}: A system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

[2] L. Adde, J. L. Helbostad, A. R. Jensenius, G. Taraldsen, and R. Støen. Using computer-based video analysis in the study of fidgety movements. *Early human development*, 85(9):541–547, 2009.

[3] L. Adde, M. Rygg, K. Lossius, G. K. Øberg, and R. Støen. General movement assessment: predicting cerebral palsy in clinical practise. *Early human development*, 83(1):13–18, 2007.

[4] M.-L. W. Ah-See, A. Makris, N. J. Taylor, M. Harrison, P. I. Richman, R. J. Burcombe, J. J. Stirling, J. A. d'Arcy, D. J. Collins, M. R. Pittam, et al. Early changes in functional dynamic magnetic resonance imaging predict for pathologic response to neoadjuvant chemotherapy in primary breast cancer. *Clinical Cancer Research*, 14(20):6580–6589, 2008.

[5] I. Ahmed and G. Jeon. A real-time person tracking system based on siammask network for intelligent video surveillance. *Journal of Real-Time Image Processing*, 18(5):1803–1814, 2021.

[6] O. S. Al-Kadi, D. Y. Chung, R. C. Carlisle, C. C. Coussios, and J. A. Noble. Quantification of ultrasonic texture intra-heterogeneity via volumetric stochastic modeling for tissue characterization. *Medical image analysis*, 21(1):59–71, 2015.

[7] P.-Y. Ancel, F. Livinec, B. Larroque, S. Marret, C. Arnaud, V. Pierrat, M. Dehan, B. Escande, A. Burguet, G. Thiriez, et al. Cerebral palsy among very preterm children in relation to gestational age and neonatal ultrasound abnormalities: the epipage cohort study. *Pediatrics*, 117(3):828–835, 2006.

[8] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan. Medical image analysis using convolutional neural networks: a review. *Journal of medical systems*, 42(11):1–13, 2018.

[9] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961, 2019.

[10] Y. Artan, D. L. Langer, M. A. Haider, T. H. Van der Kwast, A. J. Evans, M. N. Wernick, and I. S. Yetik. Prostate cancer segmentation with multispectral mri using cost-sensitive conditional random fields. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 278–281. IEEE, 2009.

[11] A. I. Baba and C. Câtoi. Tumor cell morphology. In *Comparative oncology*. The Publishing House of the Romanian Academy, 2007.

[12] R. Baghban, L. Roshangar, R. Jahanban-Esfahlan, K. Seidi, A. Ebrahimi-Kalan, M. Jaymand, S. Kolahian, T. Javaheri, and P. Zare. Tumor microenvironment complexity and therapeutic implications at a glance. *Cell Communication and Signaling*, 18(1):1–19, 2020.

[13] I. Banerjee, S. Malladi, D. Lee, A. Depeursinge, M. Telli, J. Lipson, D. Golden, and D. L. Rubin. Assessing treatment response in triple-negative breast cancer from quantitative image analysis in perfusion magnetic resonance imaging. *Journal of medical imaging*, 5(1):011008, 2017.

[14] I. Bankman. *Handbook of medical image processing and analysis*. Elsevier, 2008.

[15] A. Baraldi and F. Panniggiani. An investigation of the textural characteristics associated with gray level cooccurrence matrix statistical parameters. *IEEE transactions on geoscience and remote sensing*, 33(2):293–304, 1995.

[16] R. Barnes, C. Lehman, and D. Mulla. Priority-flood: An optimal depression-filling and watershed-labeling algorithm for digital elevation models. *Computers & Geosciences*, 62:117–127, 2014.

[17] T. Barrett and M. A. Haider. The emerging role of mri in prostate cancer active surveillance and ongoing challenges. *American Journal of Roentgenology*, 208(1):131–139, 2017.

[18] N. Bedard, M. Pierce, A. El-Naggar, S. Anandasabapathy, A. Gillenwater, and R. Richards-Kortum. Emerging roles for multimodal optical imaging in early cancer detection: a global challenge. *Technology in cancer research & treatment*, 9(2):211–217, 2010.

[19] T. Behrens, K. Rohr, and H. S. Stiehl. Robust segmentation of tubular structures in 3-d medical images by parametric object detection and tracking. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 33(4):554–561, 2003.

[20] I. Belevich, M. Joensuu, D. Kumar, H. Vihinen, and E. Jokitalo. Microscopy image browser: a platform for segmentation and analysis of multidimensional datasets. *PLoS biology*, 14(1):e1002340, 2016.

[21] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence*, 24(4):509–522, 2002.

[22] Y. Bentoutou, N. Taleb, K. Kpalma, and J. Ronsin. An automatic image registration for applications in remote sensing. *IEEE transactions on geoscience and remote sensing*, 43(9):2127–2137, 2005.

[23] P. R. Berge, L. Adde, G. Espinosa, and Ø. Stavdahl. Enigma–enhanced interactive general movement assessment. *Expert systems with applications*, 34(4):2664–2672, 2008.

[24] P. J. Besl and N. D. McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992.

[25] S. Beucher. The watershed transformation applied to image segmentation. *Scanning Microscopy*, 1992(6):28, 1992.

[26] H. Bonnefoi, S. Litière, M. Piccart, G. MacGrogan, P. Fumoleau, E. Brain, T. Petit, P. Rouanet, J. Jassem, C. Moldovan, et al. Pathological complete response after neoadjuvant chemotherapy is an independent predictive factor irrespective of simplified breast cancer intrinsic subtypes: a landmark and two-step approach analyses from the eortc 10994/big 1-00 phase iii trial. *Annals of oncology*, 25(6):1128–1136, 2014.

[27] M. Bosanquet, L. Copeland, R. Ware, and R. Boyd. A systematic review of tests to predict cerebral palsy in young children. *Developmental Medicine & Child Neurology*, 55(5):418–426, 2013.

[28] N. M. Braman, M. Etesami, P. Prasanna, C. Dubchuk, H. Gilmore, P. Tiwari, D. Plecha, and A. Madabhushi. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast dce-mri. *Breast Cancer Research*, 19(1):1–14, 2017.

[29] A. M. Bronstein, M. M. Bronstein, A. M. Bruckstein, and R. Kimmel. Analysis of two-dimensional non-rigid shapes. *International Journal of Computer Vision*, 78(1):67–88, 2008.

[30] A. Burns, B. R. Greene, M. J. McGrath, T. J. O'Shea, B. Kuris, S. M. Ayer, F. Stroiescu, and V. Cionca. Shimmer$^{\mathrm{TM}}$–a wireless sensor platform for noninvasive biomedical research. *IEEE Sensors Journal*, 10(9):1527–1534, 2010.

[31] H. B. Carter, P. C. Albertsen, M. J. Barry, R. Etzioni, S. J. Freedland, K. L. Greene, L. Holmberg, P. Kantoff, B. R. Konety, M. H. Murad, et al. Early detection of prostate cancer: Aua guideline. *The Journal of urology*, 190(2):419–426, 2013.

[32] I. Chan, W. Wells III, R. V. Mulkern, S. Haker, J. Zhang, K. H. Zou, S. E. Maier, and C. M. Tempany. Detection of prostate cancer by integration of line-scan diffusion, t2-mapping and t2-weighted magnetic resonance imaging; a multichannel statistical classifier. *Medical physics*, 30(9):2390–2398, 2003.

[33] A. Chatterjee and J. K. Erban. Neoadjuvant therapy for treatment of breast cancer: the way forward, or simply a convenient option for patients? *Gland surgery*, 6(1):119, 2017.

[34] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[35] V. H. Chen, V. Mouraviev, J. M. Mayes, L. Sun, J. F. Madden, J. W. Moul, and T. J. Polascik. Utility of a 3-dimensional transrectal ultrasound-guided prostate biopsy system for prostate cancer detection. *Technology in cancer research & treatment*, 8(2):99–103, 2009.

[36] W.-H. Chen, G.-F. Luo, and X.-Z. Zhang. Recent advances in subcellular targeted cancer therapy based on functional materials. *Advanced Materials*, 31(3):1802725, 2019.

[37] Z. Chen and S. Haykin. On different facets of regularization theory. *Neural Computation*, 14(12):2791–2846, 2002.

[38] H.-D. Cheng, X. H. Jiang, Y. Sun, and J. Wang. Color image segmentation: advances and prospects. *Pattern recognition*, 34(12):2259–2281, 2001.

[39] H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 2, pages 44–51. IEEE, 2000.

[40] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2-3):114–141, 2003.

[41] D. R. Coman and T. F. Anderson. A structural difference between the surfaces of normal and of carcinomatous epidermal cells. *Cancer Research*, 15(8):541–543, 1955.

[42] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[43] S. P. Dakua, J. Abinahed, A. Zakaria, S. Balakrishnan, G. Younes, N. Navkar, A. Al-Ansari, X. Zhai, F. Bensaali, and A. Amira. Moving object tracking in clinical scenarios: application to cardiac surgery and cerebral aneurysm clipping. *International Journal of Computer Assisted Radiology and Surgery*, 14(12):2165–2176, 2019.

[44] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, 36(5):961–1005, 1990.

[45] B. D. de Senneville, F. Z. Khoubai, M. Bevilacqua, A. Labedade, K. Flosseau, C. Chardot, S. Branchereau, J. Ripoche, S. Cairo, E. Gontier, et al. Deciphering tumour tissue organization by 3d electron microscopy and machine learning. *Communications biology*, 4(1):1–10, 2021.

[46] C. Denais and J. Lammerding. Nuclear mechanics in cancer. *Cancer biology and the nuclear envelope*, pages 435–470, 2014.

[47] K. Doi. Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology. *Physics in Medicine & Biology*, 51(13):R5, 2006.

[48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[49] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The importance of skip connections in biomedical image segmentation. In *Deep learning and data labeling for medical applications*, pages 179–187. Springer, 2016.

[50] M. Egeblad, E. S. Nakasone, and Z. Werb. Tumors as organs: complex tissues that interface with the entire organism. *Developmental cell*, 18(6):884–901, 2010.

[51] C. Einspieler and H. F. Prechtl. Prechtl's assessment of general movements: a diagnostic tool for the functional assessment of the young nervous system. *Mental retardation and developmental disabilities research reviews*, 11(1):61–67, 2005.

[52] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.

[53] T. Fjørtoft, K. H. Grunewaldt, G. C. C. Løhaugen, S. Mørkved, J. Skranes, and K. A. I. Evensen. Assessment of motor behaviour in high-risk-infants at 3 months predicts motor and cognitive outcomes in 10 years old children. *Early human development*, 89(10):787–793, 2013.

[54] D. Fleet and Y. Weiss. Optical flow estimation. In *Handbook of mathematical models in computer vision*, pages 237–257. Springer, 2006.

[55] S. Fulda, L. Galluzzi, and G. Kroemer. Targeting mitochondria for cancer therapy. *Nature reviews Drug discovery*, 9(6):447–464, 2010.

[56] F. Galli, J. V. Aguilera, B. Palermo, S. N. Markovic, P. Nisticò, and A. Signore. Relevance of immune cell and tumor microenvironment imaging in the new era of immunotherapy. *Journal of Experimental & Clinical Cancer Research*, 39(1):1–21, 2020.

[57] M. Gallowy. Texture analysis using gray level run length. *Comput Graph Image Process*, 4:172–179, 1975.

[58] S. Ge and G. Fan. Non-rigid articulated point set registration with local structure preservation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–133, 2015.

[59] F. Geisslinger, M. Müller, A. M. Vollmar, and K. Bartel. Targeting lysosomes in cancer as promising strategy to overcome chemoresistance—a mini review. *Frontiers in oncology*, page 1156, 2020.

[60] S. Ghai and M. A. Haider. Multiparametric-mri in diagnosis of prostate cancer. *Indian journal of urology: IJU: journal of the Urological Society of India*, 31(3):194, 2015.

[61] S. Giannarou and T. Stathaki. Shape signature matching for object identification invariant to image transformations and occlusion. In *International Conference on Computer Analysis of Images and Patterns*, pages 710–717. Springer, 2007.

[62] L. A. Giannuzzi et al. *Introduction to focused ion beams: instrumentation, theory, techniques and practice*. Springer Science & Business Media, 2004.

[63] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural computation*, 7(2):219–269, 1995.

[64] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[65] S. Gold, A. Rangarajan, C.-P. Lu, S. Pappu, and E. Mjolsness. New algorithms for 2d and 3d point matching: pose estimation and correspondence. *Pattern recognition*, 31(8):1019–1031, 1998.

[66] D. I. Golden, J. A. Lipson, M. L. Telli, J. M. Ford, and D. L. Rubin. Dynamic contrast-enhanced mri-based biomarkers of therapeutic response in triple-negative breast cancer. *Journal of the American Medical Informatics Association*, 20(6):1059–1066, 2013.

[67] R. Gonzalez. *Digital image processing*. Prentice Hall, Upper Saddle River, N.J, 2008.

[68] D. Gravem, M. Singh, C. Chen, J. Rich, J. Vaughan, K. Goldberg, F. Waffarn, P. Chou, D. Cooper, D. Reinkensmeyer, et al. Assessment of infant movement with a compact wireless accelerometer system. *Journal of Medical Devices*, 6(2), 2012.

[69] A. Gulli and S. Pal. *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[70] Y. Guo, S. Ruan, P. Walker, and Y. Feng. Prostate cancer segmentation from multiparametric mri based on fuzzy bayesian model. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 866–869. IEEE, 2014.

[71] M. Hadders-Algra. General movements: a window for early identification of children at high risk for developmental disorders. *The Journal of pediatrics*, 145(2):S12–S18, 2004.

[72] J. A. Hanley and B. J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843, 1983.

[73] R. M. Haralick, K. Shanmugam, and I. H. Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[74] K. Harezlak and P. Kasprowski. Application of eye tracking in medicine: A survey, research issues and challenges. *Computerized Medical Imaging and Graphics*, 65:176–190, 2018.

[75] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[76] L. Heinrich, D. Bennett, D. Ackerman, W. Park, J. Bogovic, N. Eckstein, A. Petruncio, J. Clements, S. Pang, C. S. Xu, et al. Whole-cell organelle segmentation in volume electron microscopy. *Nature*, 599(7883):141–146, 2021.

[77] F. Heinze, K. Hesels, N. Breitbach-Faller, T. Schmitz-Rode, and C. Disselhorst-Klug. Movement analysis by accelerometry of newborns and infants for the early detection of movement disorders due to infantile cerebral palsy. *Medical & biological engineering & computing*, 48(8):765–772, 2010.

[78] N. Hesse, C. Bodensteiner, M. Arens, U. G. Hofmann, R. Weinberger, and A. Sebastian Schroeder. Computer vision for medical infant motion analysis: State of the art and rgb-d data set. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[79] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes. Medical image registration. *Physics in medicine & biology*, 46(3):R1, 2001.

[80] E. Hirata and E. Sahai. Tumor microenvironment and differential responses to therapy. *Cold Spring Harbor perspectives in medicine*, 7(7):a026781, 2017.

[81] M. Holden. A review of geometric transformations for nonrigid body registration. *IEEE transactions on medical imaging*, 27(1):111–128, 2007.

[82] N. Houssami, P. Macaskill, G. von Minckwitz, M. L. Marinovich, and E. Mamounas. Meta-analysis of the association of breast cancer subtype and pathologic complete response to neoadjuvant chemotherapy. *European journal of cancer*, 48(18):3342–3354, 2012.

[83] Y. Hu, H. U. Ahmed, T. Carter, N. Arumainayagam, E. Lecornet, W. Barzell, A. Freeman, P. Nevoux, D. J. Hawkes, A. Villers, et al. A biopsy simulation study to assess the accuracy of several transrectal ultrasonography (trus)-biopsy strategies compared with template prostate mapping biopsies in patients who have undergone radical prostatectomy. *BJU international*, 110(6):812–820, 2012.

[84] W. Huang, X. Li, Y. Chen, X. Li, M.-C. Chang, M. J. Oborski, D. I. Malyarenko, M. Muzi,

G. H. Jajamovich, A. Fedorov, et al. Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: a multicenter data analysis challenge. *Translational oncology*, 2014.

[85] S. Ishii, S. Lee, H. Urakubo, H. Kume, and H. Kasai. Generative and discriminative model-based approaches to microscopic image restoration and segmentation. *Microscopy*, 69(2):79–91, 2020.

[86] L. R. Jensen, B. Garzon, M. G. Heldahl, T. F. Bathen, S. Lundgren, and I. S. Gribbestad. Diffusion-weighted and dynamic contrast-enhanced mri in evaluation of early treatment effects during neoadjuvant chemotherapy in breast cancer patients. *Journal of Magnetic Resonance Imaging*, 34(5):1099–1109, 2011.

[87] B. Jian and B. C. Vemuri. A robust algorithm for point set registration using mixture of gaussians. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1246–1251. IEEE, 2005.

[88] B. Jian and B. C. Vemuri. Robust point set registration using gaussian mixture models. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1633–1645, 2010.

[89] B. E. Johnson, A. L. Creason, J. M. Stommel, J. Keck, S. Parmar, C. B. Betts, A. Blucher, C. Boniface, E. Bucher, E. Burlingame, et al. An integrated clinical, omic, and image atlas of an evolving metastatic breast cancer. *bioRxiv*, 2020.

[90] N. Kanemaru, H. Watanabe, H. Kihara, H. Nakano, T. Nakamura, J. Nakano, G. Taga, and Y. Konishi. Jerky spontaneous movements at term age in preterm infants who later developed cerebral palsy. *Early human development*, 90(8):387–392, 2014.

[91] D. Karch, K.-S. Kang, K. Wochner, H. Philippi, M. Hadders-Algra, J. Pietz, and H. Dickhaus. Kinematic assessment of stereotypy in spontaneous movements in infants. *Gait & posture*, 36(2):307–311, 2012.

[92] M. Kaufmann, G. Von Minckwitz, E. P. Mamounas, D. Cameron, L. A. Carey, M. Cristofanilli, C. Denkert, W. Eiermann, M. Gnant, J. R. Harris, et al. Recommendations from an international consensus conference on the current status and future of neoadjuvant systemic therapy in primary breast cancer. *Annals of surgical oncology*, 19(5):1508–1516, 2012.

[93] J.-H. Kim, J.-G. Kim, Y.-H. Ji, Y.-C. Jung, and C.-Y. Won. An islanding detection method for a grid-connected system based on the goertzel algorithm. *IEEE Transactions on Power Electronics*, 26(4):1049–1055, 2011.

[94] Y. X. Kitzing, A. Prando, C. Varol, G. S. Karczmar, F. Maclean, and A. Oto. Benign conditions that mimic prostate carcinoma: Mr imaging features with histopathologic correlation. *Radiographics*, 36(1):162–175, 2016.

[95] S. Kohl, D. Bonekamp, H.-P. Schlemmer, K. Yaqubi, M. Hohenfellner, B. Hadaschik, J.-P. Radtke, and K. Maier-Hein. Adversarial networks for the detection of aggressive prostate cancer. *arXiv preprint arXiv:1702.08014*, 2017.

[96] M. Kohler et al. *Using the Kalman filter to track human interactive motion: modelling and initialization of the Kalman filter for translational motion*. Citeseer, 1997.

[97] D. Kontos, P. R. Bakic, A.-K. Carton, A. B. Troxel, E. F. Conant, and A. D. Maidment. Parenchymal texture analysis in digital breast tomosynthesis for breast cancer risk estimation: a preliminary study. *Academic radiology*, 16(3):283–298, 2009.

[98] M. Kuczma. *An introduction to the theory of functional equations and inequalities: Cauchy's equation and Jensen's inequality*. Springer Science & Business Media, 2009.

[99] R. Kvåle, B. Møller, R. Wahlqvist, S. D. Fosså, A. Berner, C. Busch, A. E. Kyrdalen, A. Svindland, T. Viset, and O. J. Halvorsen. Concordance between gleason scores of needle biopsies and radical prostatectomy specimens: a population-based study. *BJU international*, 103(12):1647–1654, 2009.

[100] S. N. Lavasani, A. Mostaar, and M. Ashtiyani. Automatic prostate cancer segmentation using kinetic analysis in dynamic contrast-enhanced mri. *Journal of Biomedical Physics & Engineering*, 8(1):107, 2018.

[101] M. Leach, B. Morgan, P. Tofts, D. Buckley, W. Huang, M. Horsfield, T. Chenevert, D. Collins, A. Jackson, D. Lomas, et al. Imaging vascular function for early stage clinical trials using dynamic contrast-enhanced magnetic resonance imaging. *European radiology*, 22(7):1451–1464, 2012.

[102] M. J. Ledesma-Carbayo, J. Kybic, M. Desco, A. Santos, and M. Unser. Cardiac motion analysis from ultrasound sequences using non-rigid registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 889–896. Springer, 2001.

[103] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. PMLR, 2015.

[104] G. Lee and H. Fujita. *Deep learning in medical image analysis: challenges and applications*, volume 1213. Springer, 2020.

[105] J. Li, W. Zhan, Y. Hu, and M. Tomizuka. Generic tracking and probabilistic prediction framework and its application in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 21(9):3634–3649, 2019.

[106] S. P. Li, A. Makris, M. J. Beresford, N. J. Taylor, M.-L. W. Ah-See, J. J. Stirling, J. A. d'Arcy, D. J. Collins, R. Kozarski, and A. R. Padhani. Use of dynamic contrast-enhanced mr imaging to predict survival in patients with primary breast cancer undergoing neoadjuvant chemotherapy. *Radiology*, 260(1):68–78, 2011.

[107] S. P. Li, A. R. Padhani, and A. Makris. Dynamic contrast-enhanced magnetic resonance imaging and blood oxygenation level-dependent magnetic resonance imaging for the assessment of changes in tumor biology with treatment. *Journal of the National Cancer Institute Monographs*, 2011(43):103–107, 2011.

[108] X. Li, L. R. Arlinghaus, G. D. Ayers, A. B. Chakravarthy, R. G. Abramson, V. G. Abramson, N. Atuegwu, J. Farley, I. A. Mayer, M. C. Kelley, et al. Dce-mri analysis methods for predicting the response of breast cancer to neoadjuvant chemotherapy: Pilot study findings. *Magnetic resonance in medicine*, 71(4):1592–1602, 2014.

[109] X. Li, H. Kang, L. R. Arlinghaus, R. G. Abramson, A. B. Chakravarthy, V. G. Abramson, J. Farley, M. Sanders, and T. E. Yankeelov. Analyzing spatial heterogeneity in dce- and dw-mri parametric maps to optimize prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. *Translational Oncology*, 7(1):14–22, 2014.

[110] G. Ligorio and A. M. Sabatini. Extended kalman filter-based methods for pose estimation using visual, inertial and magnetic sensors: Comparative analysis and performance evaluation. *Sensors*, 13(2):1919–1941, 2013.

[111] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[112] M. S. Lindström, D. Jurada, S. Bursac, I. Orsolic, J. Bartek, and S. Volarevic. Nucleolus as an emerging hub in maintenance of genome stability and cancer pathogenesis. *Oncogene*, 37(18):2351–2366, 2018.

[113] D. Linsley, J. Kim, D. Berson, and T. Serre. Robust neural circuit reconstruction from serial electron microscopy with convolutional recurrent networks. *arXiv preprint arXiv:1811.11356*, 2018.

[114] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6):e271–e297, 2019.

[115] X. Liu, D. L. Langer, M. A. Haider, Y. Yang, M. N. Wernick, and I. S. Yetik. Prostate cancer segmentation with simultaneous estimation of markov random field parameters and class. *IEEE Transactions on Medical Imaging*, 28(6):906–915, 2009.

[116] J. Ma, J. Zhao, and A. L. Yuille. Non-rigid point set registration by preserving global and local structures. *IEEE Transactions on image Processing*, 25(1):53–64, 2015.

[117] A. Machireddy, N. Meermeier, F. Coakley, and X. Song. Malignancy detection in prostate multi-parametric mr images using u-net with attention. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1520–1523. IEEE, 2020.

[118] S. O. Madgwick, A. J. Harrison, and R. Vaidyanathan. Estimation of imu and marg orientation using a gradient descent algorithm. In *2011 IEEE international conference on rehabilitation robotics*, pages 1–7. IEEE, 2011.

[119] S. Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[120] C. Marcroft, A. Khan, N. D. Embleton, M. Trenell, and T. Plötz. Movement recognition technology as a method of assessing spontaneous general movements in high risk infants. *Frontiers in neurology*, 5:284, 2015.

[121] M. Mehdy, P. Ng, E. Shair, N. Saleh, and C. Gomes. Artificial neural networks in image processing for early detection of breast cancer. *Computational and mathematical methods in medicine*, 2017, 2017.

[122] L. Meinecke, N. Breitbach-Faller, C. Bartz, R. Damen, G. Rau, and C. Disselhorst-Klug. Movement analysis in the early detection of newborns at risk for developing spasticity due to infantile cerebral palsy. *Human movement science*, 25(2):125–144, 2006.

[123] C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

[124] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[125] A. Müller, D. Schmidt, C. S. Xu, S. Pang, J. V. D'Costa, S. Kretschmar, C. Münster, T. Kurth, F. Jug, M. Weigert, et al. 3d fib-sem reconstruction of microtubule–organelle interaction in whole primary mouse β cells. *Journal of Cell Biology*, 220(2), 2021.

[126] G. Murphy, M. Haider, S. Ghai, and B. Sreeharsha. The expanding role of mri in prostate cancer. *AJR Am J Roentgenol*, 201(6):1229–1238, 2013.

[127] A. Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320. Springer, 2018.

[128] A. Myronenko and X. Song. Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence*, 32(12):2262–2275, 2010.

[129] S. Navlakha, P. Ahammad, and E. W. Myers. Unsupervised segmentation of noisy electron microscopy images using salient watersheds and region merging. *BMC bioinformatics*, 14(1):1–9, 2013.

[130] M. Nirouei, M. Pouladian, P. Abdolmaleki, S. Akhlaghpour, et al. Feature extraction and classification of breast tumors using chaos and fractal analysis on dynamic magnetic resonance imaging. *Iranian Red Crescent Med J*, 19:1–9, 2017.

[131] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[132] W. H. Organization et al. Born too soon: the global action report on preterm birth. 2012.

[133] A. R. Padhani, C. Hayes, L. Assersohn, T. Powles, A. Makris, J. Suckling, M. O. Leach, and J. E. Husband. Prediction of clinicopathologic response of breast cancer to primary chemotherapy at contrast-enhanced mr imaging: initial clinical results. *Radiology*, 239(2):361–374, 2006.

[134] G. Parlakgül, A. P. Arruda, S. Pang, E. Cagampan, N. Min, E. Güney, G. Y. Lee, K. Inouye, H. F. Hess, C. S. Xu, et al. Regulation of liver subcellular architecture controls metabolic homeostasis. *Nature*, pages 1–7, 2022.

[135] L. Peng, G. Li, M. Xiao, and L. Xie. Robust cpd algorithm for non-rigid point set registration based on structure information. *PloS one*, 11(2):e0148483, 2016.

[136] A. J. Perez, M. Seyedhosseini, T. J. Deerinck, E. A. Bushong, S. Panda, T. Tasdizen, and M. H. Ellisman. A workflow for the automatic segmentation of organelles in electron microscopy image stacks. *Frontiers in neuroanatomy*, 8:126, 2014.

[137] H. Philippi, D. Karch, K.-S. Kang, K. Wochner, J. Pietz, H. Dickhaus, and M. Hadders-Algra. Computer-based analysis of general movements reveals stereotypies predicting cerebral palsy. *Developmental Medicine & Child Neurology*, 56(10):960–967, 2014.

[138] M. D. Pickles, P. Gibbs, M. Lowry, and L. W. Turnbull. Diffusion changes precede size reduction in neoadjuvant treatment of breast cancer. *Magnetic resonance imaging*, 24(7):843–847, 2006.

[139] M. D. Pickles, M. Lowry, D. J. Manton, P. Gibbs, and L. W. Turnbull. Role of dynamic contrast enhanced mri in monitoring early response of locally advanced breast cancer to neoadjuvant chemotherapy. *Breast cancer research and treatment*, 91(1):1–10, 2005.

[140] T. M. Quan, D. G. Hildebrand, and W.-K. Jeong. Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics, 2016.

[141] A. Rasoulian, R. Rohling, and P. Abolmaesumi. Group-wise registration of point sets for statistical shape models. *IEEE transactions on medical imaging*, 31(11):2025–2034, 2012.

[142] P. Rastogi, S. J. Anderson, H. D. Bear, C. E. Geyer, M. S. Kahlenberg, A. Robidoux, R. G. Margolese, J. L. Hoehn, V. G. Vogel, S. R. Dakhil, et al. Preoperative chemotherapy: updates of national surgical adjuvant breast and bowel project protocols b-18 and b-27. *Journal of Clinical Oncology*, 26(5):778–785, 2008.

[143] S. Reardon. Rise of robot radiologists. *Nature*, 576(7787):S54–S54, 2019.

[144] J. L. Riesterer, C. S. López, E. S. Stempinski, M. Williams, K. Loftis, K. Stoltz, G. Thibault, C. Lanicault, T. Williams, and J. W. Gray. A workflow for visualizing human cancer biopsies using large-format electron microscopy. In *Methods in Cell Biology*, volume 158, pages 163–181. Elsevier, 2020.

[145] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[146] C. J. Rose, S. J. Mills, J. P. O'Connor, G. A. Buonaccorsi, C. Roberts, Y. Watson, S. Cheung, S. Zhao, B. Whitcher, A. Jackson, et al. Quantifying spatial heterogeneity in dynamic contrast-enhanced mri parameter maps. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 62(2):488–499, 2009.

[147] S. S. M. Salehi, D. Erdogmus, and A. Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017.

[148] P. Scheunders, S. Livens, G. Van de Wouwer, P. Vautrot, and D. Van Dyck. Wavelet-based texture analysis. *International Journal on Computer Science and Information Management*, 1(2):22–34, 1998.

[149] M. Seregni, C. Paganelli, P. Summers, M. Bellomi, G. Baroni, and M. Riboldi. A hybrid image registration and matching framework for real-time motion tracking in mri-guided radiotherapy. *IEEE Transactions on Biomedical Engineering*, 65(1):131–139, 2017.

[150] G. L. Serra. *Kalman filters: Theory for advanced applications*. BoD–Books on Demand, 2018.

[151] W. Shen, B. Wang, Y. Jiang, Y. Wang, and A. Yuille. Multi-stage multi-recursive-input fully convolutional networks for neuronal boundary detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2391–2400, 2017.

[152] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.

[153] M. Siam, M. Gamal, M. Abdel-Razek, S. Yogamani, M. Jagersand, and H. Zhang. A comparative study of real-time semantic segmentation for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[154] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019.

[155] R. A. Smith, K. S. Andrews, D. Brooks, S. A. Fedewa, D. Manassaram-Baptiste, D. Saslow, O. W. Brawley, and R. C. Wender. Cancer screening in the united states, 2017: a review of current american cancer society guidelines and current issues in cancer screening. *CA: a cancer journal for clinicians*, 67(2):100–121, 2017.

[156] F. Soares, F. Janela, M. Pereira, J. Seabra, and M. M. Freire. 3d lacunarity in multifractal analysis of breast tumor lesions in dynamic contrast-enhanced magnetic resonance imaging. *IEEE Transactions on image processing*, 22(11):4422–4435, 2013.

[157] A. Stahl, C. Schellewald, Ø. Stavdahl, O. M. Aamo, L. Adde, and H. Kirkerod. An optical flow-based method to predict infantile cerebral palsy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(4):605–614, 2012.

[158] G. P. Stein, O. Mano, and A. Shashua. Vision-based acc with a single camera: bounds on range and range rate accuracy. In *IEEE IV2003 intelligent vehicles symposium. Proceedings (Cat. No. 03TH8683)*, pages 120–125. IEEE, 2003.

[159] P. Suetens. *Fundamentals of medical imaging*. Cambridge university press, 2017.

[160] W. F. Symmans, F. Peintinger, C. Hatzis, R. Rajan, H. Kuerer, V. Valero, L. Assad, A. Poniecka, B. Hennessy, M. Green, et al. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *Journal of Clinical Oncology*, 25(28):4414–4422, 2007.

[161] H. Y. Tanaka and M. R. Kano. Stromal barriers to nanomedicine penetration in the pancreatic tumor microenvironment. *Cancer science*, 109(7):2085–2092, 2018.

[162] U. Tateishi, M. Miyake, T. Nagaoka, T. Terauchi, K. Kubota, T. Kinoshita, H. Daisaki, and H. A. Macapinlac. Neoadjuvant chemotherapy in breast cancer: prediction of pathologic response with pet/ct and dynamic contrast-enhanced mr imaging—prospective assessment. *Radiology*, 263(1):53–63, 2012.

[163] S. Teipel, A. Drzezga, M. J. Grothe, H. Barthel, G. Chételat, N. Schuff, P. Skudlarski, E. Cavedo, G. B. Frisoni, W. Hoffmann, et al. Multimodal imaging in alzheimer's disease: validity and usefulness for early detection. *The Lancet Neurology*, 14(10):1037–1053, 2015.

[164] J. R. Teruel, M. G. Heldahl, P. E. Goa, M. Pickles, S. Lundgren, T. F. Bathen, and P. Gibbs. Dynamic contrast-enhanced mri texture analysis for pretreatment prediction of clinical and pathological response to neoadjuvant chemotherapy in patients with locally advanced breast cancer. *NMR in Biomedicine*, 27(8):887–896, 2014.

[165] S. Thakkar, D. Sharma, K. Kalia, and R. K. Tekade. Tumor microenvironment targeted nanotherapeutics for cancer therapy and diagnosis: A review. *Acta biomaterialia*, 101:43–68, 2020.

[166] P. Thevenaz, U. E. Ruttimann, and M. Unser. A pyramid approach to subpixel registration based on intensity. *IEEE transactions on image processing*, 7(1):27–41, 1998.

[167] G. Thibault, J. Angulo, and F. Meyer. Advanced statistical matrices for texture characterization: application to cell classification. *IEEE Transactions on Biomedical Engineering*, 61(3):630–637, 2013.

[168] G. Thibault, B. Fertil, C. Navarro, S. Pereira, P. Cau, N. Levy, J. Sequeira, and J.-L. Mari. Shape and texture indexes application to cell nuclei classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 27(01):1357002, 2013.

[169] G. Thibault and I. Shafran. Fuzzy statistical matrices for cell classification. *arXiv preprint arXiv:1611.06009*, 2016.

[170] G. Thibault, A. Tudorica, A. Afzal, S. Y. Chui, A. Naik, M. L. Troxell, K. A. Kemmer, K. Y. Oh, N. Roy, N. Jafarian, et al. Dce-mri texture features for early prediction of breast cancer therapy response. *Tomography*, 3(1):23–32, 2017.

[171] P. S. Tofts, G. Brix, D. L. Buckley, J. L. Evelhoch, E. Henderson, M. V. Knopp, H. B. Larsson, T.-Y. Lee, N. A. Mayr, G. J. Parker, et al. Estimating kinetic parameters from dynamic contrast-enhanced t1-weighted mri of a diffusable tracer: standardized quantities and symbols. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 10(3):223–232, 1999.

[172] P. S. Tofts and A. G. Kermode. Measurement of the blood-brain barrier permeability and leakage space using dynamic mr imaging. 1. fundamental concepts. *Magnetic resonance in medicine*, 17(2):357–367, 1991.

[173] A. Tudorica, K. Y. Oh, S. Y. Chui, N. Roy, M. L. Troxell, A. Naik, K. A. Kemmer, Y. Chen, M. L. Holtorf, A. Afzal, et al. Early prediction and evaluation of breast cancer response

to neoadjuvant chemotherapy using quantitative dce-mri. *Translational oncology*, 9(1):8–17, 2016.

[174] A. Tzalavra, K. Dalakleidi, E. I. Zacharaki, N. Tsiaparas, F. Constantinidis, N. Paragios, and K. S. Nikita. Comparison of multi-resolution analysis patterns for texture classification of breast tumors based on dce-mri. In *International Workshop on Machine Learning in Medical Imaging*, pages 296–304. Springer, 2016.

[175] A. G. Tzalavra, E. I. Zacharaki, N. N. Tsiaparas, F. Constantinidis, and K. S. Nikita. A multiresolution analysis framework for breast tumor classification based on dce-mri. In *2014 IEEE International Conference on Imaging Systems and Techniques (IST) Proceedings*, pages 246–250. IEEE, 2014.

[176] M. Tzelepi and A. Tefas. Semantic scene segmentation for robotics applications. In *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–4. IEEE, 2021.

[177] A. Umar and S. Atabo. A review of imaging techniques in scientific research/clinical diagnosis. *MOJ Anat. Physiol.*, 6:175–183, 2019.

[178] C. Urrea and R. Agramonte. Kalman filter: historical overview and review of its use in robotics 60 years after its creation. *Journal of Sensors*, 2021, 2021.

[179] B. Van Ginneken, M. B. Stegmann, and M. Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical image analysis*, 10(1):19–40, 2006.

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

[181] H. M. Vergara, C. Pape, K. I. Meechan, V. Zinchenko, C. Genoud, A. A. Wanner, K. N. Mutemi, B. Titze, R. M. Templin, P. Y. Bertucci, et al. Whole-body integration of gene expression and single-cell morphology. *Cell*, 184(18):4819–4837, 2021.

[182] M. T. Vlaardingerbroek and J. A. Boer. *Magnetic resonance imaging: theory and practice*. Springer Science & Business Media, 2013.

[183] G. Von Minckwitz, M. Untch, J.-U. Blohmer, S. D. Costa, H. Eidtmann, P. A. Fasching, B. Gerber, W. Eiermann, J. Hilfrich, J. Huober, et al. Definition and impact of pathologic

complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. *J Clin oncol*, 30(15):1796–1804, 2012.

[184] H. Wadell. Volume, shape, and roundness of quartz particles. *The Journal of Geology*, 43(3):250–280, 1935.

[185] D. F. Walnut. *An introduction to wavelet analysis*. Springer Science & Business Media, 2002.

[186] Y. Wang, B. Georgescu, T. Chen, W. Wu, P. Wang, X. Lu, R. Ionasec, Y. Zheng, and D. Comaniciu. Learning-based detection and tracking in medical imaging: a probabilistic approach. In *Deformation Models*, pages 209–235. Springer, 2013.

[187] R. A. Willoughby. Solutions of ill-posed problems (an tikhonov and vy arsenin). *SIAM Review*, 21(2):266, 1979.

[188] X. Yang, C. Liu, Z. Wang, J. Yang, H. Le Min, L. Wang, and K.-T. T. Cheng. Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric mri. *Medical image analysis*, 42:212–227, 2017.

[189] Z. Yang, Y. Wei, and Y. Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34, 2021.

[190] T. E. Yankeelov, M. Lepage, A. Chakravarthy, E. E. Broome, K. J. Niermann, M. C. Kelley, I. Meszoely, I. A. Mayer, C. R. Herman, K. McManus, et al. Integration of quantitative dce-mri and adc mapping to monitor treatment response in human breast cancer: initial results. *Magnetic resonance imaging*, 25(1):1–13, 2007.

[191] T. E. Yankeelov, D. A. Mankoff, L. H. Schwartz, F. S. Lieberman, J. M. Buatti, J. M. Mountz, B. J. Erickson, F. M. Fennessy, W. Huang, J. Kalpathy-Cramer, et al. Quantitative imaging in cancer clinical trials. *Clinical Cancer Research*, 22(2):284–290, 2016.

[192] T. E. Yankeelov, W. D. Rooney, X. Li, and C. S. Springer Jr. Variation of the relaxographic "shutter-speed" for transcytolemmal water exchange affects the cr bolus-tracking curve shape. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 50(6):1151–1169, 2003.

[193] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[194] J. Yu, A. Fulcher, M. Turner, C. Cockrell, E. Cote, and T. Wallace. Prostate cancer and its mimics at multiparametric prostate mri. *The British journal of radiology*, 87(1037):20130659, 2014.

[195] A. L. Yuille and N. M. Grzywacz. A mathematical analysis of the motion coherence theory. *International Journal of Computer Vision*, 3(2):155–175, 1989.

[196] M. Zambetti, M. Mansutti, P. Gomez, A. Lluch, C. Dittrich, C. Zamagni, E. Ciruelos, L. Pavesi, V. Semiglazov, E. De Benedictis, et al. Pathological complete response rates following different neoadjuvant chemotherapy regimens for operable breast cancer according to er status, in two parallel, randomized phase ii trials with an adaptive study design (ecto ii). *Breast cancer research and treatment*, 132(3):843–851, 2012.

[197] T. Zeng, B. Wu, and S. Ji. Deepem3d: approaching human-level performance on 3d anisotropic em image segmentation. *Bioinformatics*, 33(16):2555–2562, 2017.

[198] G. Zhang and H. Chanson. Application of local optical flow methods to high-velocity free-surface flows: Validation and application to stepped chutes. *Experimental Thermal and Fluid Science*, 90:186–199, 2018.

[199] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

[200] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[201] Y. Zheng and D. Doermann. Robust point matching for nonrigid shapes by preserving local neighborhood structures. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):643–649, 2006.

[202] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[203] D. Zink, A. H. Fischer, and J. A. Nickerson. Nuclear structure in cancer cells. *Nature reviews cancer*, 4(9):677–687, 2004.

[204] J. G. Zwicker. Motor impairment in very preterm infants: implications for clinical practice and research. 2014.

Appendix A

Reproducing Kernel Hilbert Spaces and Representer Theorem

The aim of regularization is to find a solution to an ill-posed problem by making the result depend smoothly on the data by restricting the hypothesis space H. The hypothesis space is a set of all possible hypotheses that approximate an unknown function and the goal of a learning process is to find the final hypothesis in this space that best approximates the unknown function. One way of imposing regularization is to introduce a regularization term in the error minimization equation, which will force the minimization to seek for simpler functions, which incur less penalty. The most commonly used regularization method to solve ill-posed problems was introduced by Tikhonov and can be written as,

$$E(f) = Err(f) + \frac{\lambda}{2}\|f\|_H^2 \tag{A.1}$$

where $Err(f)$ is the empirical error term, λ is the regularization parameter and $\|.\|$ is the norm of the function in the hypothesis space H. This is similar to Equation 3.16 in our formulation, where velocity field ν is the function being evaluated. In order to solve such problems, some hypothesis spaces with specific desirable properties have been defined. One such space is the Reproducing Kernel Hilbert Space. We are discussing only the basic theory here and detailed explanation can be found in [37, 123, 63, 98, 187]. Hilbert space is a space, which extends the concept of Euclidean space to infinite dimensions. It is a complete metric space with respect to a distance function and has an inner product.

Definition A.1. *An evaluation functional over the Hilbert space H is a linear functional F that evaluates each function in the space at the point x,*

$$F_t[f] = f(x) \quad \forall f \in H. \tag{A.2}$$

Definition A.2. *A Hilbert space H is a Reproducing Kernel Hilbert Space, if the evaluation*

functionals are bounded, i.e. if for all x there exists some $M > 0$, such that

$$|F_t[f]| = |f(x)| \leq M\|f\|_H. \tag{A.3}$$

Definition A.3. *A function K is a reproducing kernel if it is symmetric i.e. $K(x, y) = K(y, x)$ and positive definite*

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \geq 0 \tag{A.4}$$

for any $n \in N$ and choice of $x_1, x_2, \ldots, x_n \in X$ and $c_1, c_2, \ldots, c_n \in R$. A reproducing kernel K reproduces the value of a function $f \in H$ at a point $x \in X$. This gives the reproducing property which states that for every $x \in X$ there exists a function $K_x \in H$, called the representer of x, such that $F_t[f] = \langle K_x, f\rangle = f(x)$.

Theorem A.1. *A RKHS defines a corresponding reproducing kernel. Conversely a reproducing kernel defines a unique RKHS.*
Given a reproducing kernel K, a unique RKHS can be defined as $H = \sum_{n=1}^{N} c_i K(x, x_i)$, whose norm is defined by the inner product,

$$\langle f, g\rangle = \sum_{i=1}^{s}\sum_{j=1}^{s'} c_i c_j K(x, x'_j) \tag{A.5}$$

where $f(x) = \sum_{i=1}^{s} c_i K_{x_i}(x)$ and $g(x) = \sum_{i=1}^{s'} c_i K_{x'_i}(x)$.

Theorem A.2. *(The Representer Theorem) The optimal solution $\hat{f}$ that minimizes a regularized error functional of the form*

$$E(f) = Err(f) + \frac{\lambda}{2}\|f\|_H^2 \tag{A.6}$$

can be represented as

$$\hat{f} = \sum_{i=1}^{N} c_i K(x, x_i) \tag{A.7}$$

Proof. Given a reproducing kernel K, we can define a unique RKHS H_0 in the subspace of H as,

$$H_0 = \left\{ f \in H | f = \sum_{i=1}^{N} c_i K(x, x_i) \right\}. \tag{A.8}$$

Let $H_0^{\perp}$ be a linear subspace of H orthogonal to H_0, i.e.

$$H_0^{\perp} = \{g \in H |\, \langle g, f\rangle = 0\} \quad \forall f \in H_0. \tag{A.9}$$

As H_0 is finite dimensional and closed we can write, $H = H_0 \oplus H_0^{\perp}$. Now, every function $f \in H$ can be uniquely decomposed into a component along H_0, denoted by f_0, and a component perpendicular to $H_0^{\perp}$, denoted by $f_0^{\perp}$ i.e., $f = f_0 + f_0^{\perp}$. By orthogonality, $\|f_0 + f_0^{\perp}\|^2 = \|f_0\|^2 + \|f_0^{\perp}\|^2$ and by the reproducing property, $Err(f_0 + f_0^{\perp}) = Err(f_0)$ (since evaluating $f(x) = f_0(x) + f_0^{\perp}(x)$ to

compute the empirical error requires taking the inner product with the representer K_{x_i}, and doing so nullifies the $f_0^{\perp}$ term while preserving the f_0 term). Combining these two facts, we can see that,

$$E(f) = Err(f) + \frac{\lambda}{2}\|f\|_H^2 = Err\left(f_0 + f_0^{\perp}\right) + \frac{\lambda}{2}\|f_0 + f_0^{\perp}\|_H^2 \tag{A.10}$$

$$= Err(f_0) + \frac{\lambda}{2}\|f_0\|_H^2 + \frac{\lambda}{2}\|f_0^{\perp}\|_H^2 \tag{A.11}$$

$$\geq Err(f_0) + \frac{\lambda}{2}\|f_0\|_H^2. \tag{A.12}$$

Hence, the optimal minimum of the regularized error functional comes from the space H_0 and has the form $\hat{f} = \sum_{i=1}^{N} c_i K(x, x_i)$. In our solution for the velocity field ν, we have used a Gaussian as the reproducing kernel as it helps extract the high frequency content of the function, which is minimized to obtain a smooth result. □