BIOLOGICAL CELL COUNTING USING DEEP LEARNING

M.Tech. Thesis

By

SAURABH KUMAR

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

MAY 2025

BIOLOGICAL CELL COUNTING USING DEEP LEARNING

A THESIS

Submitted in partial fulfillment of the
requirements for the award of the degree
of
Master of Technology

by
SAURABH KUMAR

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE
MAY 2025

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **BIOLOGICAL CELL COUNTING USING DEEP LEARNING** in the partial fulfillment of the requirements for the award of the degree of **MASTER OF TECHNOLOGY** and submitted in the **DEPARTMENT OF ELECTRICAL ENGINEERING, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from July 2023 to May 2025 under the supervision of Prof. Vivek Kanhangad.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Saurabh Kumar

29-05-2025

(SAURABH KUMAR)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

29 May 2025

Signature of the Supervisor of M.Tech. thesis (with date)

(PROF. VIVEK KANHANGAD)

SAURABH KUMAR has successfully given his M.Tech. Oral Examination held on **7th May, 2025**.

29 May 2025

Signature of Supervisor of M.Tech. thesis

Date:

Saptarshi Ghosh

Convener, DPGC

Date: 31-05-2025

ACKNOWLEDGEMENTS

First and foremost, I express my heartfelt gratitude to my supervisor, **Prof. Vivek Kanhangad**, for his exceptional mentorship, continuous encouragement, and insightful guidance throughout my M.Tech research at the Indian Institute of Technology Indore. His invaluable suggestions, high academic standards, and unwavering support have played a pivotal role in the successful completion of this thesis.

I am sincerely thankful to Mr. Pawan Soni for his consistent support and valuable technical assistance at various stages of this work. His willingness to help and constructive inputs were instrumental in overcoming several challenges.

I also extend my appreciation to my lab mate, Mr. Garv Jain, for his camaraderie, collaborative spirit, and insightful discussions.

I gratefully acknowledge the Ministry of Education, Government of India, for the financial support that enabled my academic and research activities during the M.Tech program.

Lastly, my deepest thanks go to my family for their unconditional love, motivation, and enduring patience. Their unwavering belief in me and moral support have been the foundation of this journey.

I am truly grateful to all who supported me—directly and indirectly—in the successful completion of this work.

Saurabh Kumar

dedicated to my family

Abstract

Accurate biological cell counting plays a pivotal role in numerous biomedical applications, yet conventional manual and rule-based approaches struggle with dense, overlapping, and morphologically diverse cells. This thesis presents a hybrid deep learning framework for automated cell counting using both convolutional and transformer-based architectures.

Initially, a Dual Cascaded Network (DCNet) is proposed, combining a VGG16-based encoder with a U-Net decoder to generate high-resolution density maps from microscopy images. To address limitations in crowded cells, a transformer-based alternative—Restormer—is employed, offering improved global context modeling through attention mechanisms and specialized components such as Multi-Dconv Head Transposed Attention and Gated Feed-Forward Networks.

The study introduces Focal Inverse Distance Transform (FIDT) maps to enhance localization precision in dense cell environments. Additionally, a SALW strategy is integrated to dynamically balance learning difficulty across spatial regions. Evaluated on diverse datasets—including synthetic bacterial, bone marrow, and adipose tissue images—the proposed models demonstrate robust performance, achieving competitive accuracy across varying imaging conditions. This work highlights the effectiveness of hybrid architectures and attention-guided learning in advancing the state-of-the-art in cell counting.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

Acronym	Expansion
DNNs	Deep Neural Networks
FCNs	Fully Connected Networks
RGB	Red Green Blue
Conv	Convolution
ReLU	Rectified Linear Unit
ViT	Vision Transformer
MHSA	Multi-Head Self-Attention
FFN	Feed Forward Network
GELU	Gaussian Error Linear Unit
ASPP	Atrous Spatial Pyramid Pooling
GPU	Graphical Processing Unit
YOLO	You Only Look Once
R-CNN	Region based Convolutional Neural Network
MAE	Mean Absolute Error
MSE	Mean Squared Error
GAME	Grid Average Mean Average Error
VGG	Synthetic Fluorescence Microscopy
MBM	Modified Bone Marrow
ADI	Adipocyte Tissue

H&E	Hematoxylin and Eosin
GTEx	Genotype-Tissue Expression
SALW	Self Adaptive Loss Weighting
SVMs	Support Vector Machines
C-FCRN	Concatenated Fully Connected Regression Network
FCRN	Fully Connected Regression Network
FIDT	Focal Inverse Distance Transform
VGG16	Visual Geometry Group (16 Layers)
DCNet	Dual Cascaded Network
Restormer	Restoration Transformer
MDTA	Multi-Dconv Head Transposed Attention
GDFN	Gated-Dconv Feed Forward Network
D-conv	Dilated Convolution
LoG	Laplacian of Gaussian
LoIG	Laplacian of Inverse Gaussian

SYMBOLS

Symbol	Meaning
Q	Query
K	Key
V	Value
d	Distance
N	Number
C_i^{pred}	predicted count for i^{th} image
C_i^{gt}	ground truth count for i^{th} image
$\mathbb{R}$	Real Number
γ	Gamma
$*$	Convolution
σ	Sigma
Ω	Omega
π	Pi
x	x coordinate
y	y coordinate
p	Spatial Space
L	Loss
exp	Exponential
a	Learnable Parameter

∇	Laplacian Operator
I	Image

Chapter 1

Introduction

1.1 Why cell counting?

Accurate cell counting in microscopy images is vital in various biomedical and clinical applications, including disease diagnosis, drug discovery, and understanding cellular mechanisms. Traditionally performed manually, this process is not only time-consuming but also prone to subjective errors and inconsistencies, particularly when dealing with densely packed or overlapping cells. These limitations have catalyzed the adoption of deep learning methods, which offer automation, scalability, and improved accuracy.

Deep learning, especially convolutional and transformer-based neural networks, has revolutionized cell counting by enabling the estimation of cell densities through density maps. These models are capable of capturing complex spatial patterns and variations in cell morphology, even under challenging imaging conditions. Moreover, they significantly reduce the labor-intensive nature of manual annotation while ensuring consistent performance across large datasets. Given the increasing volume of biomedical image data and the demand for high-throughput analysis, integrating deep learning for cell counting is not just advantageous—it is essential for modern biological research and healthcare advancements.

1.1.1 Fundamentals of a Deep Neural Network

A DNN consists of multiple interconnected layers, each designed to perform specific roles in extracting and processing information from the input data. The first component of a DNN is the input layer, which receives raw data such as grayscale or RGB microscopy images. Each neuron in this layer corresponds to a single pixel or a group of pixels from the input. For example, in the context of cell counting, the input layer may process a 256×256 fluorescence microscopy image in which cells appear as bright regions against a dark background.

The central part of a DNN is composed of its hidden layers, where most of the computational learning takes place. Convolutional layers, often referred to as Conv layers, are among the most critical components within this section. They specialize in identifying spatial patterns like edges, textures, and structures by applying small filters (kernels) that slide across the image to generate feature maps. These feature maps allow the network to detect vital cellular structures, such as nuclei and boundaries, which are essential in biomedical image analysis. Activation functions are then applied to the outputs of these convolutional layers to introduce non-linearity into the network, enabling it to capture complex relationships. Commonly used activation functions include ReLU (Rectified Linear Unit), which enhances training efficiency by suppressing negative values, along with alternatives like Tanh and Sigmoid.

To further optimize performance and computational efficiency, pooling layers are incorporated to reduce the dimensionality of the feature maps. Techniques such as max pooling and average pooling are widely used to retain the most salient features while decreasing the data volume, thus aiding in the recognition of cells even when their positions vary slightly. Normalization layers, such as batch normalization, are introduced to maintain consistent activation distributions throughout the network, resulting in faster and more stable training. Furthermore, dropout layers play a critical role in regularizing the model by randomly deactivating neurons during training, which helps prevent overfitting—particularly important when dealing with small-scale biomedical datasets.

As the data progresses through the network, it reaches the fully connected or dense layers, which are responsible for synthesizing and interpreting the extracted features to make a final prediction. Each neuron in a dense layer is connected to all neurons in the previous layer, allowing for comprehensive integration of information. The final component, the output layer, is specifically configured based on the task. For classification tasks, it assigns class labels, while for detection or localization, it outputs spatial coordinates or segmentation masks. In the

case of cell counting, the output may be a scalar representing the total count, a density map indicating concentration, or a count map showing spatial distribution. Depending on the required output, the activation function at this stage might be linear for regression purposes or softmax for classification tasks.

1.1.2 Why DNNs for Biological Cell Counting?

Traditional cell counting techniques have predominantly relied on manually crafted image processing methods, including thresholding, edge detection, and morphological operations. While these approaches were foundational, they often suffer from sensitivity to noise, inconsistent staining procedures, and variations in cell morphology, limiting their generalizability across different datasets and imaging conditions. In contrast, DNNs bring significant advantages to cell counting tasks, beginning with automated feature learning. Unlike traditional methods that require manual filter design, DNNs are capable of learning discriminative features directly from raw data, thereby reducing the reliance on human intervention.

Furthermore, DNNs demonstrate robust performance across a range of imaging modalities, including fluorescence and brightfield microscopy, making them well-suited for diverse biomedical applications. Their adaptability is another notable strength, as these models can be fine-tuned with relatively minimal changes to work effectively on different datasets. This is especially useful in biomedical research, where imaging conditions and specimen types may vary significantly. Additionally, DNN architectures such as U-Net [11], SAU-Net [3], and CSRNet [10] have shown considerable scalability, with the capability to be extended to accommodate high-resolution images, three-dimensional (3D) volumes, or even time-lapse sequences. Finally, by leveraging both spatial and contextual cues from the input images, DNNs consistently outperform traditional cell counting methods in terms of both precision and recall, making them a more accurate and reliable choice for modern biomedical image analysis.

1.1.3 Application in Cell Counting Pipelines

In deep learning-based biological cell counting, the overall workflow typically follows a structured pipeline designed to handle the complexities of microscopy image analysis. The process begins with input image preprocessing, which may include operations such as normalization and resizing to standardize the data and prepare it for efficient processing. This is followed by feature extraction, where convolutional layers are employed to identify and capture critical patterns within the image, such as cell edges, textures, and spatial arrangements. The network then generates intermediate representations, which may take the form of segmentation masks or density maps, providing a detailed visualization of cell locations and distributions. These representations are interpreted to estimate either the total number of cells or their specific positions within the image. Finally, a postprocessing step is applied to refine the predictions, which is particularly important in densely populated regions where cells may overlap or be closely clustered.

DNNs have become foundational to modern biomedical image analysis due to their ability to learn and generalize from complex datasets. Their layered, hierarchical architecture, which resembles the blob-like object detection mechanism, enables them to tackle the high precision demands of cell counting tasks even under challenging imaging conditions. As microscopy technologies continue to evolve, producing increasingly large and intricate datasets, the role of DNNs in automating and enhancing quantitative analysis in cell biology is expected to grow even more significant.

1.2 Transformer

Transformers have emerged as a transformative deep learning architecture initially developed for natural language processing but now gaining significant traction in computer vision, including biomedical image analysis and cell counting tasks. Unlike convolutional networks

that rely on local receptive fields, Transformers excel in modelling global dependencies using attention mechanisms. This capability is particularly valuable when spatial context plays a crucial role, such as in high-resolution biological images where cells exhibit varying density, shape, and arrangement.

1.2.1 Basic Architecture of a Transformer

The standard Transformer architecture, as introduced in the seminal paper *"Attention is All You Need"* by Vaswani et al. (2017) [14], follows an encoder-decoder structure originally designed for natural language processing. However, for vision-centric tasks such as cell counting or segmentation, the encoder-only variant—popularized through Vision Transformers (ViTs)—has been more commonly adopted due to its suitability for spatial data processing.

In this adaptation to vision applications, the first component is the input embedding layer. Here, a two-dimensional image is divided into fixed-size patches (typically 16×16 pixels). Each patch is then flattened into a vector and passed through a linear projection layer to form token embeddings. To retain spatial information, which is essential in visual data, positional encodings are added to these tokens.

Following the embedding layer is the Multi-Head Self-Attention (MHSA) mechanism. This module enables the model to attend to multiple spatial regions simultaneously, allowing for a more comprehensive understanding of the image context. The MHSA mechanism involves computing query Q, key K, and value V matrices from the input embeddings. The attention scores are then computed using the scaled dot-product formula:

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right) \qquad (1)$$

This multi-head formulation empowers the model to capture diverse semantic relationships across different image regions, which is especially useful in biomedical imaging where relevant features such as cell centres or boundaries may appear in varying positions and forms.

The output of the attention layer is passed through a Feed-Forward Network (FFN), which typically consists of two fully connected layers with a non-linear activation function, most commonly the Gaussian Error Linear Unit (GELU). This component independently transforms each token, enhancing the representational capacity of the model.

To ensure training stability and efficient learning, each sub-layer in the Transformer encoder includes layer normalization and residual connections. These mechanisms play a critical role in enabling consistent gradient flow across layers, stabilizing training dynamics, and facilitating the learning of identity mappings, which can accelerate convergence.

Positional encoding is another essential element of the Transformer, especially in vision applications. Since self-attention mechanisms do not inherently consider the order or position of tokens, positional encodings are integrated to inject spatial order information into the model. These encodings can be sinusoidal, learned, or based on relative positioning, depending on the implementation.

Finally, the design of the output head varies based on the specific vision task. For classification tasks, a special class token is passed through a linear layer to predict category labels. In segmentation tasks, the output is reshaped into a feature map that aligns with the input image. In the case of cell counting, the Transformer's final outputs are decoded into density maps or scalar counts, thereby translating learned spatial and contextual representations into quantitative biological information.

1.2.2 Advantages of Transformers in Cell Counting

Transformers offer several advantages that make them particularly well-suited for biological cell counting and related vision tasks. One of their most significant strengths lies in their ability to capture global contextual information. The self-attention mechanism allows the model to understand relationships between spatially distant regions within an image, which is particularly beneficial in scenarios involving overlapping or densely clustered cells. This global perspective enables

the model to differentiate between individual cells even when boundaries are ambiguous.

Another advantage is the scalability of Transformers to large image inputs. Unlike convolutional neural networks (CNNs), where the receptive field is constrained by the kernel size and network depth, Transformers can process an entire image context simultaneously without requiring deeper architectures. This property allows them to analyse large-scale microscopy images more effectively. Moreover, Transformers tend to generalize better across different datasets. Since they are less reliant on local texture features compared to CNNs, they are more adaptable to variations in imaging conditions, making them ideal for applications in biomedical domains where dataset heterogeneity is common.

The flexibility of Transformer architectures further enhances their utility. They can be seamlessly integrated with other neural components such as convolutional layers, Atrous Spatial Pyramid Pooling (ASPP), and attention gates. This modularity has led to the development of powerful hybrid models like SAU-Net [3] and Restormer [12], which combine the strengths of different architectural paradigms for improved performance. Lastly, Transformers demonstrate superior effectiveness in challenging scenarios such as crowded or low-contrast microscopy images. In these situations, where cell boundaries may be faint or indistinct, the attention mechanism enables the model to focus on subtle but biologically significant features, thereby improving detection and counting accuracy.

1.2.3 Limitations of Transformers

Despite their powerful capabilities, Transformers also present several limitations when applied to image-based biomedical tasks. One major drawback is their high computational cost. The self-attention mechanism inherent to Transformers scales quadratically with the input size, meaning that as image dimensions increase, so does the demand for computational resources. This makes training on high-resolution

biomedical images particularly demanding in terms of GPU memory and processing power.

Another significant limitation is their data-hungry nature. Transformers typically require large-scale datasets to achieve optimal performance, which poses a challenge in the biomedical field where obtaining extensive, well-annotated datasets is often infeasible due to time, cost, and domain expertise constraints. This reliance on large datasets can hinder their applicability in medical scenarios with limited labelled data.

Additionally, Transformers lack the inductive biases that are inherently present in CNNs. CNNs are designed to be translation-invariant and spatially aware, enabling them to efficiently process image data with fewer training samples. In contrast, Transformers must learn these spatial relationships and patterns from scratch, which not only increases the complexity of training but also requires more data and time to achieve comparable performance.

Furthermore, the effectiveness of Transformers heavily depends on the method used for positional encoding, which is essential for embedding spatial information into the model. In biomedical applications where the precise location and morphology of cells are crucial, any inadequacy in encoding spatial relationships can adversely impact performance. Therefore, the reliance on positional encoding adds another layer of sensitivity and potential instability to Transformer-based models in medical imaging tasks.

1.2.4 Transformers in Biomedical Imaging

In recent research, various Transformer-based architectures have been effectively applied to microscopy-based cell counting, demonstrating notable improvements over traditional methods. For instance, SAU-Net [3] incorporates self-attention modules into the widely used U-Net [11] architecture, enhancing the model's ability to focus on the foreground regions, particularly the cells. This integration allows for more precise localization and counting in dense cell populations. Similarly, Restormer [12] utilizes efficient Transformer blocks designed for high-

resolution image restoration. These blocks have been adapted for improving the quality of low-resolution or noisy microscopy images, thereby aiding in more accurate downstream cell analysis tasks.

These advanced architectures leverage the Transformer's strength in modelling both local and global features, which is particularly advantageous for complex tasks like 2D and 3D cell counting. Transformers represent a significant paradigm shift from conventional convolution-based image analysis approaches. Their capacity to capture long-range dependencies across an image makes them exceptionally suitable for analysing dense, cluttered, or high-resolution cellular imagery. Although they come with certain challenges, such as high computational demands and a need for large training datasets, ongoing research into lightweight Transformers, hybrid network architectures, and attention-enhanced CNNs is steadily addressing these limitations. Consequently, Transformers are expected to become a cornerstone technology in the future landscape of automated biological cell counting and image-based biomedical analysis.

1.3 Cell Counting

Cell counting is a fundamental task in many biological and biomedical research applications, including cancer diagnosis, stem cell therapy, drug screening, tissue analysis, and neuroscience. Accurately quantifying the number of cells in microscopy images provides critical information for assessing cell proliferation, viability, density, and overall health. Despite its importance, traditional manual counting is time-consuming, subjective, and prone to human error, particularly in large-scale or high-throughput experiments.

With the advancement of computational techniques and the advent of deep learning, automated cell counting has evolved into a sophisticated and reliable alternative. This section explores the conceptual

foundations, traditional challenges, and modern deep learning-based approaches for cell counting.

Cell counting serves as a cornerstone in several experimental and diagnostic workflows. In tissue engineering, it determines cell proliferation rates; in cancer studies, it measures tumour growth or regression; in stem cell research, it evaluates differentiation and regeneration; in drug discovery, it assesses cytotoxicity of drug compounds; and in immunology, it quantifies immune response through changes in cell populations. Accurate and reproducible counting methods are critical for ensuring experimental validity, reproducibility, and scaling up clinical research.

Historically, cell counting has been performed through manual counting and classical image processing. Manual counting, often carried out by experts using a hemacytometer or by annotating microscopy images, is time-consuming and labour-intensive. It is also subject to inter-observer and intra-observer variability and is infeasible for large-scale image datasets. Classical image processing employs techniques such as thresholding, edge detection (e.g., Sobel, Canny), morphological operations (dilation, erosion), and watershed segmentation. These methods work well for images with good contrast and minimal noise but are sensitive to lighting and staining variations and struggle with overlapping cells. Moreover, they require task-specific hand-engineered features and generalize poorly across different datasets.

The shift toward deep learning has addressed many shortcomings of traditional approaches. Instead of relying on handcrafted rules, deep learning models learn feature representations directly from annotated data. Depending on the nature of the output, modern cell counting methods can be categorized into detection-based methods, regression-based methods, and density map estimation.

Detection-based methods treat cell counting as a detection problem by identifying cell centres or nuclei using bounding boxes or circular masks. These methods employ object detection networks like Faster R-

CNN [13], YOLO [18], or variants of U-Net [11], [3]. However, they face challenges in densely populated or overlapping cell regions and require precise localization for each cell. Regression-based methods predict the total number of cells in an image without identifying individual locations, requiring simpler image-level labels but lacking spatial information about cell distribution. The most common and accurate method in recent years has been density map estimation. In this approach, each annotated cell, typically represented as a dot, is converted into a Gaussian blob. The network learns to regress a density map such that its integral equals the total cell count. This method effectively handles overlapping and crowded cells and does not require precise segmentation or bounding boxes.

Several architectures have been proposed specifically or adapted for cell counting. CSRNet [10], a dilated convolutional network, preserves spatial resolution while enlarging the receptive field, making it effective for highly congested scenes. Count-ception [4] introduces redundant counting through a fully convolutional network, using overlapping receptive fields to count the same cells multiple times and averaging predictions to reduce errors. SAU-Net [3] is an attention-augmented U-Net [11] that incorporates self-attention modules and supports both 2D and 3D data, enabling volumetric cell counting. The Two-Path Network [20] employs a dual-stream architecture, with one path capturing spatial details and the other focusing on semantic context, combining their outputs to produce accurate density maps.

To assess the performance of cell counting algorithms, several standard metrics are used. Mean Absolute Error (MAE) is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|C_i^{pred} - C_i^{gt}\right| \qquad (2)$$

Where:

N is the number of test images,

C_i^{pred} is the predicted count for i^{th} image,

C_i^{gt} is the ground truth count for i^{th} image.

A lower MAE indicates better performance.

MAE measures the average absolute difference between predicted and ground truth counts. Mean Squared Error (MSE) is given by:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(C_i^{pred} - C_i^{gt}\right)^2 \qquad (3)$$

This metric is more sensitive to large errors and emphasizes robustness. The Grid Average Mean Absolute Error (GAME) divides the image into grids and computes the error per grid, helping assess spatial accuracy. The R^2 score or correlation coefficient measures how well the predicted count fits the actual trend.

We have used MAE for cell counting due to several reasons. MAE is highly interpretable, providing a direct and intuitive sense of average error. For example, an MAE of 3 indicates an average discrepancy of three cells, which is easily understandable for practitioners. It is symmetric and robust, treating overestimation and underestimation equally, an important aspect in biomedical contexts. Additionally, MAE is less sensitive to outliers compared to MSE, making it suitable for datasets with variable densities. It aligns well with the primary objective of cell counting, which is to accurately estimate the total number of cells. Furthermore, MAE is a standard benchmark metric in literature, used widely in models such as CSRNet [10], Count-ception [4], and SAU-Net [3], enabling consistent comparison across different studies. For density map-based methods, where the total count is derived by integrating over the predicted density map, MAE effectively captures prediction discrepancies, making it ideal for such approaches.

Despite the progress made, several challenges remain in cell counting. Occlusions and overlapping cells, staining and imaging variability, sparse or inconsistent annotations, and the need for 3D microscopy handling are significant hurdles. Additionally, in many datasets, only approximate or noisy labels are available, posing further difficulties.

Emerging trends and future directions are addressing these limitations. Few-shot and transfer learning are being explored to train models with limited data and adapt to new cell types or imaging modalities. For example, FamNet developed by Ranjan et al. (2021) [21] applies few-shot learning to count novel object types using minimal samples. Self-supervised learning is being utilized to leverage unlabelled data for pretext tasks before fine-tuning for cell counting. Transformer-based models are gaining traction due to their ability to leverage global context, providing better accuracy in high-density cell images. Uncertainty modelling is being integrated to handle label noise and rater disagreement. Furthermore, 3D and multimodal integration—combining fluorescence, phase contrast, and volumetric data—is improving model robustness and generalizability. As biological datasets grow in complexity and scale, intelligent models capable of understanding spatial patterns, handling imperfect labels, and generalizing across diverse domains will become increasingly essential. Deep learning, through architectures like CSRNet [10], Count-ception [4], and Transformer-driven models, has firmly established itself as the foundation of modern, scalable, and accurate cell quantification systems.

1.4 Datasets

To evaluate the performance of the proposed deep learning-based cell counting model, three publicly available benchmark datasets were utilized. These datasets represent a diverse range of imaging scenarios and cell morphologies, making them suitable for assessing the generalizability and robustness of cell counting algorithms. Each dataset presents unique challenges in terms of image quality, cell density, shape variation, and background complexity. The following section provides a comprehensive description of each dataset.

1.4.1 Synthetic Bacterial Cells

The Synthetic Bacterial Cell Dataset, commonly referred to as the VGG dataset, was developed by Lempitsky et al. (2010) [7] using a simulation platform initially introduced by Lehmussola et al. (2007) [22], This dataset is hosted and made available by the Visual Geometry Group at the University of Oxford. It comprises 200 synthetic RGB fluorescent microscopy images, each of size 256×256 pixels, collectively containing 35,192 simulated bacterial cells.

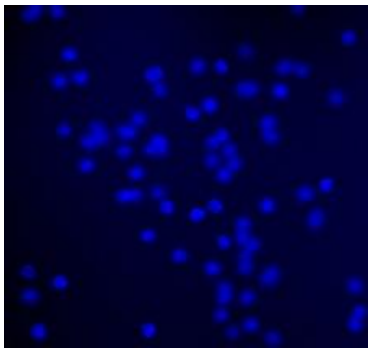

Figure 1.1: VGG Sample Input Image [38]

The dataset was specifically designed to replicate the challenges commonly encountered in automatic cell counting. It incorporates various complex imaging conditions, such as cell clustering, overlaps, and focal depth variations, which emulate real-world microscopy data. Despite being synthetically generated, the dataset maintains high visual fidelity and statistical similarity to real microscopy images, making it an excellent benchmark for evaluating cell counting models in terms of accuracy and generalization. The consistent annotation quality and controlled synthetic environment allow researchers to systematically study model behaviour under challenging scenarios.

1.4.2 Modified Bone Marrow Cells

The Bone Marrow Cell Dataset (MBM) is based on clinical microscopic images and was constructed by Paul et al. (2017) [4] through modifications of an earlier dataset described by another research group. This dataset contains 44 high-resolution RGB images of size 600×600 pixels, derived from bone marrow samples of healthy human subjects.

These samples were stained using Haematoxylin and Eosin (H&E), a widely adopted histological staining method.

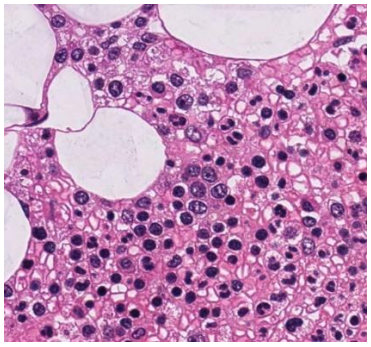

Figure 1.2: MBM Sample Input Image [38]

One of the major challenges associated with the MBM dataset is the complexity of the background, which includes staining artifacts and variations in texture and illumination. These factors make it difficult to isolate and count individual cells. Furthermore, bone marrow samples typically exhibit heterogeneous cell types with varying sizes and densities, further complicating the segmentation and density estimation tasks. The dataset includes a total of 5,553 manually annotated cells, providing a solid ground truth for evaluating the performance of cell counting models under noisy and uneven conditions.

1.4.3 Human Subcutaneous Adipose Tissues

The Adipose Tissue Dataset (ADI) was curated from the Genotype-Tissue Expression (GTEx) Consortium, a large-scale initiative aimed at understanding gene expression across various human tissues. The dataset focuses on subcutaneous adipose tissue and was later adapted and down sampled by Paul et al. (2017) [4] for use in cell counting experiments. The final version of the dataset used in this study consists of 200 RGB microscopy images, each resized to 150×150 pixels, encompassing a total of 29,684 annotated cells.

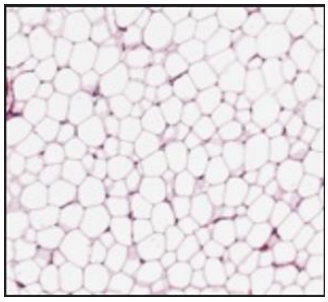

Figure 1.3: ADI Sample Input Image [38]

This dataset poses significant challenges due to the close packing and high density of adipose cells, along with intra-class variability in cell shape and size. Unlike the VGG dataset, which features simulated data, and the MBM dataset, which includes stained histological samples, the ADI dataset highlights the model's ability to generalize to real biological variance. The high visual similarity among adjacent cells and minimal separation boundaries create difficulty in distinguishing and counting individual cells accurately. This makes ADI an important benchmark for testing a model's fine-grained discrimination capabilities in complex tissue environments.

Table I: Dataset details

Dataset	ADI	MBM	VGG
Scenario	Real	Real	Synthetic
Image Size	150 ×150 × 3	600 × 600 × 3	256 ×256 × 3
# of Images	200	44	200

1.5 Organization of the Thesis

This thesis is organized into six chapters, each addressing a key component of the study and development of cell counting using deep learning and transformer-based approaches.

Chapter 1 introduces the fundamental concepts that underpin this work. It begins with an overview of DNNs, followed by a brief introduction to Transformer architectures. The chapter then discusses the significance of cell counting in biomedical imaging and outlines the motivation

behind the study, Datasets used in the project. Finally, the chapter concludes with the organization of the thesis.

Chapter 2 presents a detailed review of the existing literature on object and cell counting methods, covering both classical and modern deep learning-based approaches. This chapter also formulates the core problem addressed in the thesis and identifies the research gaps that this work aims to fill.

Chapter 3 focuses on a deep learning-based approach for cell counting. It provides a comprehensive description of the proposed model architecture, data preprocessing techniques, training strategy, and hyperparameter tuning. The training results are analysed to understand the model's learning behaviour and performance.

Chapter 4 explores a transformer-based approach to cell counting using the Restormer model. It details the model architecture, and the preprocessing pipeline tailored for this method. A special focus is given to the Self-Adaptive Loss Weighting (SALW) technique used to enhance learning. This chapter also discusses the training procedure and evaluates the results.

Chapter 5 consolidates the results from both the deep learning and transformer-based approaches. A comparative analysis is carried out to assess their performance across various datasets and evaluation metrics. This chapter offers insights into the strengths and limitations of each method.

Chapter 6 concludes the thesis by summarizing the key findings and contributions. It also highlights the potential directions for future work, including suggestions for improving model accuracy and generalization, and expanding the approach to other biomedical applications.

The thesis is followed by appendices (if any) that provide supplementary material and technical details and concludes with a comprehensive list of references that support the research conducted in this work.

Chapter 2

Literature Review and Problem Formulation

This section explores recent developments in cell counting methods with help of CNNs and includes a brief discussion on transformer-based approaches.

2.1 Detection based Counting

Detection-based counting methods, such as [7], [23], [20] identify and quantify objects by localizing their centroids or bounding boxes within an image. Traditional techniques relied on handcrafted features and thresholding, but deep learning has significantly improved accuracy. Early approaches combined feature extraction with machine learning models like SVMs and Random Forests developed by Lempitsky et al. (2010) [7], introduced a density map-based method that estimated object positions rather than detecting them directly, later refined by Arteta et al. (2016) [24] using structured learning. However, these methods struggled with occlusions and variations in object appearance.

CNN-based architecture like [25], [18], and [26] improved object localization but faced challenges in dense environments. Rodriguez-Vazquez et al. (2016) [19] addressed this using adversarial training, while [23] developed a deeply supervised CNN (C-FCRN) with auxiliary networks for feature refinement. Xie et al. (2018) [8] enhanced generalization with deep regression models that analyzed spatial relationships

To tackle occlusion-related challenges, researchers explored FCNs and attention mechanisms. Count-ception [4] introduced by [11], which used multiple receptive fields to improve localization, while [17] developed an uncertainty-aware detection model leveraging multi-rater annotations. Falk et al. (2019) [16] demonstrated the effectiveness of U-

Net [11] based segmentation-driven counting. Other studies integrated segmentation with detection for improved accuracy— Cheng et al. (2022) [27] used a spatially relaxed CNN to reduce density map noise.

Hybrid models combining detection and density estimation also emerged. Jiang et al. (2021) [20] proposed a two-path network that extracted spatial details and contextual information before merging them for density estimation. Xue et al. (2016) [28] combined deep regression networks with detection pipelines to enhance accuracy under occlusions. Recent advancements include context-aware detection models and iterative refinement strategies. Guo et al. (2021) [3] extended U-Net [11] with a self-attention mechanism for improved microscopy image localization, while Paulauskaite-Taraseviciene et al. (2019) [9] validated Mask R-CNN [13] for detecting overlapping objects. Despite progress, detection-based methods still struggle with occlusions, low contrast, and irregular object shapes. Future research is exploring transformer-based models Vaswani et al. (2017) [14] and self-supervised learning for better adaptability. Multi-modal fusion approaches are also under investigation, with Zhang et al. (2022) [5] demonstrating the potential of vision transformers for counting tasks.

2.2 Cell Counting by Regression

Regression-based counting methods estimate object counts by mapping input features directly to numerical values, eliminating explicit object detection. This approach is particularly effective for high-density, occluded, and irregularly distributed objects. Early methods relied on handcrafted features with linear regression models, but deep learning significantly improved accuracy. Lempitsky et al. (2010) [7] introduced a density map-based approach using dot-annotated images, while Fiaschi et al. (2012) [29] refined it with structured regression for cluttered environments. Arteta et al. (2014) [24] further enhanced accuracy by incorporating spatial constraints into density maps. With deep learning, CNNs became central to regression-based counting. Xie

et al. (2018) [8] proposed a fully convolutional regression network (FCRN) to estimate density maps directly, while Cohen et al. (2017) [4] introduced Count-ception, leveraging multiple receptive fields to reduce localization errors. He et al. (2021) [23] developed C-FCRN, integrating auxiliary CNNs for feature refinement. Liu and Yang (2017) [30] explored multi-scale CNNs to handle varying object sizes, and Wang et al. (2020) [6] improved generalization with an attention-based multi-scale regression network. Hybrid models have further advanced this approach. Cheng et al. (2022) [27] used a spatially relaxed CNN with Gaussian kernels to reduce density map noise. Walach and Wolf (2016) [31] introduced a deeply supervised network with iterative feedback for refined density estimates. Sindagi et al. (2019) [32] proposed a context-aware regression framework utilizing global and local contextual features for improved density map refinement. Interactive learning has further refined regression-based models. Liu et al. (2023) [30] developed an adaptive density map generator that dynamically adjusted annotations during training. Xue et al. (2016) [28] incorporated feedback mechanisms for iterative prediction refinement, while Zou et al. (2021) [33] introduced uncertainty estimation for better interpretability.

Recent advancements focus on attention mechanisms and improved feature extraction. Guo et al. (2021) [3] extended U-Net [11] with self-attention modules for better density estimation. Optimization techniques and loss function improvements have also enhanced handling of high-density distributions. Future directions include multi-scale learning and transformer-based models for improved adaptability and accuracy in regression-based counting.

2.3 Transformer based Counting

Transformer-based models have significantly improved cell counting in biomedical imaging by effectively handling dense and overlapping objects. Unlike CNNs, they utilize self-attention to capture long-range

dependencies, ensuring robust performance in microscopy images. Their global context modeling enhances accuracy, especially in varied cell distributions. Zhang et al. (2022) [5] developed a vision transformer-based framework that excelled across different cell densities. Cheng et al. (2022) [27] combined transformers with CNNs, achieving state-of-the-art results. Guo et al. (2021) [3] integrated self-attention into U-Net [11], improving segmentation and density estimation. Additionally, Restormer, a transformer model optimized for image restoration, employs a patch-based technique to efficiently process biomedical images, enhancing precision in cell counting tasks. Transformers also improve computational efficiency by processing entire images in parallel, making them ideal for high-resolution medical analysis. However, their high resource demand limits real-time applications. Future efforts should focus on lightweight architecture and self-supervised learning to enhance adaptability, particularly in scenarios with limited labeled data. By offering superior accuracy and scalability, transformers are set to revolutionize biomedical image analysis, enabling more precise and automated cell counting solutions.

2.4 Problem Formulation

To address the critical limitations observed in current methods, the research aims to solve the following condensed challenges:

1. Design a model that accurately counts cells under varying densities and morphological complexities.
2. Ensure robustness against occlusions, overlapping instances, and low-contrast imaging conditions.
3. Achieve computational efficiency suitable for processing high-resolution microscopy images.

This thesis proposes a hybrid deep learning framework that combines convolutional and transformer-based architectures, aiming to extract

multi-scale spatial features, leverage global context via attention mechanisms, and produce precise density maps for robust and scalable cell counting in biomedical images.

Chapter 3

Deep Learning Based Cell Counting

3.1 Data Preprocessing

3.1.1 Focal Inverse Distance Transform Map

In traditional cell counting approaches based on density estimation, Gaussian-based ground truth maps are commonly employed. While this method effectively captures spatial cell distributions in moderately dense images, it suffers from overlapping distributions and diminished localization accuracy in highly clustered regions. To address this limitation, the Focal Inverse Distance Transform (FIDT) map described by Liang et al. (2022) [34] has been introduced as a more precise alternative for annotating cell centroids, particularly in cases of dense and overlapping cell structures.

3.1.2 Limitations of Gaussian Maps

Gaussian maps represent each cell as a localized 2D Gaussian kernel centred at its centroid. For a given point (x_i, y_i) the Gaussian density map $D(x, y)$ is computed as:

$$D(x,y) = \sum_{i=1}^{N} \frac{1}{2\pi\sigma^2}\left(-\frac{(x-x_i)^2 + (y-y_i)^2}{2\sigma^2}\right) \qquad (4)$$

Where:

- N is the total number of annotated cell centers,
- σ is the standard deviation controlling the spread of the kernel.

Although effective in sparse scenes, this method poses two main issues in dense scenarios:

1. Kernel Overlap: When cells are closely packed, the Gaussian kernels tend to overlap significantly, leading to imprecise localization and skewed density representations.

2. Lack of Spatial Contrast: The uniform spread of Gaussians reduces gradient strength around centroids, making it harder for the network to learn sharp and confident peak responses.

3.1.3 Concept and Formulation of FIDT Maps

The FIDT map redefines the way cell centroids are encoded by leveraging inverse distance transforms with an adaptive focal mechanism. It emphasizes the pixel-wise distance to the nearest ground truth centroid and assigns higher values to pixels close to the cell centre, creating sharper and more distinguishable peaks compared to Gaussian maps.

Let Ω denote the spatial domain of the image and let $\mathcal{P} = (x_i, y_i)_{i=1}^{N}$ represent the set of ground truth cell centroids. For each pixel $p \in \Omega$, its distance to the closest point in $\mathcal{P}$ is:

$$d(p) = \min_{(x_i,y_i)\in\mathcal{P}} \left|\left|p - (x_i, y_i)\right|\right|_2 \qquad (5)$$

Then, the FIDT map $M(p)$ is defined using a focal inverse transform:

$$M(p) = \left(\frac{1}{d(p) + 1}\right)^{\gamma} \qquad (6)$$

Where:

- $d(p)$ is the Euclidean distance from pixel p to the nearest ground truth centroid,
- $\gamma > 0$ is a hyperparameter called the focal factor, which controls the steepness and focus of the response around the centroid.

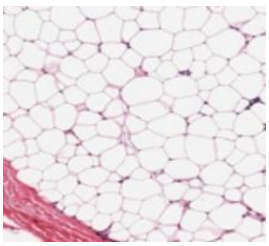

Figure3.1: Input ADI Image [38]

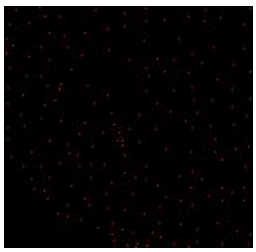

Figure 3.2: Dot Annotation [38]

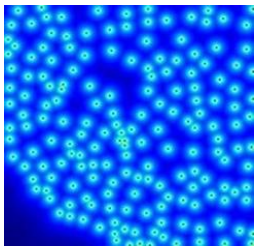

Figure 3.3: FIDT Map

This formulation guarantees that the closer a pixel is to a cell centre, the higher its response in the map. Unlike Gaussian maps, which rely on a fixed spread (σ), the FIDT map dynamically adapts based on pixel-level distances, leading to more robust and distinct peaks.

3.1.4 Mathematical Intuition Behind FIDT

The key innovation of FIDT lies in how it integrates the focal principle—originally proposed in Focal Loss for handling class imbalance—into spatial representation. The exponent γ in the FIDT equation plays a similar role by emphasizing hard (close-to-centre) pixels and down-weighting easier (distant) ones. This introduces spatial adaptivity and sharpens the local maxima corresponding to cell centroids, thereby making learning more effective.

As γ increases:

- The value of $M(p)$ decreases rapidly for pixels farther from the centroid,
- The map becomes more concentrated around the centroid, improving spatial discrimination.

3.1.4 Advantages of FIDT

The use of Focal Inverse Distance Transform (FIDT) maps, as opposed to traditional Gaussian-based maps, offers several key advantages in dense cell counting scenarios. One of the primary benefits is sharper localization, as FIDT maps produce well-defined, focused peaks precisely at cell centres. This characteristic enhances the model's capability to differentiate between closely packed cells (as shown in Figure 3.5), which is particularly important in crowded microscopy images. Additionally, the inverse distance-based formulation of FIDT inherently reduces overlap between adjacent cell representations. Unlike Gaussian maps, which spread density over a larger area and may cause interference (as shown in Figure 3.4), FIDT maps maintain clearer boundaries between neighbouring cells.

Another significant advantage of FIDT is its parameter robustness. Traditional Gaussian density maps require manual tuning of the kernel width parameter (σ), which can vary significantly depending on the dataset. In contrast, FIDT maps eliminate this requirement by defining the spatial profile purely based on distance, simplifying the target generation process and enhancing generalizability across different imaging conditions. Moreover, the sharp gradients produced near the centroids in FIDT maps offer a stronger and more informative learning signal for the model. This leads to faster and more stable convergence during training, ultimately improving the accuracy and efficiency of cell counting systems.

Figure 3.4 illustrates a Gaussian density map, a widely used approach for cell counting where each annotated cell centre is blurred using a Gaussian kernel. This method smooths the spatial distribution and allows for straightforward integration to estimate total cell count. However, as shown in Figure 3.4, the overlapping Gaussian blobs in high-density regions can lead to ambiguity, making it difficult to accurately localize individual cells.

In contrast, Figure 3.5 presents the FIDT map, which encodes spatial information more distinctly by incorporating inverse distance transforms. This results in sharper, non-overlapping peaks even in crowded scenes. As evident in Figure 3.5, FIDT maps preserve spatial resolution and better highlight individual cell centres.

Visualization Example

Figure 3.4: Gaussian Map

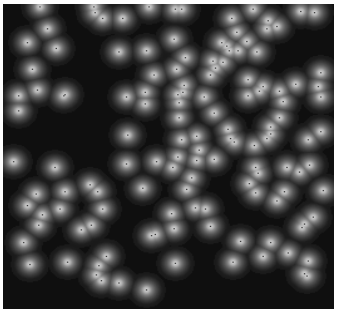

Figure 3.5: FIDT Map

3.1.5 Data Augmentation Strategy

To enhance the robustness and generalizability of the deep learning model, a systematic data augmentation pipeline was applied during the preprocessing stage. The goal of this augmentation was to artificially increase the diversity of the training samples without collecting additional data, which is particularly valuable in biomedical imaging where annotated samples are often limited.

For each original image, three rotation operations were performed at angles of 90°, 180°, and 270°, effectively simulating different orientations of cells that may naturally occur during microscopic slide preparation. Additionally, horizontal and vertical flipping transformations were applied to generate mirror-image variants of the input. To further extend variability, the same three rotation operations (90°, 180°, and 270°) were also applied to each of the horizontally and vertically flipped versions of the image. As a result, a single input image yielded a total of nine augmented variants — three rotations of the original, three rotations of the horizontally flipped image, and three rotations of the vertically flipped image.

This approach not only increased the size of the dataset significantly but also encouraged the model to learn rotation- and flip-invariant features, which is critical for consistent performance across diverse imaging conditions. The augmentation process was carefully designed to ensure that the ground truth annotations (such as cell coordinates or density

maps) were adjusted accordingly, maintaining alignment with the transformed images.

3.2 Model Architecture

In the early phase of this research, a hybrid deep learning framework was implemented to address the challenges inherent in automated cell counting from microscopy images. The proposed architecture integrates a pretrained VGG16 model as the backbone for feature extraction, followed by a cascaded U-Net [11] structure to perform pixel-level prediction through fine-tuned spatial localization. This composite approach was specifically tailored to leverage the generalization ability of pretrained convolutional networks along with the precise segmentation capabilities of encoder-decoder-based architectures, thereby improving the accuracy and robustness of the cell counting task.

3.2.1 VGG16 Feature Extractor

The model begins with a feature extraction block based on VGG16, a convolutional neural network originally introduced by Simonyan et al. (2014) [35]. VGG16 is widely recognized for its simplicity, depth, and effectiveness in learning hierarchical features from image data. It consists of 13 convolutional layers followed by 3 fully connected layers (not used in this case), with all convolutional operations using small 3×3 filters and ReLU activations. In the proposed cell counting framework, the convolutional blocks of VGG16 are used up to the final convolutional layer, excluding the classification head. This allows the model to extract rich semantic representations of cellular structures such as membranes, nuclei, and cytoplasmic regions. These representations serve as a high-level abstraction of the raw input, facilitating more efficient learning during subsequent processing stages.

Using pretrained VGG16 weights, originally learned on the ImageNet [1], [26] dataset, provides a strong initialization, especially beneficial when training data is limited. This technique, known as transfer

learning, helps accelerate convergence and reduces the risk of overfitting, while ensuring that lower-level features such as edges and textures are effectively captured from the start.

3.2.2 Cascaded U-Net Architecture

Following the VGG16 backbone, the model incorporates a U-Net-based encoder-decoder pipeline to reconstruct high-resolution density maps from the abstracted feature maps. U-Net [11], introduced by Ronneberger et al. (2015) [11], was designed specifically for biomedical image segmentation and has since become a cornerstone model in medical imaging tasks due to its strong performance in localization and segmentation. The hallmark of U-Net [11] is its symmetrical structure, consisting of a contracting path (encoder) and an expansive path (decoder), connected through skip connections.

In the encoder path, a series of convolutional layers combined with down sampling operations (e.g., max pooling) progressively reduce the spatial dimensions while increasing the depth of the feature maps, allowing the network to encode semantic information over increasingly larger receptive fields. The decoder path then up samples these feature maps using transpose convolutions (also called up-convolutions) and refines them through additional convolutional operations. Importantly, skip connections link each encoder block with its corresponding decoder block, facilitating the reuse of spatially precise features that may otherwise be lost during down sampling.

This design is particularly advantageous for cell counting, as it ensures the network maintains fine-grained localization information necessary for distinguishing individual cells in high-density or overlapping regions. Additionally, the decoder blocks in this model are configured to generate smooth density maps, enabling accurate estimation of the number and distribution of cells across the image.

3.2.3 Dual Cascaded Network

The uniqueness of this implementation lies in the cascaded U-Net [11] structure, which includes multiple depth and concatenation operations

to refine and merge features from different stages of the network. This design introduces multiple paths for feature flow, allowing the model to learn both global and local context more effectively. Intermediate outputs from early convolutional stages and later encoding stages are concatenated at various points in the decoding process, enriching the model's capacity to discriminate between overlapping or morphologically diverse cells.

Furthermore, the architecture (see Figure 3.6) includes additional layers for transposed convolutions, dropout (for regularization), and batch normalization (to stabilize training). These enhancements contribute to the network's ability to generalize well on unseen data.

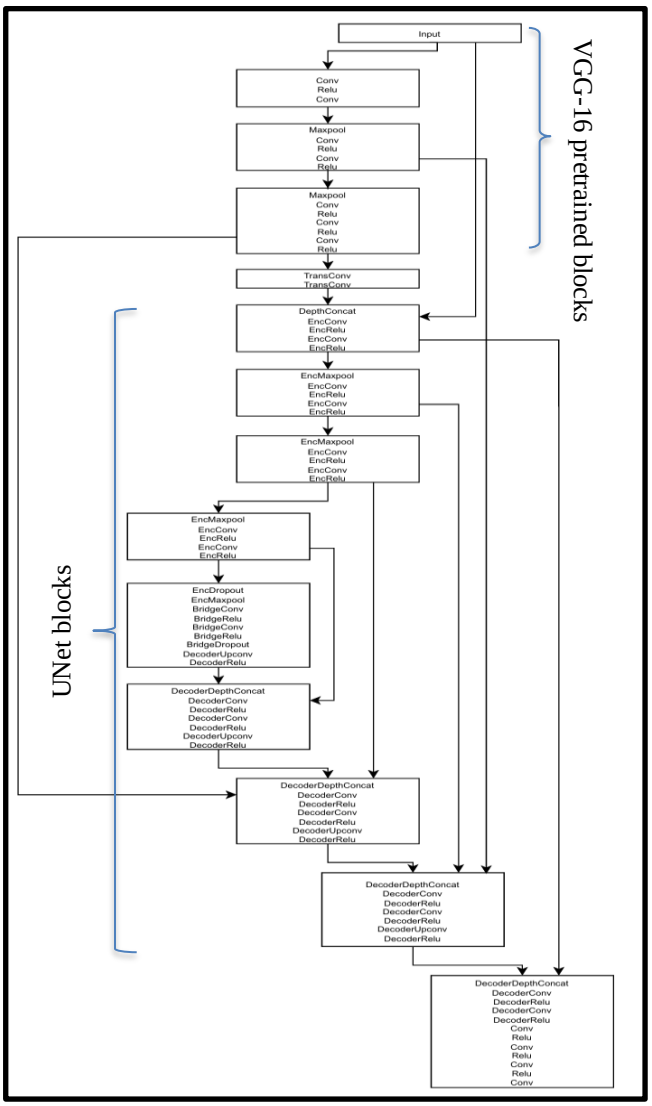

Figure 3.6: Dual Cascaded Network

3.2.4 Key Benefits of the Proposed Model

Despite its limitations in performance, the proposed cascaded VGG16-U-Net model [11] provides several foundational advantages that justify its inclusion in the early stages of this study. One of the key strengths of this architecture lies in its ability to leverage pretrained VGG16 weights, enabling efficient transfer learning. This significantly reduces training time and facilitates faster convergence, particularly when working with limited training data—a common constraint in biomedical imaging. Additionally, the U-Net structure [11], with its characteristic skip connections, plays a critical role in preserving spatial information. This architectural feature ensures that important spatial characteristics, such as cell boundaries and morphological details, are retained throughout the network, which is vital for accurate cell detection.

Furthermore, the model supports end-to-end density map estimation, allowing it to predict continuous-valued density maps directly from input images. This capability is especially beneficial in scenarios where detailed individual cell annotations are unavailable or impractical to obtain. By predicting density maps rather than discrete cell locations, the model can still produce accurate count estimations while circumventing the need for exhaustive manual labelling. These features make the VGG16-U-Net model a valuable baseline for exploring more advanced architectures in the context of automated cell counting.

3.3 Model Training and Hyperparameter Tuning

In the training phase of the proposed deep learning model for cell counting, a range of key hyperparameters were systematically adjusted to optimize predictive performance. The model was trained using supervised learning, where each input image was associated with a corresponding ground truth density map. The experiments focused on the impact of different values of batch size, loss functions, learning rate, and the number of skip connections used in the cascaded DCNet architecture.

Batch size, which determines the number of samples processed before model parameters are updated, was varied across a set of values including 20, 30, 32, 40, 45, and 50. This helped analyse the trade-off between computational efficiency and convergence behaviour. While smaller batch sizes allowed for more granular updates to the model weights, they also increased training time. In contrast, larger batch sizes offered faster iterations but sometimes led to suboptimal convergence due to noisier gradient estimates.

In addition, multiple loss functions were explored, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Huber Loss. MAE penalizes all errors equally and is less sensitive to outliers, whereas MSE penalizes larger errors more heavily, making it suitable for emphasizing high-deviation predictions. Huber Loss serves as a compromise between the two by behaving like MSE near the minimum and like MAE for outliers, making it a balanced choice in cases where data may contain noise or inconsistencies.

Another crucial factor in the optimization process was the learning rate, which controls the size of updates to the model's weights during training. Different learning rates were tested to determine an optimal setting that ensured stable convergence without overshooting the minimum of the loss function. Smaller learning rates led to smoother convergence but required more training epochs, while larger values accelerated convergence at the risk of instability or divergence. By carefully tuning this parameter, the model achieved a more balanced and controlled learning trajectory.

Finally, architectural tuning involved modifying the number of skip connections in the DCNet model. These skip connections are instrumental in preserving spatial and contextual information from earlier layers of the network. By varying the number and placement of these connections, it was possible to assess their influence on the model's ability to reconstruct detailed density maps, particularly in the presence of overlapping or densely packed cells.

Overall, this systematic hyperparameter tuning process helped identify configurations that, while not optimal in all scenarios, contributed to a more stable and interpretable training procedure, laying the groundwork for further refinements in future model iterations.

After extensive experimentation with various combinations of hyperparameters, the most effective configuration was identified. The model yielded its best performance when trained using two skip connections within the cascaded DCNet architecture, a batch size of 45, and the Huber loss function. The use of two skip connections proved to be a balanced choice—it preserved essential spatial features from the encoder layers without overcomplicating the network structure, which could otherwise lead to redundant or conflicting information being passed forward. The batch size of 45 struck an optimal trade-off between stability in gradient updates and training efficiency, enabling smoother convergence while effectively utilizing available computational resources. Moreover, the Huber loss function offered the best results among the tested loss metrics, thanks to its hybrid nature that combines the robustness of Mean Absolute Error (MAE) and the sensitivity of Mean Squared Error (MSE). It particularly improved the model's performance in handling outlier predictions, which are common in cell counting tasks due to the variability in cell density and overlapping structures. This configuration served as the final training setup for subsequent evaluations and benchmarking across the selected datasets.

3.4 Results

During extensive experimentation with various architectural and training configurations, a notable outcome was achieved using a combination of two skip connections, a batch size of 45, and the Huber loss function as the training objective. This particular setup yielded the third-best performance across all tested configurations, with a Mean Absolute Error (MAE) of 11.0.

The use of two skip connections helped in effectively retaining spatial information and preserving fine-grained features through the network layers. Skip connections play a critical role in addressing the vanishing gradient problem and enable the model to learn residual mappings, which in turn enhance the learning of subtle cell structures in high-density regions. Specifically, having two skip connections provided a balanced trade-off between computational complexity and performance, allowing sufficient gradient flow without overcomplicating the model.

The batch size of 45 contributed to stable gradient estimates during training. Larger batch sizes tend to smooth out gradient noise, which can lead to more consistent convergence, while still being small enough to fit within GPU memory constraints and maintain model generalization.

The choice of the Huber loss function further contributed to this performance by providing robustness against outliers. Unlike Mean Squared Error (MSE), which can be overly sensitive to large deviations, the Huber loss behaves like MSE for small errors and like MAE for large errors. This dual nature allowed the model to focus on minimizing smaller, frequent errors while being less influenced by occasional large deviations, which are common in challenging cell counting datasets. The resulting MAE of 11.0 signifies a reasonably accurate prediction, especially in scenarios with dense cell populations and varying imaging conditions. Though not the best overall result, this configuration proved to be highly competitive, suggesting that the combination of skip connections, careful batch sizing, and robust loss functions is effective for deep learning-based cell counting tasks.

The result discussed in this section corresponds to the ADI (Adipose Tissue Imaging) dataset. As part of the evaluation, a series of experiments were conducted to examine the impact of different batch sizes, skip connection configurations, and loss functions on model performance. The configuration that combined two skip connections, a batch size of 45, and the Huber loss function achieved a Mean Absolute Error (MAE) of 11.0, ranking as the third-best result among all tested

setups. This finding highlights the importance of careful selection of architectural components and training parameters in optimizing counting accuracy. Table II shows a comprehensive summary of the results obtained using varying batch sizes, skip connections, and loss functions, allowing for a clear comparison of how each factor contributes to overall model performance on the ADI dataset.

Table II: Experimental Result of DCNet

Batch Size	Number of skip connections	Loss Function	MAE
20	2 skip connections	MSE	14.047
		MAE	12.511
		Huber	15.171
	3 skip connections	MSE	12.418
		MAE	13.565
		Huber	11.398
30	2 skip connections	MSE	12.977
		MAE	13.646
		Huber	11.394
	3 skip connections	MSE	13.199
		MAE	12.478
		Huber	13.64
45	2 skip connections	Huber	11

3.5 Conclusion

In this chapter, we presented a deep learning-based approach for cell counting using a custom-designed model named Dual Cascaded Network (DCNet). The DCNet architecture is a cascaded framework that integrates the feature extraction capabilities of VGG16 with the spatial reconstruction strengths of U-Net [11], enhanced through strategically placed skip connections. This design enables the model to capture both high-level semantic information and low-level spatial features, which is critical for accurate cell localization and counting in complex microscopy images.

The chapter began with an overview of the data preprocessing pipeline, where raw microscopy images were normalized, resized, and converted into density maps using dot annotations. These steps ensured that the

input data was well-prepared for training and that the density maps provided rich spatial supervision.

Next, the model architecture of DCNet was detailed. The encoder is based on VGG16, which effectively captures hierarchical features, while the decoder follows the U-Net [11] structure to reconstruct high-resolution density maps. The incorporation of skip connections between encoder and decoder blocks helps in preserving spatial information and facilitates the learning of fine cell boundaries.

We also discussed the training procedure and hyperparameter tuning, wherein multiple configurations of batch size, skip connections, and loss functions were explored. Among these, the combination of two skip connections, a batch size of 45, and Huber loss proved particularly effective, striking a balance between stability and robustness to outliers.

The training results, evaluated on the ADI dataset, demonstrated that this configuration achieved the third-best performance across all tested setups, with a Mean Absolute Error (MAE) of 11.0. These findings underscore the importance of architectural choices and training strategy in achieving high accuracy in cell counting tasks. A comparative summary of various experimental settings is provided in Table II for reference.

In summary, this chapter demonstrated the effectiveness of the proposed DCNet architecture for cell counting in microscopy images. The cascaded use of VGG16 and U-Net [11], combined with thoughtful training configurations, enabled strong performance across key evaluation metrics. The insights gained here serve as a solid foundation for the next chapter, which investigates a transformer-based alternative—Restormer—to further enhance the model's ability to capture long-range dependencies and global context.

Chapter 4

Transformer Based Cell Counting

4.1 Data Preprocessing

To ensure robust training and accurate generalization of the Restormer-based cell counting model, a comprehensive data preprocessing pipeline was established. Given the inherent differences in resolution, staining, and image quality across the datasets (described in section 1.4), careful normalization, augmentation, and target map preparation steps were performed. These preprocessing techniques aimed to harmonize the data characteristics and enhance the model's ability to detect and count cells under varying imaging conditions.

4.1.1 Image Normalization and Standardization

All input images were resized to fixed dimensions compatible with the Restormer architecture. Pixel intensities were normalized to a standard range to remove variations due to illumination and sensor-specific characteristics. This helped in maintaining consistency during training and ensured stable learning dynamics.

4.1.2 Data Augmentation Strategies

To enhance the robustness of the model and reduce the risk of overfitting, a comprehensive set of data augmentation techniques was employed during the training phase. These augmentations included both geometric and photometric transformations, ensuring that the model could generalize well across diverse imaging conditions and cell morphologies.

Geometric augmentation involved applying random rotations to the input images at angles of 90°, 180°, and 270°. This helped the model develop invariance to cell orientation and positional variations, which is crucial in microscopy images where cells can appear in any direction.

Photometric augmentation was implemented to address variations in staining protocols and illumination conditions observed across the

datasets. Several colour-based adjustments were applied to simulate these differences. Brightness adjustment was used to mimic changes in image exposure by randomly increasing or decreasing the overall brightness, allowing the model to become more resilient to lighting inconsistencies. Colour jittering introduced random variability in hue and saturation levels, which aided the model in learning features that are invariant to differences in staining. Additionally, specific shifts in hue and saturation were included to further improve generalization, enabling the model to adapt to the colour distortions typically introduced by different imaging setups. These augmentations collectively contributed to a more generalized and reliable performance across multiple datasets.

These augmentations were applied randomly during training, ensuring that the model encountered a wide variety of input styles and conditions.

To address the issue of boundary-region cells being underrepresented during training, padding was applied to all images prior to density map generation. Without padding, cells located near the image edges were often partially excluded from the receptive field of the network, leading to inaccurate or incomplete density predictions. By extending the image borders through symmetric padding, we ensured that edge-region cells were fully included in both the input and corresponding FIDT maps. This strategy improved the model's ability to learn from the entire spatial extent of the image, including border regions where cells are frequently present but previously overlooked.

4.1.3 Super-Resolution Enhancement (for ADI Dataset)

The ADI dataset, due to its low original resolution and poor visibility of cell boundaries, posed a significant challenge for learning fine spatial features. To address this, a super-resolution [36] module was applied as a preprocessing step. This module enhanced the image resolution, allowing the model to better observe and learn from subtle details such as cell contours, textures, and edge information. By improving boundary visibility, super-resolution enabled more accurate feature extraction during training, ultimately leading to better performance on ADI data.

This step was applied only to the ADI dataset, as the MBM and VGG datasets already had sufficient resolution for accurate processing.

4.1.4 Density Map Generation Using FIDT Maps

Instead of relying on traditional Gaussian-based density maps, this study utilized the Focal Inverse Distance Transform (FIDT) maps for target generation, as proposed by Liang et al. (2022) [34]. The FIDT methodology enables the creation of density maps that better capture the true spatial distribution of cellular regions, particularly in images with dense or overlapping cell populations. This approach adapts the density spread dynamically based on the local proximity of cells, leading to a more informative and context-aware representation of cellular arrangements.

One of the key advantages of using FIDT maps is their ability to preserve spatial distribution information by modulating the density spread relative to nearby cell locations. This characteristic allows for a more precise representation of both isolated and clustered cells within the same image. Additionally, the FIDT-based maps enhance the quality of the supervision signal, thereby simplifying the learning process for models tasked with accurate density estimation.

The generation process begins with dot annotations, which mark the centroids of individual cells. These annotations are then convolved with a distance-adaptive kernel as defined by the FIDT algorithm. The resulting FIDT map serves as the ground truth for the regression task, with the integral over the entire map representing the total cell count for the image. Compared to conventional fixed-kernel approaches, FIDT maps offer a richer and more flexible depiction of cell distributions, making them particularly suitable for datasets such as the Modified Bone Marrow (MBM) and Adipose Tissue (ADI) datasets, where cell layouts are often irregular and non-uniform.

4.1.5 Final Input-Target Pipeline

Following all preprocessing steps, the final paired data used for training comprised two main components: the input image and the corresponding

target map. The input image consisted of an augmented microscopy image, and in the case of the ADI dataset, it also underwent super-resolution enhancement to address challenges related to low visibility and indistinct cellular structures. The target map was a Focal Inverse Distance Transform (FIDT)-based density map, designed to accurately represent the spatial distribution of cells within the image. These image-target pairs were employed to train the Restormer model in a supervised manner using a pixel-wise regression loss function.

The preprocessing pipeline integrated a combination of strategies tailored to accommodate the varying characteristics of the ADI, MBM, and VGG datasets. The inclusion of super-resolution techniques for the ADI dataset significantly improved the visibility and definition of cell features, which are often compromised in lower-quality images. Simultaneously, the use of FIDT maps in place of conventional Gaussian-based density maps provided the model with richer and more informative supervisory signals that reflected the actual distribution of cellular regions more accurately. In addition, extensive photometric augmentation—encompassing modifications in brightness, colour, hue, and saturation—was applied to simulate variations in staining protocols and imaging conditions. These augmentations enabled the model to develop robustness and generalization capabilities across different imaging artifacts. Collectively, these preprocessing methods established a robust and comprehensive foundation for training the Restormer-based cell counting architecture.

4.2 Restormer Model Architecture

4.2.1 Introduction to Restormer

Restormer (Restoration Transformer) [12] is a novel transformer-based architecture introduced to address the challenges of high-resolution image restoration tasks such as denoising, deraining and deblurring (as shown in Figure 4.4). Unlike traditional CNNs, which have a limited receptive field and often struggle to model long-range dependencies in

images, Restormer leverages the power of self-attention mechanisms to model global context while maintaining computational efficiency.

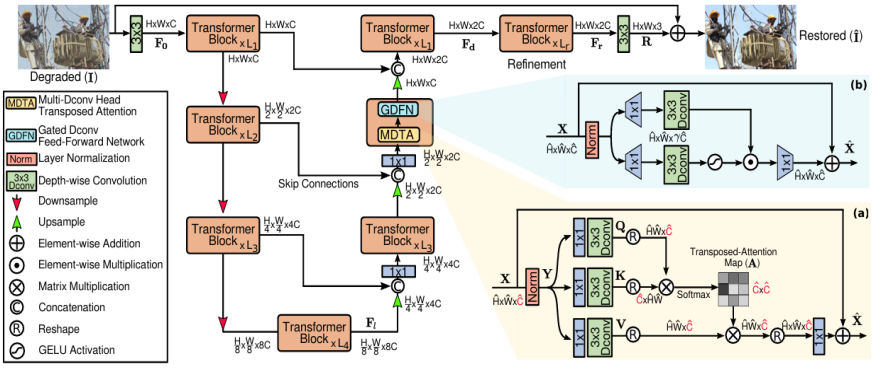

Figure 4.1: Restormer Architecture [12]

In the domain of cell counting, especially in biomedical microscopy images, challenges such as dense cell populations, varying cell morphology, noise, and overlapping structures make accurate prediction complex. Traditional CNN-based models may not fully capture the spatial relationships between distant but correlated regions. Restormer's architecture, with its attention-based mechanism, offers a strong alternative by providing a global understanding of the image while still preserving fine spatial details.

4.2.2 Restormer Architectural Components

Restormer is founded on a hierarchical encoder-decoder Transformer architecture, designed to efficiently handle high-resolution image restoration tasks, and repurposed in this study for cell counting. One of its core innovations is the Multi-Dconv Head Transposed Attention (MDTA) module. Unlike traditional Transformers, which compute attention across all pixel pairs—leading to high computational overhead for large images—MDTA integrates depth-wise convolution (D-Conv) within the attention mechanism. This design enables the model to capture local spatial context before computing attention, offering a more scalable alternative to vanilla self-attention. By operating in a spatially-aware manner, MDTA significantly reduces computational complexity. In the context of cell counting, particularly in microscopy images where cells are small and densely clustered, MDTA enables the model to

effectively focus on biologically relevant features while being sensitive to local variations in cell morphology, intensity, and density. This capability greatly enhances both localization and counting accuracy.

Each Transformer block in Restormer also includes a specialized feed-forward module known as the Gated Depth-wise Convolutional Feed-forward Network (GDFN). This component consists of depth-wise separable convolutions, which extract spatial information in a channel-wise manner, and a gating mechanism that dynamically controls the flow of information. The gating mechanism plays a crucial role in filtering out irrelevant background noise and emphasizing important image features such as cell edges and centres. For cell counting tasks, this targeted attention to biologically meaningful structures results in improved precision, especially in images affected by noise, variable illumination, or blur.

Restormer employs a hierarchical encoder-decoder framework similar to that of U-Net [11], which supports multi-scale feature learning. The encoder path progressively down samples the input image to extract deep semantic features, while the decoder path up samples the data to reconstruct the original spatial resolution. Crucially, skip connections bridge the encoder and decoder layers at corresponding levels, ensuring that high-frequency spatial information—such as edges and textures—is preserved throughout the network. This architectural design is particularly advantageous for cell counting, as cells exhibit a wide range of sizes and intensities. The hierarchical approach enables detection of both small, faint cells and larger cellular structures, while the skip connections help retain the fine-grained details necessary for accurate localization.

Additionally, Restormer replaces standard position encodings with Gated Positional Encodings, allowing the model to learn positional relevance in a dynamic fashion rather than relying on fixed embeddings. This is coupled with Layer Normalization to improve training stability. In microscopy images, where cells often appear in complex spatial

configurations or are partially overlapping, the model's ability to understand spatial relationships dynamically is essential for precise and reliable cell count estimation.

4.2.3 Why Restormer Is Effective for Cell Counting

Although Restormer was originally designed for image restoration tasks, it offers several advantages that make it suitable and highly effective for cell counting in biomedical images:

1. Global Context Understanding

Cell counting often involves recognizing patterns in cell distribution across the entire image. Unlike CNNs, which only process local regions at a time, Restormer can attend to features across the full spatial range, enabling better estimation of cell counts even in images with uneven cell distribution or overlapping regions.

2. Fine Detail Preservation

The model is highly capable of preserving high-frequency details such as cell boundaries, shapes, and edges—features that are essential for differentiating individual cells, particularly in dense clusters.

3. Robustness to Noise and Artifacts

Microscopy images are prone to imaging artifacts, intensity variations, and noise due to experimental limitations. Given its origin in restoration tasks, Restormer naturally handles noise well, allowing it to generate cleaner density maps and reducing false positives/negatives in cell counting.

4. Efficient Processing of High-Resolution Images

Cell datasets frequently consist of high-resolution images to preserve intricate cellular details. However, traditional Transformer models become impractical at such scales due to their extensive memory requirements. Restormer overcomes this limitation by incorporating efficient attention mechanisms such as Multi-Dconv Head Transposed

Attention (MDTA) and utilizing depth-wise convolutions, which significantly reduce computational complexity. These design choices enable the model to scale effectively to high-resolution inputs while maintaining a manageable computational load.

4.2.4 Adaptation for Cell Counting in This Work

In this research, the Restormer architecture was adapted to perform density map regression for the purpose of cell counting. The model was trained using microscopy images annotated with dot annotations representing individual cell locations. The primary objective was to predict a continuous-valued density map, where the integral over any region of the image accurately estimates the number of cells present.

Several key adaptations were made to tailor Restormer for this task. First, the traditional restoration target used in image restoration tasks was replaced with density maps, enabling the model to learn to predict spatial cell distributions rather than denoised images. Additionally, SALW [37] was employed during training. This technique dynamically adjusts the loss contributions from different regions of the image, allowing the model to focus more on complex or high-error areas, which is particularly beneficial in dense or noisy microscopy images. The model was further fine-tuned on domain-specific datasets such as ADI, MBM, and VGG, which encompass a wide range of cell types and real-world imaging challenges.

Restormer's architecture, which combines global attention with computational efficiency and precise spatial feature preservation, presents a novel and powerful solution for the cell counting problem. Its ability to simultaneously capture fine local detail and broader spatial context provides a distinct advantage over conventional CNN-based approaches. The subsequent sections of this chapter delve into the training methodology, dataset-specific evaluations, and performance comparisons with other state-of-the-art models.

4.3 Self Adaptive Loss Weighting

4.3.1 Motivation and Background

In supervised deep learning tasks like cell counting, the goal is to minimize a loss function that quantifies the difference between the model's predictions and the ground truth. In many real-world applications—including microscopy-based cell counting—the quality of the input data can vary significantly across samples and datasets due to factors such as image resolution, staining protocols, contrast variability, and background noise. Some regions of an image may be easy for the model to learn (e.g., clear, well-defined cells), while others are ambiguous (e.g., overlapping cells, faint boundaries, or low-intensity regions).

Using a fixed, static loss weight across all training samples or loss components fails to account for these disparities in learning difficulty. In such cases, the model may overly prioritize regions it finds easy to learn, while underfitting the more challenging ones. To overcome this imbalance, we integrate a SALW [37] approach, which allows the model to dynamically and continuously adjust how much attention it pays to the loss at each stage of training.

This dynamic adjustment is achieved by introducing a learnable parameter into the loss function that controls the scale of the loss based on the model's confidence in the prediction. The mathematical framework is inspired by probabilistic modelling and uncertainty weighting introduced by Kendall et al. (2018) [15].

4.3.2 Theoretical Foundation

The adaptive loss function is defined as:

$$L_{total} = exp(-a) \cdot L_{main} + a \qquad (7)$$

Where:

- L_{main} is the primary task loss (e.g., Mean Absolute Error or MSE computed between the predicted and ground truth density maps).
- $a \in \mathbb{R}$ is a scalar parameter learned through backpropagation.
- $exp(-a)$ dynamically adjusts the weight of the loss.
- The additive term a acts as a regularizer, preventing a from becoming too large or too small during optimization.

This formulation arises from modelling the output of the network as a Gaussian distribution and optimizing its log-likelihood. Here, $exp(-a)$ corresponds to the inverse variance (i.e., the precision or confidence) in the model's predictions. The idea is that if the model is uncertain (i.e., has higher variance), it should penalize that prediction less, and vice versa.

This approach is rooted in probabilistic principles and allows the model to adaptively scale the loss without manual intervention.

4.3.3 Intuition Behind the Adaptive Term

The key strength of SALW [37] lies in its ability to make the loss scale learnable. In this approach, the parameter a is optimized jointly with the model weights through gradient descent. This dynamic adjustment allows the network to autonomously assess and adapt its confidence in predictions over the course of training. If the model encounters difficulty in reducing the primary loss in specific regions of the image—such as areas affected by noise or blur—it will learn to increase the value of a, thereby reducing the penalty for prediction errors in those challenging regions. On the other hand, when the model performs well in cleaner, more reliable sections of the image, it decreases a, amplifying the contribution of those areas to the total loss. This mechanism effectively introduces a form of automatic curriculum learning, enabling the

network to progressively focus on more difficult parts of the task while maintaining balanced learning pressure across the image.

4.3.4 Application in Restormer-Based Cell Counting

In this work, the SALW mechanism [37] is integrated into the Restormer-based cell counting model, which is trained on a range of datasets including ADI, MBM, and VGG. These datasets present varying levels of image quality and structural complexity, necessitating a flexible learning approach. SALW is specifically applied to the loss function used for training the model to predict FIDT-based density maps, which are designed to represent the spatial distribution of cells within each image. Given that FIDT maps are highly sensitive to factors such as annotation quality, image resolution, and local cell density, the difficulty associated with learning from them can vary significantly across different image patches or entire datasets.

The SALW-enhanced loss is implemented using an adaptive formulation in which the total loss L_{total} is computed with a learnable parameter a. This parameter is initialized to a default value (e.g., 0.0) and is subsequently updated during training through gradient backpropagation alongside the other model parameters. This setup enables the model to dynamically adjust the importance assigned to different regions or tasks throughout training, depending on the level of prediction uncertainty.

For instance, in the ADI dataset—characterized by low-resolution images and indistinct cell boundaries even after super-resolution enhancement—the model learns to reduce the weight of regions with high uncertainty. Conversely, in the VGG dataset, which contains synthetic, high-resolution images with clearly defined cell structures, the model increases the emphasis on confident predictions. This leads to more stable learning and improved convergence, as the model adaptively focuses on reliable data while mitigating the impact of ambiguous or noisy regions.

4.3.5 Benefits in Cell Counting Context

Integrating SALW [37] into the cell counting pipeline introduces several notable advantages. One of the most significant benefits is the model's ability to dynamically focus on challenging regions within the image. Without requiring any manual annotation or predefined region-specific weighting, SALW enables the network to automatically allocate more attention to harder-to-learn areas, such as regions with overlapping cells, low contrast, or inconsistent staining. This adaptability is particularly valuable in biomedical image analysis, where dataset characteristics often vary widely in terms of resolution, contrast, and cellular morphology. As a result, SALW contributes to better generalization across diverse datasets.

Another advantage lies in the reduction of manual tuning typically associated with traditional loss-weighting strategies. Conventional methods often rely on extensive hyperparameter searches to determine static loss weights, which can be both time-consuming and suboptimal. In contrast, SALW learns the optimal loss scaling dynamically during training, streamlining the process and improving performance. Moreover, SALW introduces uncertainty-aware learning by explicitly modelling and responding to prediction uncertainty. This capability enhances the model's robustness and reliability when applied to real-world microscopy data, which often includes noise, artifacts, and ambiguous cellular structures.

4.3.6 Summary

Self-Adaptive Loss Weighting is a principled and effective strategy to dynamically balance learning focus during model training. In our work, its integration into the Restormer-based architecture enhances the model’s ability to learn from noisy, low-resolution, or visually ambiguous images common in biological microscopy. It supports more stable and efficient training and leads to better overall performance in terms of cell count accuracy and density map quality.

4.4 Model Training and Hyperparameter Tuning

4.4.1 Training Overview

The training of the proposed Restormer-based cell counting model was conducted with the goal of achieving accurate and generalizable density map predictions across multiple microscopy datasets (ADI, MBM, and VGG). The model was trained for a total of 300,000 iterations, allowing adequate time for convergence even on complex, high-resolution biomedical images.

A base learning rate of 0.00001 was used, optimized using a cosine annealing learning rate schedule. This strategy gradually decreases the learning rate, promoting stable convergence and avoiding abrupt gradient oscillations. Additionally, specific learning rate milestones were defined at [92,000, 150,000, 200,000, 250,000, 300,000] to control the decay curve more precisely during the training process.

4.4.2 Architecture Configuration

To fully leverage the capabilities of the Restormer architecture, several architectural hyperparameters were carefully configured. The patch size was set to a progressive hierarchy of [32, 64, 64, 128, 256], enabling effective hierarchical feature extraction from fine to coarse levels of resolution. This configuration facilitates multi-scale analysis, which is crucial in capturing both small and large cellular features. A consistent batch size of [2, 2, 2, 2, 2] was maintained across all training stages to ensure training efficiency while accommodating the high memory requirements of processing high-resolution images.

Each stage of the model was designed with [2, 2, 2, 2, 2] Transformer blocks, providing adequate depth to learn complex spatial dependencies present in microscopy images. The number of attention heads was also fixed at [2, 2, 2, 2, 2] for each stage, striking a balance between capturing diverse attention patterns and maintaining computational efficiency. Channel dimensions were incrementally set to [64, 128, 256, 512], allowing for increased feature representation capacity at deeper layers. An expansion factor of 2 was applied in the feed-forward networks,

enhancing the intermediate feature space and improving the model's ability to encode rich information. Additionally, two refinement stages were incorporated to iteratively refine the predicted density maps, leading to more accurate and spatially coherent cell count estimations.

4.4.3 Loss Function Evaluation

To determine the most effective supervisory signal for training the Restormer-based cell counting model, several regression loss functions were explored. These included Mean Absolute Error (MAE or L1 Loss), Mean Squared Error (MSE or L2 Loss), and Huber Loss. MAE focuses on minimizing the average absolute difference between the predicted and ground truth density values, making it straightforward and interpretable. MSE, on the other hand, penalizes larger deviations more severely, making it particularly useful for emphasizing and correcting substantial prediction errors. Huber Loss blends the strengths of both MAE and MSE, offering robustness against outliers while maintaining smooth optimization behaviour.

Through extensive experimentation, it was observed that while each of these loss functions produced reasonable outcomes, the combination of MAE (L1 Loss) with the SALW mechanism [37] consistently yielded superior results in terms of both training stability and model accuracy. SALW played a critical role by dynamically adjusting the influence of the loss throughout the training process. This enabled the model to concentrate more effectively on challenging or uncertain regions within the image, such as overlapping cells or areas characterized by low contrast. The adaptive behaviour facilitated by SALW was especially advantageous when working with diverse datasets, where variations in input quality and annotation consistency could otherwise hinder model performance.

4.4.4 Optimization Strategy and Implementation

Training was performed using the AdamW optimizer, known for its adaptive gradient updates and efficiency in deep architectures. To improve training throughput, 8 parallel workers were utilized for data

loading. Additionally, a seed value of -1 was used to introduce randomized initialization, contributing to the robustness of the training process.

We have observed the self adaptive loss weighting trainable parameter (described by equation 7) and got decreasing value pattern (shown in figure 4.5).

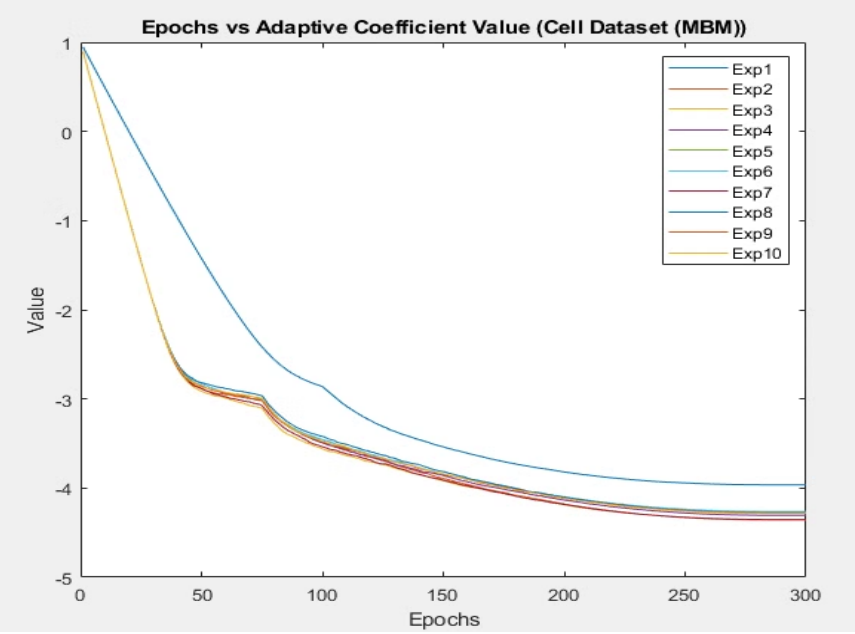

Figure 4.2: Adaptive Parameter Value Learning Trend

The above trend is observed for MBM dataset and similar type of trend is observed another datasets ADI and VGG.

In conclusion, the training procedure for the Restormer model was carefully designed to balance depth, efficiency, and generalizability. The use of a cosine annealing learning schedule, hierarchical architectural design, and dynamic loss weighting via SALW [37] created a highly effective training pipeline. Among all tested loss functions, L1 Loss combined with SALW emerged as the most effective, leading to improved prediction accuracy and robustness across datasets. This setup allowed the model to learn both localized and distributed cell patterns efficiently, enabling accurate and consistent cell counting in varied imaging conditions.

4.5 Counting Algorithm

4.5.1 Introduction

In addition to deep learning and transformer-based approaches, we also implemented a classical image processing method for cell counting based on a modified Laplacian of Gaussian (LoG) as used in [19], technique. We name our approach LoIG — which stands for Laplacian of Inverse-Gaussian. This method aims to enhance the contrast of faint or poorly visible cell structures by introducing an inverse operation between Gaussian smoothing and Laplacian edge detection.

The key idea is to suppress the influence of bright background areas and emphasize dark cellular regions before applying edge enhancement. This adjustment helps improve blob detection accuracy, especially in images where cells appear as dark regions against a brighter or uneven background.

4.5.2 Method Overview

The standard LoG [19] algorithm works by first smoothing the image with a Gaussian filter to reduce noise, followed by applying the Laplacian operator to detect regions of rapid intensity change, i.e., potential blob centres.

In the LoIG algorithm, we modify this pipeline as follows:

1. **Apply Gaussian Filter**:
    - The input image is first smoothed using a Gaussian filter with standard deviation σ to reduce high-frequency noise.

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \qquad (8)$$

Let the smoothed image be $I_G(x, y)$.

2. **Inverse Operation**:

    - After smoothing, the image is inverted to highlight low-intensity regions (typically cells) and suppress background areas.

$$I_{inv}(x, y) = 1 - I_G(x, y) \qquad (9)$$

This step is particularly useful for datasets where cells appear darker than the background, as it improves contrast before edge detection.

3. **Apply Laplacian Operator**:

    - The Laplacian operator is then applied to the inverted image to detect intensity transitions and highlight potential cell regions.

$$LoIG(x, y) = \nabla^2 I_{inv}(x, y) \qquad (10)$$

Where ∇^2 denotes the Laplacian operator (second spatial derivative).

4.5.3 Combined LoIG Equation

Combining all steps, the complete **LoIG transformation** can be expressed as:

$$LoIG(x, y) = \nabla^2\left[1 - \left(I * G(x, y, \sigma)\right)\right] \qquad (11)$$

Where:

- I is the input grayscale image,
- $G(x, y, \sigma)$ is the Gaussian kernel,
- $*$ denotes convolution,
- ∇^2 is the Laplacian operator.

4.5.4 Cell Counting Using LoIG

After computing the Laplacian of Inverted Gaussian (LoIG) map, the cell counting process proceeds through a series of classical image

processing steps. First, thresholding is applied—either globally or adaptively—to convert the LoIG map into a binary image. This step is essential for distinguishing foreground cellular structures from the background. Next, blob detection is performed using methods such as connected components analysis or local maxima detection to identify distinct blobs that correspond to individual cells. Finally, the number of detected blobs is counted to estimate the total number of cells present in the image.

The LoIG algorithm offers several notable advantages and specific use cases. One of its key benefits is its enhanced performance in low-contrast conditions. The inverse operation enhances the visibility of dark or faint cellular structures, which are frequently encountered in microscopy images. Additionally, the method is computationally simple, requiring no training or manual annotation, making it accessible for use in low-resource settings. It is particularly effective in images where cell boundaries are rich in edge information and become more pronounced following enhancement.

However, the method is not without limitations. It is highly sensitive to parameter tuning; for instance, the sigma value used in the Gaussian filter and the threshold level must be carefully chosen to ensure optimal performance. Moreover, the LoIG approach is not well-suited for images with densely overlapping cells or highly irregular shapes, where it may fail to separate adjacent structures accurately. Another major drawback is the lack of learning capability; the algorithm cannot adapt to variations in imaging styles unless the parameters are manually re-tuned for each new condition.

Despite these constraints, the LoIG algorithm presents a lightweight and interpretable alternative to more complex, data-driven cell counting models. By applying an inverse operation after Gaussian smoothing, it effectively enhances low-intensity cellular features and improves edge detection. While it may not achieve the same level of accuracy as deep learning-based approaches in challenging scenarios, it provides a

practical and effective solution in controlled environments and serves as a valuable baseline for evaluating more advanced methods.

4.6 Results

One of the core objectives of this research was to determine the most effective training configuration for accurate and generalizable cell counting across a variety of microscopy datasets. Through extensive experimentation involving multiple loss functions—namely Mean Absolute Error (L1 Loss), Mean Squared Error (L2 Loss), and Huber Loss—we found that the L1 loss combined with SALW [37] consistently provided the most stable training and superior performance across datasets.

Unlike traditional loss functions with fixed weights, the SALW mechanism dynamically adjusts the influence of the loss during training based on prediction uncertainty. This adaptiveness proved highly beneficial in microscopy images, where different regions of the image may vary significantly in complexity, contrast, and cell distribution. L1 loss, being robust to outliers and focused on minimizing absolute error, worked synergistically with SALW to help the model emphasize difficult-to-learn regions while not being overly influenced by isolated errors.

4.6.1 Dataset-wise Results

The effectiveness of the proposed configuration, combining Mean Absolute Error (L1 Loss) with SALW [37], was evaluated across three distinct datasets—ADI, MBM, and VGG—each offering unique visual and structural characteristics. In the ADI dataset, which comprises low-resolution fluorescence microscopy images with faint and poorly defined cell boundaries, the L1 + SALW configuration demonstrated robust performance by securing the third-best result among all tested variants. The application of super-resolution preprocessing techniques played a crucial role in enhancing the visual quality of the input images,

while SALW contributed to better learning outcomes in regions with limited visibility and weak signal contrast.

In the MBM dataset, which features moderate-to-high density bone marrow images with a wide range of cell sizes and significant overlap between cells, the model configured with L1 + SALW delivered the second-best performance. The dynamic loss weighting enabled the model to effectively adapt to the heterogeneous distribution of cells and prevented overfitting to highly clustered regions, ensuring more generalized predictions.

The VGG dataset, composed of synthetic, high-resolution images with clearly defined cell boundaries and consistent intensity distributions, presented an ideal testing environment. Under these optimal conditions, the L1 + SALW configuration achieved the best overall performance. The model was able to converge rapidly and accurately, leveraging the uniformity and precision of the dataset's annotations. This cross-dataset evaluation underscores the adaptability and effectiveness of the L1 + SALW combination across a range of biomedical imaging scenarios.

Table III: Result of Restormer Model

Model	ADI (MAE) N=50	MBM (MAE) N=15	VGG (MAE) N=50
Ciampi et al. (2022) [17]	**8.7±0.8**	**5.7±0.9**	2.5±0.1
Count-ception, Paul et al. (2017) [4]	19.4±2.2	8.3±2.3	2.3±0.4
Jiang and Yu (2021) [6]	10.6±0.3	7.5±0.7	2.2±0.2
Rodriguez-Vazquez et al. (2022) [19]	17.3±3.6	**4.2±2.4**	**2.2±0.5**
Ours	**11±0.2**	5.3±1.0	**2.09±0.08**

The use of L1 loss optimized with Self-Adaptive Loss Weighting enabled the Restormer model to adapt to the challenges posed by each dataset. The performance across ADI, MBM, and VGG clearly demonstrates (in Table III) the versatility and generalization capability of this approach. While other loss functions like MSE and Huber showed reasonable performance, they lacked the adaptive robustness required to handle the varied levels of noise, resolution, and complexity present in real-world microscopy images. The result underscores the effectiveness of combining absolute error minimization with uncertainty-aware training dynamics for cell counting tasks.

4.7 Conclusion

This chapter presented a transformer-based approach for cell counting using the Restormer architecture, originally designed for image restoration tasks. Through architectural adaptation and extensive training, we demonstrated the model's potential in handling the unique challenges of microscopy image analysis, including varying resolution, noise levels, and cell densities.

The hierarchical structure of Restormer, equipped with Multi-Dconv Head Transposed Attention (MDTA) and Gated Depth-wise Feed-forward Networks (GDFNs), allowed the model to capture both fine-grained local details and global spatial dependencies effectively. These capabilities proved essential in accurately predicting cell density maps, particularly in cases of overlapping or faintly stained cells.

A robust data preprocessing pipeline was designed to standardize inputs across three datasets—ADI, MBM, and VGG—which varied significantly in terms of resolution and visual quality. Key steps such as super-resolution (for ADI), padding, and extensive augmentations helped enhance model generalization. The use of FIDT maps as the regression target provided a more adaptive and spatially aware supervision signal than conventional Gaussian-based density maps.

To further optimize the training process, we employed SALW based on uncertainty modelling. This strategy allowed the model to dynamically adjust its focus during training, placing more emphasis on harder-to-learn regions. Among various loss functions evaluated, the combination of L1 loss with SALW yielded the most consistent and accurate results across datasets.

Through rigorous hyperparameter tuning—including variations in patch sizes, transformer depth, attention heads, and channel dimensions—the model achieved competitive performance: best results on the VGG dataset, second-best on MBM, and a strong third-best on the challenging ADI dataset, even with its initial low-resolution limitations.

In summary, this chapter validated the feasibility and effectiveness of adapting a transformer-based architecture for the task of cell counting. The integration of attention mechanisms, adaptive loss strategies, and customized preprocessing steps collectively contributed to high accuracy and generalizability, setting a strong foundation for further research and enhancements in biomedical image analysis.

Chapter 5

Results and Discussion

This chapter presents the experimental results and comparative analysis of two distinct approaches to biological cell counting: the deep learning-based **DCNet** architecture and the transformer-based **Restormer** model. The performance evaluation is based on Mean Absolute Error (MAE) across three microscopy datasets—ADI, MBM, and VGG—with additional discussion on the qualitative aspects, training dynamics, and generalization capabilities of each method.

5.1 DCNet Performance: A Deep Learning-Based Baseline

The Dual Cascaded Network (DCNet), comprising a VGG16 encoder and a cascaded U-Net-style decoder [11], was initially introduced as a baseline architecture for learning fine spatial features through skip connections and performing end-to-end density map regression. Although DCNet's architecture appeared promising due to its structured use of skip connections and dense feature propagation, it fell short in practical performance across multiple evaluation settings. On the ADI dataset, which consists of densely populated and low-contrast adipose tissue microscopy images, DCNet achieved a best Mean Absolute Error (MAE) of 11.0 under its most favourable configuration—using two skip connections, a batch size of 45, and Huber loss. The results on the MBM and VGG datasets were 5.5 and 9.2, respectively. These errors reflected limitations in accurate cell localization and count estimation, especially in visually cluttered or complex regions.

The relatively poor performance of DCNet across all datasets can be attributed to several architectural constraints. One of the key limitations is its restricted receptive field, which hampers the model's ability to grasp global context—an essential requirement for understanding spatial distributions in microscopy images. Additionally, DCNet showed tendencies to overfit, particularly when applied to real-world datasets

characterized by significant morphological variability. Furthermore, the architecture demonstrated high sensitivity to parameter tuning, necessitating a delicate balance between batch size, choice of loss function, and network depth for optimal performance. While DCNet effectively preserved spatial granularity through its skip connections, it lacked the ability to model the complex and long-range dependencies that are critical for accurate analysis of crowded cellular images.

5.2 Transformer-Based Counting with Restormer

In contrast, the Restormer model—originally developed for high-resolution image restoration—was adapted to perform density-based cell counting. Incorporating a transformer backbone allowed the model to capture global spatial dependencies and contextual relationships that CNNs typically overlook.

Restormer consistently outperformed DCNet across all datasets:

In addition to lower MAEs, Restormer produced smoother and more coherent density maps, even under conditions of high cell overlap or poor contrast. The integration of SALW further enhanced training by automatically adjusting the focus on difficult regions, resulting in better generalization across heterogeneous datasets.

5.3 Comparative Insights

The results clearly indicate that DCNet is not suitable for complex cell counting tasks (shown in Table 2), particularly in real-world biomedical images where spatial context and morphological variability are critical. On the other hand, Restormer delivers comparatively better accuracy (as shown in Table IV), making it a more viable solution for practical deployments.

Table IV: Comparative Result of DCNet vs Restormer

(Best result is in **Bold**, second best result is Underlined and third best result is **Bold Underlined**)

Model	ADI (MAE) N=50	MBM (MAE) N=15	VGG (MAE) N=50
DCNet	11.0±12.639	5.50±4.1	9.2±3.1
Ciampi et al. (2022) [17]	**8.7±0.8**	**5.7±0.9**	2.5±0.1
Count-ception, Paul et al. (2017) [4]	19.4±2.2	8.3±2.3	2.3±0.4
Jiang and Yu (2021) [6]	10.6±0.3	7.5±0.7	2.2±0.2
Rodriguez-Vazquez et al. (2022) [19]	17.3±3.6	**4.2±2.4**	**2.2±0.5**
Ours	**11±0.2**	5.3±1.0	**2.09±0.08**

5.4 Summary

In summary, the DCNet model, despite incorporating cascaded skip connections and VGG16-based features, could not deliver competitive accuracy in dense and noisy biological datasets. While its design preserved spatial features and leveraged transfer learning through VGG16, it lacked the capacity to model complex, long-range dependencies and exhibited sensitivity to hyperparameter tuning, limiting its performance in real-world microscopy scenarios.

In contrast, the Restormer transformer model, equipped with a global attention mechanism and adaptive loss modulation via SALW [37], demonstrated consistently superior results across all tested datasets. Its ability to handle high-resolution images, model both local and global contexts, and dynamically focus on challenging regions contributed to its effectiveness in cell counting tasks.

Future cell counting solutions in biomedical imaging should therefore prioritize architectures that integrate both local precision and global spatial awareness. Transformer-based models like Restormer exemplify this balance and represent a promising direction for developing robust, scalable, and generalizable approaches to automated biological cell analysis.

Chapter 6

Conclusions and Scope for Future Work

6.1 Conclusions

This thesis presented a comprehensive exploration of automated biological cell counting using both deep learning and transformer-based models. Initially, a convolutional architecture—DCNet—was implemented by combining a VGG16 encoder with a cascaded U-Net [11] decoder. While the model incorporated skip connections and robust feature extraction, its performance was limited, particularly in complex datasets like ADI. The DCNet model struggled to generalize across varying cell densities and morphologies, highlighting the constraints of convolutional models in capturing global spatial dependencies.

To overcome these limitations, a transformer-based model, Restormer, was adapted for the cell counting task. Leveraging self-attention mechanisms and a hierarchical encoder-decoder structure, Restormer demonstrated superior performance across diverse microscopy datasets (ADI, MBM, and VGG). Its ability to model long-range dependencies, combined with robust preprocessing (including FIDT maps and super-resolution enhancement), led to smoother and more accurate density predictions. Additionally, the integration of SALW [37] allowed dynamic adjustment of learning focus, enhancing performance on challenging regions of microscopy images.

Quantitative evaluations revealed that the transformer-based approach consistently outperformed the DCNet baseline in terms of Mean Absolute Error (MAE), achieving better accuracy and generalization. The results affirm that attention-based models are more suited to address the spatial complexity and variability inherent in biological cell images.

6.2 Scope for Future Work

In summary, the DCNet model, despite incorporating cascaded skip connections and VGG16-based features, could not deliver competitive accuracy in dense and noisy biological datasets. While its design preserved spatial features and leveraged transfer learning through VGG16, it lacked the capacity to model complex, long-range dependencies and exhibited sensitivity to hyperparameter tuning, limiting its performance in real-world microscopy scenarios.

In contrast, the Restormer transformer model, equipped with a global attention mechanism and adaptive loss modulation via SALW [37], demonstrated consistently superior results across all tested datasets. Its ability to handle high-resolution images, model both local and global contexts, and dynamically focus on challenging regions contributed to its effectiveness in cell counting tasks.

Future cell counting solutions in biomedical imaging should therefore prioritize architectures that integrate both local precision and global spatial awareness. Transformer-based models like Restormer exemplify this balance and represent a promising direction for developing robust, scalable, and generalizable approaches to automated biological cell analysis.

Despite the encouraging results, several avenues remain open for further investigation and enhancement. One key area is the development of lightweight Transformer architectures. Although Transformer models are highly accurate, they are often computationally intensive. Future research could focus on exploring efficient variants or hybrid CNN-Transformer models that reduce inference time and memory usage, making them more suitable for real-time clinical applications.

Another promising direction is the integration of self-supervised and few-shot learning methods. Given the scarcity of annotated biomedical data, these techniques could allow models to pre-train on unlabelled data and then adapt with minimal supervision to different imaging conditions

or novel cell types. This would significantly broaden the applicability of cell counting models in varied laboratory and clinical settings.

Additionally, incorporating uncertainty estimation into cell counting models could enhance both the interpretability and the reliability of predictions, particularly in clinical environments where decision-making heavily depends on the confidence of the automated systems. This would help users better understand and trust the model's outputs.

Moreover, current models are typically focused solely on cell counting. A valuable extension would be to integrate cell counting with other downstream biomedical tasks such as cell classification, tracking, or segmentation. This would help build a comprehensive and unified pipeline for cellular analysis, streamlining workflows in biomedical research and diagnostics.

Finally, enhancing cross-domain generalization remains a critical challenge. Testing models on completely unseen tissue types or staining protocols would provide valuable insights into their robustness and could drive the development of domain-agnostic cell counting systems, further improving their real-world utility and scalability.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.

[2] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[3] Y. Guo, O. Krupa, J. Stein, G. Wu, and A. Krishnamurthy, "SAU-Net: A unified network for cell counting in 2D and 3D microscopy images," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 19, no. 4, pp. 1926–1937, Jul.–Aug. 2022.

[4] J. P. Cohen, G. Boucher, C. A. Glastonbury, H. Z. Lo, and Y. Bengio, "Count-ception: Counting by fully convolutional redundant counting," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 18–26.

[5] H. Zhang, H. Chen, J. Qin, B. Wang, G. Ma, P. Wang, D. Zhong, and J. Liu, "MC-ViT: Multi-path cross-scale vision transformer for thymoma histopathology whole slide image typing," *Frontiers in Oncology*, vol. 12, p. 925903, 2022.

[6] Y. Wang et al., "Learning from synthetic data for crowd counting in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8198–8207.

[7] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2010, pp. 1324–1332.

[8] W. Xie, J. A. Noble, and A. Zisserman, "Microscopy cell counting with fully convolutional regression networks," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 6, no. 3, pp. 283–292, 2018.

[9] A. Paulauskaite-Taraseviciene, K. Sutiene, J. Valotka, V. Raudonis, and T. Iesmantas, "Deep learning-based detection of overlapping cells,"

in *Proc. 2019 3rd Int. Conf. Adv. Artif. Intell.*, pp. 217–220, 2019, doi: 10.1145/3369114.3369120.

[10] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany, Oct. 2015, vol. 9351, pp. 234–241, Springer.

[12] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5728–5739.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[14] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.

[15] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.

[16] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald, et al., "U-Net: Deep learning for cell counting, detection, and morphometry," *Nat. Methods*, vol. 16, no. 1, pp. 67–70, 2019.

[17] L. Ciampi, F. Carrara, V. Totaro, G. Amato, F. Falchi, and C. Gennaro, "Counting cells in microscopy images using a density map regression approach," *Med. Image Anal.*, vol. 80, p. 102500, 2022.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[19] J. Rodriguez-Vazquez, A. Alvarez-Fernandez, M. Molina, et al., “Counting objects in images using blob-based representations,” *Neural Netw.*, vol. 145, pp. 155–163, 2022.

[20] N. Jiang and F. Yu, “A two-path network for cell counting,” *IEEE Access*, vol. 9, pp. 70806–70815, 2021.

[21] V. Ranjan, U. Sharma, T. Nguyen, and M. Hoai, “Learning to count everything,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 1–10, doi: 10.1109/CVPR46437.2021.00340

[22] A. Lehmussola, P. Ruusuvuori, J. Selinummi, H. Huttunen, and O. Yli-Harja, “Computational framework for simulating fluorescence microscope images with cell populations,” *IEEE Trans. Med. Imaging*, vol. 26, no. 7, pp. 1010–1016, 2007.

[23] S. He, K. T. Minn, L. Solnica-Krezel, M. A. Anastasio, and H. Li, “Deeply-supervised density regression for automatic cell counting in microscopy images,” *Med. Image Anal.*, vol. 68, p. 101892, 2021.

[24] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman, “Detecting overlapping instances in microscopy images using extremal region trees,” *Med. Image Anal.*, vol. 27, pp. 3–16, 2016.

[25] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *Adv. Neural Inf. Process. Syst.*, vol. 9199, pp. 2969239–2969250, 2015.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.

[27] Z.-Q. Cheng, Q. Dai, H. Li, J. Song, X. Wu, and A. G. Hauptmann, “Rethinking spatial invariance of convolutional networks for object counting,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19638–19648.

[28] Y. Xue, N. Ray, J. Hugh, and G. Bigras, “Cell counting by regression using convolutional neural network,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 274–290.

[29] L. Fiaschi, U. Köthe, R. Nair, and F. A. Hamprecht, “Learning to count with regression forest and structured labels,” in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, 2012, pp. 2685–2688.

[30] F. Liu and L. Yang, "Multi-objective convolutional learning for cell detection in microscopy images," *Pattern Recognition*, vol. 61, pp. 639–649, Jan. 2017. doi: 10.1016/j.patcog.2016.07.027

[31] E. Walach and L. Wolf, “Learning to count with CNN boosting,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 660–676.

[32] V. A. Sindagi and V. M. Patel, “HA-CCN: Hierarchical attention-based crowd counting network,” *IEEE Trans. Image Process.*, vol. 29, pp. 323–335, 2019.

[33] Z. Zou, X. Qu, P. Zhou, S. Xu, X. Ye, W. Wu, and J. Ye, “Coarse to Fine: Domain Adaptive Crowd Counting via Adversarial Scoring Network,” in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2021, pp. 2185–2194.

[34] D. Liang, W. Xu, Y. Zhu, and Y. Zhou, “Focal inverse distance transform maps for crowd localization,” *IEEE Trans. Multimedia*, vol. 25, pp. 6040–6052, 2022.

[35] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[36] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.

[37] L. D. McClenny and U. M. Braga-Neto, "Self-adaptive physics-informed neural networks," *J. Comput. Phys.*, vol. 474, p. 111722, 2023.

[38] https://github.com/ieee8023/countception