

SINGLE IMAGE CROWD COUNTING USING DEEP LEARNING MODELS

M.Tech Thesis

By

PAWAR SHUBHAM RAJEBHAU



**DISCIPLINE OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE**

JUNE, 2023

SINGLE IMAGE CROWD COUNTING USING DEEP LEARNING MODELS

A THESIS

*Submitted in partial fulfillment of the
requirements for the award of the degree
of*
MASTER OF TECHNOLOGY

by
PAWAR SHUBHAM RAJEBHAU



**DISCIPLINE OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE**
JUNE, 2023

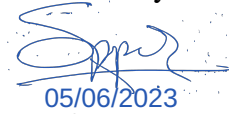


INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Exploring Crowd Counting Field With Deep Learning Models** in the partial fulfillment of the requirements for the award of the degree of **MASTER OF TECHNOLOGY** and submitted in the **DISCIPLINE OF ELECTRICAL ENGINEERING, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from June, 2022 to June, 2023 under the supervision of Prof. Vivek Kanhangad, HOD, Department of Electrical Engineering, IIT Indore.

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.


05/06/2023

Signature of the student with date
(PAWAR SHUBHAM)

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



Signature of Thesis Supervisor

Date: 5 June 2023

PROF. VIVEK KANHANGAD

PAWAR SHUBHAM has successfully given his **M.Tech.** Oral Examination held on **15/05/2023**.



Signature of Thesis Supervisor
(Prof. Vivek Kanhangad)

Date: 5 June 2023

Signature of PSPC Member
(Prof. Amod C. Umarikar)

Date:

Signature of Convener, DPGC
(Dr. Swaminathan R.)

Date:

Signature of PSPC Member
(Dr Neminath Hubballi)

Date:

Acknowledgments

First of all, I would like to express my sincere gratitude to my thesis supervisor, Prof. Vivek Kanhangad, for his constant support and guidance throughout the learning process of this master's thesis. Furthermore, I would also like to express my gratitude to my PSPC committee members, Prof. Amod C. Umarikar and Dr Neminath Hubballi, for their valuable suggestions.

I would also like to thank all the faculty members and the staff at IIT Indore for their cooperation throughout my study and thesis work. I would like to thank the Discipline of Electrical Engineering for providing all the facilities, resources, and research environment required for the completion of this work.

I would like to thank all the members of the Pattern Recognition Image Analysis Lab for their assistance during this work. In particular, I would like to thank Mr Siddharth Singh Savner for his assistance at various levels of this work.

I am especially grateful to my family members for their invaluable support and strong belief in me. I dedicate my master's thesis to my great parents for their countless sacrifices.

Pawar Shubham Rajebhau

Abstract

This MTech. thesis presents multiple approaches for crowd counting in single images using deep learning models. The research investigates multiple techniques to improve the accuracy of crowd counting, with a particular focus on high-density crowd images where existing models often yield suboptimal results.

The frequency domain approach was employed to provide compact and effective supervision for the network. By transforming the density map into the frequency domain using Fast Fourier Transform (FFT), global spatial information was incorporated without relying on external algorithms. This approach enabled the network to leverage frequency-based representations for crowd counting.

A classification-based strategy was pursued to address the challenges of high-density crowd images. The dataset was divided into three density groups, and separate models were trained for each group. A classifier was then employed to select the appropriate model for a given input. Although initial results from this approach were unsatisfactory due to the limitations of existing models, further improvements were achieved by adopting a multi-scale architecture. Multiple regression heads were incorporated in parallel, akin to the MCNN model, and dilated convolution kernels were employed to reduce the additional parameters introduced by the extra layers. This modification resulted in a more robust and stable model capable of better crowd-counting performance.

Additionally, self-supervised training techniques were explored to enhance the model's capabilities further. Autoencoding was employed, wherein the model learned to encode the input into a lower-dimensional representation by utilizing the same input for supervision. This allowed the model to capture salient features and

patterns within the data. Furthermore, rotation prediction was employed, where the model learned to estimate the angle of rotation applied to an input. By leveraging self-supervised pre-training and multitasking, the model acquired a deeper understanding of the underlying structure of the data. Overall, quality questions will be seen while this problem is being solved, motivating one to find the answers.

Contents

1. Introduction	7
1.1. Introduction to the field of crowd analysis	7
1.2. Motivation	9
1.3. Major contribution	11
1.4. Organization of the thesis	11
2. Literature survey	12
2.1. Introduction	12
2.2. Traditional Methods	14
2.3. Density map estimation	15
2.3.1. Network architectures	18
2.3.2. Learning approaches	24
2.3.3. Loss function	24
2.4. Datasets	26
2.5. Evaluation metrics	27
2.6. Conclusion	28
3. Proposed frameworks	29
3.1. Crowd counting in the frequency domain.	29
3.1.1. Network architecture	30
3.2. Classification approach	31
3.2.1. Introduction	31
3.2.2. Network architecture	32
3.3. Multi-scale model approach	34
3.3.1. Network architecture	34
3.4. Self-supervised approach	37
3.4.1. Network architecture	39

4. Results and discussion	42
4.1. Frequency domain analysis	42
4.2. Classification approach	43
4.3. Multiscale analysis	44
4.3.1. Multiscale model 1	44
4.3.2. Multiscale model 2	45
4.4. Self-supervised	46
5. Conclusion and future work	47
5.1. Conclusions	47
5.2. Future works	47

List of Figures

2.1	Fixed σ of the Gaussian kernel [27]	17
2.2	The σ is equal to the distance to the nearest neighbour. [27]	17
2.3	The σ is computed as the average of the distances to the three nearest neighbours, divided by 10 [27]	18
2.4	Single-column model [31]	18
2.5	multi-column model [31]	19
2.6	Mixed model [31]	19
2.7	MSB blob [31]	21
2.8	SANet module [31]	22
2.9	Models with Transfer Learning [31]	22
2.10	Vision transformer [31]	23
3.1	Chf-loss [25]	29
3.2	Fft-loss	30
3.3	FFT-model	30
3.4	Initial pre-training stage	32
3.5	Classifier training stage	32
3.6	Final stage	33
3.7	Base-model	34
3.8	Vgg-16	35
3.9	Multiscale model 1	35
3.10	Multiscale model 2	36
3.11	Dilated convolution [16]	36
3.12	Masked autoencoder [29]	40
3.13	supervised MAE [30]	40
3.14	Multitask model.	41

Chapter 1

Introduction

This chapter introduces the crowd analysis field, emphasizing the significance of crowd counting and analysis in computer vision and image processing. The motivation behind selecting a crowd-counting project for this master's thesis is outlined, highlighting the research advancements, cross-domain applications, and the opportunity to gain insights into deep learning models. The major contribution of the thesis, focusing on improving crowd-counting accuracy through innovative approaches, is presented. The chapter concludes by providing an overview of the organization of the remaining sections of the thesis.

1.1 Introduction to the field of crowd analysis

Crowd counting and crowd analysis are essential research areas in computer vision and image processing. They involve estimating the number of people in a crowd and analysing their behaviour, movement patterns, and interactions. This field has seen significant advancements in recent years due to its broad applications in various domains, such as crowd management, surveillance, urban planning, and event organization.

The development of crowd-counting techniques has gained traction due to several reasons. Firstly, the abundance of surveillance cameras and the availability of large-scale image and video datasets have provided researchers with comprehensive data to train and evaluate crowd-counting algorithms. Additionally, the increasing need for efficient crowd management, security, and safety measures in public spaces has fuelled the demand for accurate and real-time crowd analysis systems. By accurately estimating crowd sizes and

understanding their dynamics, authorities can take proactive measures to ensure public safety, prevent stampedes, manage traffic, and optimize resource allocation.

Several research advancements have been made in the field of crowd-counting. Traditional methods often relied on handcrafted features and simple density estimation techniques. However, with the emergence of deep learning, convolutional neural networks (CNNs) have revolutionized crowd counting by achieving state-of-the-art performance. These deep learning models leverage their ability to learn complex patterns and spatial dependencies from large-scale datasets, enabling more accurate crowd estimation.

Recent developments in crowd counting also involve exploring novel techniques such as attention mechanisms, multi-scale analysis, and generative adversarial networks (GANs). Attention mechanisms allow models to focus on relevant regions in an image, leading to better counting accuracy. The multi-scale analysis involves extracting features at multiple scales to capture different crowd density levels. GANs, on the other hand, can generate synthetic crowd images to augment training datasets and improve the robustness of crowd-counting models. Numerous projects depend on crowd counting and analysis research. Here are a few examples:

- **Crowd management and safety:** Public venues like stadiums, train stations, and airports use crowd-counting systems to monitor and manage large gatherings, prevent overcrowding, and ensure the safety of individuals.
- **Traffic monitoring and optimization:** Traffic authorities employ crowd-counting techniques to estimate the number of pedestrians at busy intersections or public transport hubs. This information aids in optimizing traffic flow and designing efficient transportation systems.

- **Event planning:** Organizers of events like concerts, festivals, and rallies use crowd analysis to estimate attendance, plan logistics, and allocate resources effectively.
- **Security and surveillance:** Crowd counting algorithms are crucial in surveillance systems to detect and track suspicious activities, identify anomalies, and ensure public safety in crowded areas.
- **Urban planning:** Crowd analysis assists city planners in understanding the movement patterns of pedestrians, identifying congested areas, and designing public spaces that accommodate large crowds efficiently.

In conclusion, crowd-counting and analysis have witnessed significant advancements due to the availability of data, the need for crowd management, and the emergence of deep learning techniques.

1.2 Motivation

The remarkable advancements and continuous research in this field have played a pivotal role in selecting a crowd-counting project for this master's thesis. The abundance of high-quality work and the active exploration of cutting-edge models and innovative ideas by numerous researchers have sparked significant interest. Additionally, the potential application of crowd-counting techniques in diverse domains, including cell counting in tissues, agricultural crop estimation, and object detection problems, has strongly motivated this project.

- **Research advancements:** The field of crowd counting has made substantial progress, with researchers continuously pushing the boundaries of what is possible. By working on this project, the aim is to tap into the wealth of existing work

and benefit from the latest advancements in the field. The abundance of research literature and the constant flow of new models and techniques provide strong motivation to contribute to this area's growing body of knowledge.

- **Cross-domain applications:** Crowd counting techniques have demonstrated their versatility and potential for adaptation in various domains. The skills acquired during this project can be seamlessly transferred to other areas, such as cell counting in tissue samples or estimating agricultural crop yields. This multidisciplinary nature of crowd-counting research broadens the scope of its applications. It encourages exploration beyond the immediate domain, providing an opportunity to contribute to diverse scientific and practical challenges.
- **Deep learning insights:** Undertaking a crowd-counting project offers us a valuable opportunity to gain in-depth knowledge of deep-learning models and their inner workings. Exploring these models' different components and working blocks provides a solid foundation for understanding the principles underlying their success. From convolutional neural networks (CNNs) to attention mechanisms, this project will allow us to grasp these advanced models' intricacies, facilitating the application of similar techniques to various other computer vision problems.
- **Loss functions and model training:** Crowd counting projects explore different loss functions and model training methodologies explicitly tailored for counting tasks. Understanding the nuances of these techniques is crucial for achieving accurate and robust crowd estimation. Through this project, the aim is to gain expertise in selecting appropriate loss functions, exploring novel training strategies, and optimising model performance. Such

knowledge is transferable to other deep learning applications, providing a comprehensive skill set to tackle various computer vision challenges.

1.3 Major contribution

This MTech thesis focuses on multiple approaches to improve crowd counting accuracy in high-density crowd images. A frequency domain approach was introduced by utilizing Fast Fourier Transform (FFT) to incorporate global spatial information. A classification-based strategy with a multi-scale architecture and self-supervised training techniques enhanced the model's performance.

1.4 Organization of the thesis

The rest of this thesis is organized as follows:

Chapter 2 presents the literature survey consists of different state-of-the-art techniques for crowd counting.

Chapter 3 presents the proposed methods.

Chapter 4 presents the results and discussion of the proposed methods in depth.

Chapter 5 includes the conclusion and scope of the future work of this thesis.

Chapter 2

Literature survey

2.1 Introduction

Crowd counting is a vital task that plays a crucial role in numerous real-world applications. It involves estimating the number of individuals in a crowd or gathering by analysing visual data such as images or videos. This field has garnered significant attention due to its relevance in public safety, event management, transportation planning, and urban infrastructure design.

Accurate crowd counting provides valuable insights for situational awareness, resource allocation, and crowd management. It aids in ensuring public safety during large-scale events, optimizes transportation systems, and enhances public spaces' efficiency. Moreover, it enables a better understanding of crowd dynamics, crowd behaviour analysis, and the impact of various factors on crowd movements.

However, crowd counting poses several challenges due to the inherent complexities associated with crowded scenes. Factors such as occlusions, varying crowd densities, perspective distortion, and scale variations make accurate counting a non-trivial task. Traditional manual counting methods are labour-intensive, time-consuming, and prone to errors, necessitating the development of automated and reliable crowd-counting techniques.

In recent years, computer vision and deep learning advancements have revolutionized the crowd counting field. Deep learning models, particularly convolutional neural networks (CNNs), have demonstrated superior performance in accurately estimating crowd counts from visual data. These models learn to extract intricate

patterns and features from images or videos, enabling them to handle complex crowd scenes more effectively.

This literature survey section will provide a comprehensive overview of crowd-counting methods, encompassing traditional approaches and the latest state-of-the-art methods. Additionally, relevant aspects such as crowd-counting datasets, evaluation metrics, and loss functions will be explored. This comprehensive examination will enable a deeper understanding of crowd-counting techniques' evolution and current landscape.

Traditionally, two main approaches have been employed for crowd counting: detection-based and regression-based. In crowded scenes with severe occlusions, the accuracy of the detection-based approach, which utilizes computer vision techniques to identify individual objects, heads, or body parts and determine their total count in an image, tends to decline. This method also incurs the highest labelling cost, requiring complete identification and outlining of each object. Conversely, the regression-based approach takes a different approach by directly estimating the count by establishing a relationship with the characteristics of the image. This method is more accurate than the detection-based approach in crowded scenes. However, it has limitations regarding spatial information and interpretability, which restrict its applicability in localization studies. Regression-based methods do not require annotating individual objects, resulting in a lower annotation cost as only the total object count needs to be provided. Recently, a contemporary deep learning technique known as density map estimation has emerged as a promising solution for crowd counting. This method has proven to achieve impressive accuracy in crowded scenes while retaining spatial information regarding the distribution of individuals. Unlike the detection-based approach, density map estimation only necessitates indicating the locations of people's heads, thereby striking a balance in labelling cost between the detection-based and regression-based approaches. By understanding these crowd counting approaches' characteristics and

annotation requirements, researchers and practitioners can make informed decisions when selecting the most suitable method for their specific application. This comprehensive overview sets the stage for further exploration and advancements in crowd counting.

2.2 Traditional methods

The overview shows that traditional methods primarily rely on a total count approach. These approaches utilize image processing techniques to identify hand-crafted features. e.g., Z. Lin et al. [1] introduced an approach for detecting and segmenting humans simultaneously. They combined local part-based and global shape-template-based methods to achieve this. Then they used support vector machines (SVMs) [2] classifiers to separate human/nonhuman patterns. Lin et al. [3] introduced a method for recognizing head-like contours and estimating crowd size. Their approach involved extracting the featured area of the head-like contour using the Haar wavelet transform. Subsequently, a support vector machine was employed to classify these featured areas as either the contour of a head or something else. Nevertheless, when it comes to images with dense crowds, the accuracy of these methods noticeably diminishes due to various challenges like occlusions, low resolution, perspective effects, and more. Regression-based methods have gained popularity over detection-based approaches because they can estimate the total count from images or image patches. Instead of focusing on specific body parts or shapes, these methods utilize global features like texture, foreground, and gradients. This enables them to provide estimates of the total count at the image or patch level. Ke Chen et al. [4] used low-level imagery features from each local cell region, including local foreground, edges and texture features. They proposed using local rather than global features, which lack local context information and information sharing among these local regions.

Dalal et al. [5] presented a method that utilizes grids of Histograms of oriented gradient (HOG) descriptors. This approach incorporates several key elements, including fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization. As demonstrated in their work, these components contribute to the method's effectiveness. While regression-based methods effectively address challenges like occlusions, low resolution, and perspective effects, they often exhibit poor performance when dealing with high-density crowd images.

2.3 Density map estimation

Density estimation refers to using a CNN model to predict the density map of a crowd scene, going beyond simply predicting the head count. The density map provides the total headcount and additional location information about the crowd scene. By generating a density map, the model captures a finer level of detail regarding the distribution and density of individuals within the scene. This richer information can be valuable for various applications in crowd analysis and understanding crowd behaviour. As this is the modern deep learning-based method, much research has been done and is still ongoing. To cover all aspects of the literature survey, it was divided into subparts focusing on network architectures, loss functions, and training methods. Before diving deep into the world of these classical models, let us first understand what a density map is. How are they created? Moreover, what are their pros and cons? To generate density maps, dot-annotated ground truths are utilized. Each object within an image is annotated with a single dot typically placed on the person's head. However, these sparse dot maps make it challenging to train neural networks effectively. Therefore, a conversion transforms these dot maps into density maps. For instance, let us consider x_i as a pixel representing

the head position, which can be represented by the delta function $\delta(x-x_i)$. The delta function is convolved to generate the density map with a Gaussian kernel $G\sigma$, resulting in a smooth distribution of values indicating the head density across the image. This convolution operation helps generate a more informative and continuous density map from the sparse dot annotations.

$$Y = \sum_{i=1}^K (\delta(x - x_i) \cdot G\sigma) \quad (2.1)$$

K represents the total number of annotated points corresponding to the total headcount. The integral of the density map provides an estimation of the total headcount by summing up the values across the entire map. The σ values of the Gaussian kernels used in density map generation are typically fixed [6]. However, this fixed approach does not account for scale variations in different images, limiting its effectiveness. To address this, adaptive methods were proposed. These adaptive approaches calculate the value of based on the average distance to the K -nearest neighbouring head annotations. This allows for a lower degree of Gaussian blur for dense crowd regions and a higher degree for regions with sparse density. Another variation involves taking the average of the three nearest neighbours to determine the value of σ . This adaptive approach helps capture the variations in crowd density across different image regions. The images below illustrate the concept of using adaptive values and the resulting density maps.

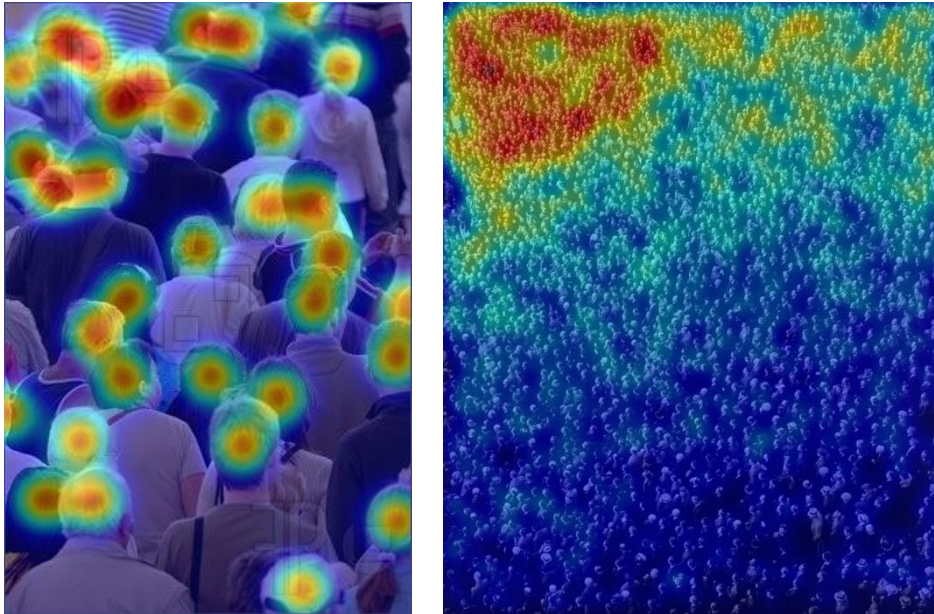


Figure 2.1: Fixed σ of the Gaussian kernel [27]

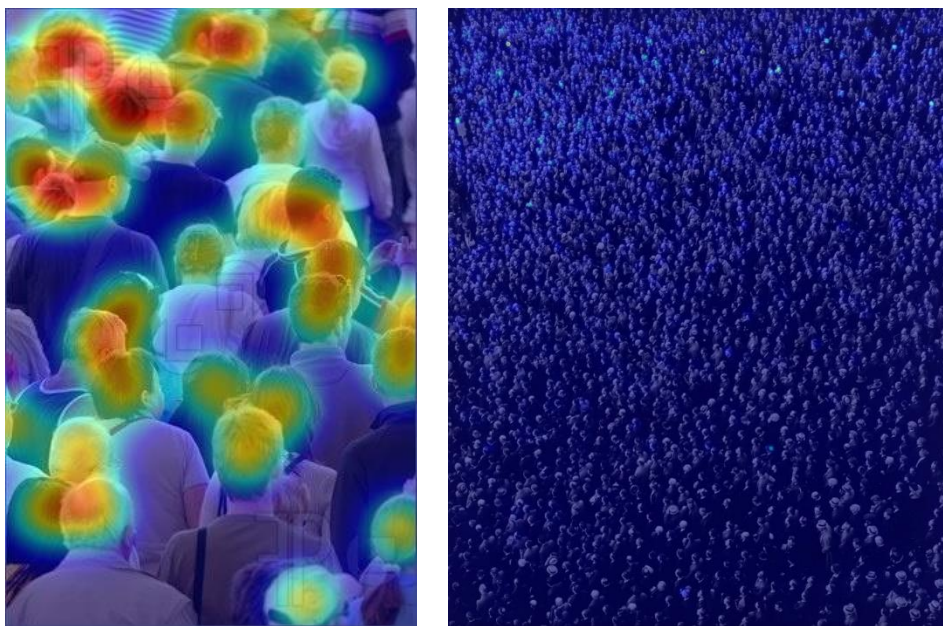


Figure 2.2: The σ is equal to the distance to the nearest neighbour. [27]



Figure 2.3: The σ is computed as the average of the distances to the three nearest neighbours, divided by 10. [27]

2.3.1 Network architectures

Different deep neural network models used for crowd counting were analysed and classified into three categories: single-column, multi-column, and hybrid designs. Although some similarities exist with other CNN models, understanding the unique architecture of these models is crucial for accurate crowd counting.

Single-column models: These are characterized by a sequence of cascaded convolution layers arranged in a single column. The general structure is as follows:

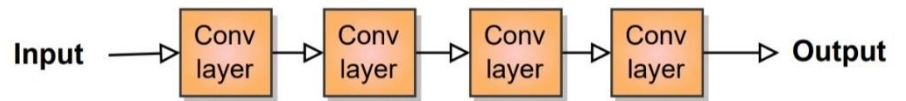


Figure 2.4: Single-column model [31]

Zhang et al. introduced the Crowd CNN [7] model, which consists of three convolutions (Conv) layers with distinct kernel sizes (7×7 , 7×7 , and 5×5). These Conv layers are then followed by three

fully connected (FC) layers. The model takes 72 patches cropped from an image as input and produces a density map of size 18×18 as the output. The model's performance is assessed using the UCF CC 50 [8] dataset with an MAE of 467.0 and MSE of 498.5. Compact single-column models exhibit reduced accuracy when applied to dense images. Additionally, they encounter challenges in handling scale variations within images.

Multi-column models:

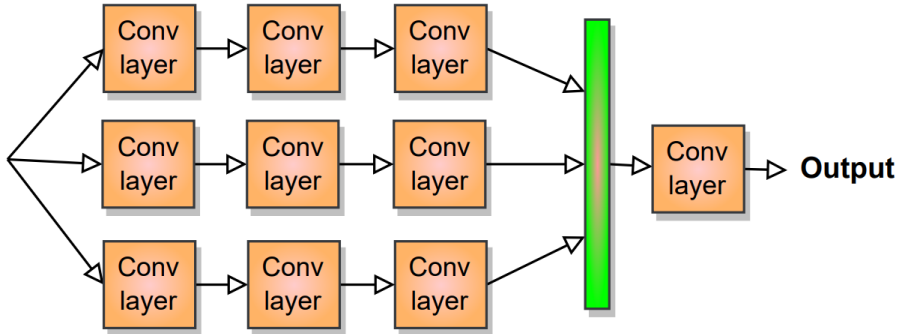


Figure 2.5: Multi-column model [31]

Multi-column approaches are preferred due to their ability to learn features that can detect multi-scale variations and are independent of perspective. Zhang et al. [9] proposed the first multi-column CNN (MCNN), an architecture consisting of three CNN layers. Each column within the MCNN has different receptive fields and generates a density map that matches the shape of the ground truth map. The outputs from each column are combined by concatenation to get the density map finally. Notably, MCNN can process images of any size, making it versatile in its application. The model's performance is assessed using the UCF CC 50 and Shanghaitech A and B dataset. MSE and MAE for part-A are 173.2 and 110.2, respectively. For part B, MSE and MAE are 41.3 and 26.4, respectively. On UCF CC 50, the MAE and MSE are 377.6 and 509.1, respectively. Sindagi et al. introduced a simpler two-column network called cascaded multi-task learning (CMTL) [10]. This

architecture enables simultaneous prediction of images' density map and counts density. In the two-column setup, the first is designed to estimate a high-level prior, such as the total headcount, while the second is dedicated to estimating the density map. By jointly training these two columns, the CMTL model aims to improve the accuracy of crowd-counting tasks. Liu et al. propose a Decide-Net [11] designed for simultaneous detection and density estimation. Decide-Net comprises three columns, each performing a distinct task. In the proposed architecture, the first column, Reg-Net, employs a 5-layer CNN to predict the density map in cases where the target density map is unavailable. The second column, Det-Net, utilizes the Faster R-CNN network [12] to predict bounding boxes and generate a density map based on the detections. Lastly, the third column, Quality-Net, takes the density maps generated by Reg-Net and Det-Net and predicts the final density map. This three-column setup enables a comprehensive approach to density map estimation by leveraging different techniques and combining their outputs. This multi-column architecture enables Decide-Net to estimate the count and density maps effectively. Multi-column models have the potential to handle scale variations. Although multi-column models can effectively handle object scales, their adaptability is limited by the number of columns they contain. While multi-scale models can mitigate scale variations, they often require higher computational resources, increasing computational costs. Therefore, there is a trade-off between the model's ability to handle scale variations and the computational efficiency of the approach. This is because multiple columns need to be trained in parallel, resulting in increased computational requirements.

Mixed models: This network architecture combines the advantages of single and multi-column models. It incorporates specialized multi-path or multi-column modules into a single-column network; this architecture effectively addresses the limitations of both approaches. It enables improved detection of multi-scale features

while minimizing the increase in model size. This allows for improved detection of multi-scale features without significantly increasing the model size.

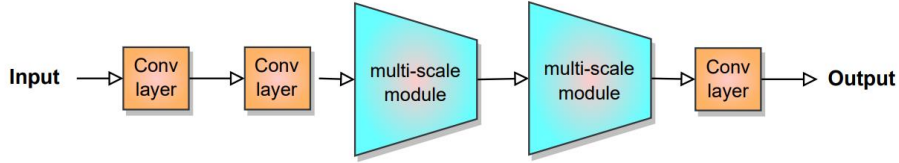


Figure 2.6: Mixed model [31]

The first crowd-counting model to adopt this architecture was the multi-scale CNN [13], proposed by Zeng et al. MSCNN is a single-column CNN network that incorporates three multi-scale blobs. These MSBs draw inspiration from the Inception model [14] and consist of a naive Inception module. These MSBs allow MSCNN to capture multi-scale information effectively and improve its performance in crowd-counting tasks. The MSB incorporates Conv filters of varying sizes, including 3×3 , 5×5 , 7×7 , and 9×9 , to capture information at different scales.

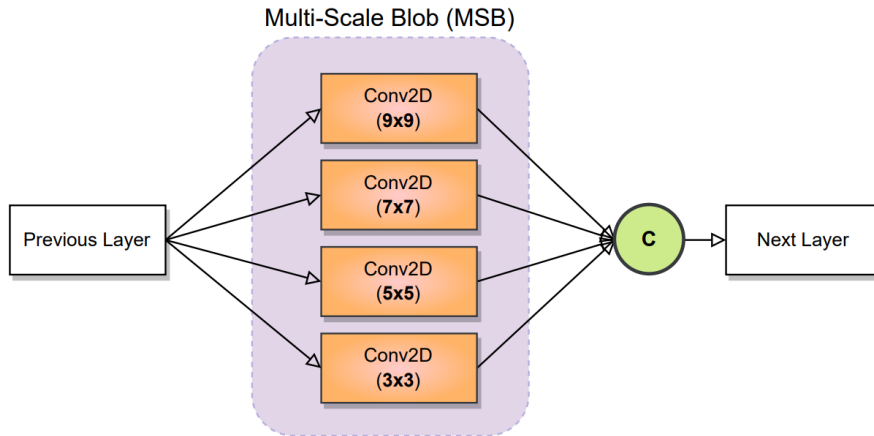


Figure 2.7: MSB blob

Cao et al. proposed a little different multi-scale module called SANet [15]

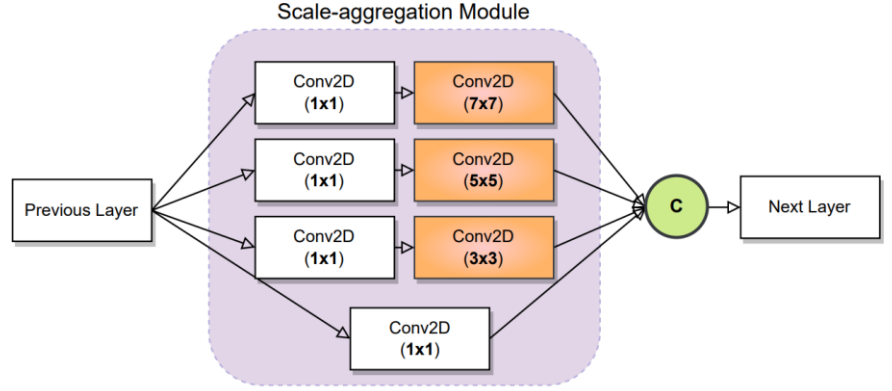


Figure 2.8: SANet module

Models with transfer learning: Compact crowd-counting models often experience a decline in accuracy when applied to high-density scenes. Transfer learning has emerged as a promising approach to address this issue and enhance the performance of crowd models in densely populated areas. By leveraging pre-trained models or knowledge from related tasks, transfer learning enables the model to benefit from prior learning experiences and generalize well to challenging scenarios with high crowd density.

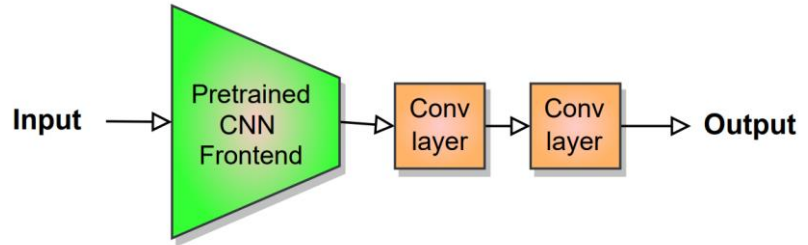


Figure 2.9: Models with Transfer Learning

Li et al. proposed a CSR-Net [16] which utilizes the initial ten layers of the VGG-16 model, which is pre-trained on the ImageNet dataset [17], as the front-end convolutional neural network (CNN). CSR-Net employs dilated convolutions in its back-end CNN, unlike traditional pooling operations, resulting in an all-convolutional network with larger receptive fields. This architectural choice facilitates easier training. Additionally, transfer learning is commonly employed in various architecture types, including single-

column, multi-column, and encoder-decoder models, to leverage pre-existing knowledge and improve performance.

Transformer-based models: Transformer models, known for their self-attention mechanisms, have achieved impressive natural language and speech processing results. This success has sparked interest in the computer vision field, leading to their application in various vision tasks.

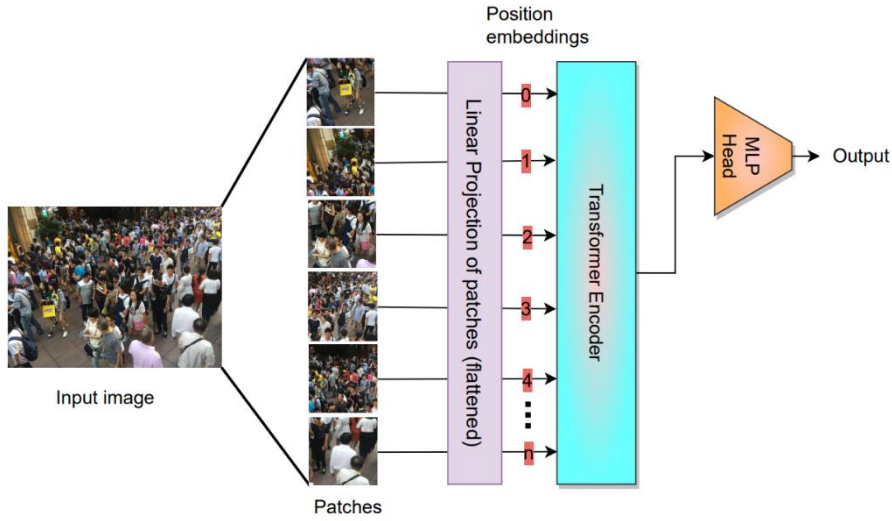


Figure 2.10: Vision transformer

Liang et al. proposed Trans-Crowd [18], a pure transformer architecture that is employed for crowd-counting tasks. Initially sized at (1152x768), each image is divided into six fixed-size patches of (384x384). These patches are then flattened into sequences, and then they are passed through the encoder of the Vision transformer (ViT) model [19], which has been pre-trained on the ImageNet dataset [17]. Then the output is given to a regression head which predicts the total count. Tian et al. proposed CCTrans [20], a transformer-based model designed for density map prediction. CCTrans shares the same input pipeline as the previous module but employs a distinct model architecture. It utilizes Twins [21] as the backbone feature encoder, transforming the 1D output into 2D feature maps. These feature maps are up sampled to 1/8 of

the input patch size and passed through regression heads to generate density maps.

2.3.2 Learning approaches

Current SOTs in crowd density estimation predominantly rely on point-level supervision, which involves fully supervised learning. While this approach yields highly reliable and accurate results, it demands costly annotations. All the models observed until this point have been trained in a fully supervised manner. To make crowd-counting more cost-effective, some studies have investigated the use of weakly supervised learning methods. These methods leverage total headcount annotations, which are more affordable and easier to obtain. In recent advancements, transformer models have been introduced to enable weakly supervised learning. These models automatically capture semantic information from crowd images using their self-attention mechanism. An example is the Trans-Crowd model [18], which is built upon the ViT architecture [19]. Trans-Crowd exclusively relies on count-level supervision to predict the crowd count, showcasing the potential of transformer models in weakly supervised crowd-counting tasks.

2.3.3 Loss function

Loss functions play a crucial role in crowd counting, significantly impacting the performance of image-based crowd-counting systems. Pixel-wise L2 loss focuses on the individual pixel values and does not consider the spatial relationships or contextual information between pixels. It treats each pixel independently and assigns equal importance to all pixels in the density maps. However, pixelwise L2 loss may only partially exploit the position information or the spatial structure present in the ground truth data. It does not consider the relationships between neighbouring pixels

or the global context of the crowd. So, to address these limitations, alternate loss functions have been proposed, such as Bayesian loss [22] introduced by Ma et al., which takes a different approach. Instead of generating a density map, it uses the ground truth dot map, which represents the positions of individuals in the crowd, to calculate class conditional distributions (CCD) for each position. The CCD provides information about the likelihood of a person being present at a specific location within the image. They used a simple pre-trained VGG19 backbone and three convolutional layered regression head network architectures. Wang et al. proposed a DM-count model [23] which uses OT loss. The traditional pixel-wise L2 loss only considers the corresponding pixel in the ground truth. In contrast, the Optimal Transport (OT) loss considers the influence of nearby pixels based on their distances. The OT loss addresses the global optimization problem by considering the transport of all pixels, utilizing the position information of the ground truth for more precise supervision. However, using the OT loss requires external algorithms, such as the Sinkhorn algorithm [24], to extract spatial information from the ground truth, which can be computationally inefficient. Shu et al. proposed Chf-Loss [25] as Extracting global spatial information without the aid of external algorithms, such as the Sinkhorn algorithm [24] for the OT loss, can be challenging. However, by transforming the finite measure into the frequency domain, the spatial information becomes hierarchically organized within a compact range around the origin, and they use characteristic functions to do that. Literature survey table for sot results-take it from any survey paper

2.4 Datasets

NWPU-crowd: This NWPU dataset [26] has 5109 images which contain 2,133,375 annotated instances, and labels are both point and box. The advantages of this dataset are presence of negative samples as they refer to instances or regions within an image that does not contain the object of interest. Including negative samples in the dataset helps the deep learning model to discriminate between the object of interest and other irrelevant parts of the image. Fair evaluation could be achieved by ensuring that the dataset covers a wide range of scenarios and that the annotations are accurate and consistent, as it is essential while objectively comparing the performance of different models and techniques. High-resolution images as it allows the model to capture more intricate features, leading to more accurate and nuanced predictions.

JHU crowd: This dataset [27] has 4372 images which contain 1.51 million annotations with an average resolution of 1430x910. The dataset contains numerous images that incorporate weather-related degradations and variations in illumination, presenting a highly challenging dataset. It offers labels at the head level, including dots, approximate bounding boxes, and blur-level, as well as image-level labels indicating scene type and weather conditions.

UCF-QNRF: This dataset [28] contains 1535 images with 1,251,642 annotated instances with an average resolution of 2013x2902. The advantage of this dataset is that it contains buildings, vegetation, sky and roads usually present in realistic captured scenarios. This makes the dataset more realistic and challenging. Also, it is very diverse in terms of image resolutions, perspectives and crowd density. Also, it contains images from all parts of the world.

UCF-CC-50: This dataset [8] has 50 images with 63,974 annotated instances with an average resolution of 2101x2888. It contains images of a highly dense crowd, ranging between 94 to 4543. The

small size and high density make it more challenging.

Shanghaitech part A and part B: Part A of this dataset [9] contains 482 images (300 for training, 182 for testing) and high-density crowds collected from the Internet. Part B contains 716 images (400 for training, 316 for testing) and is captured from busy streets in urban areas of Shanghai. The average resolution of images in Shanghaitech part A is 589x868 pixels. The scenes in part B are less crowded than those in part A.

2.5 Evaluation metrics

Mean absolute error (MAE): MAE measures the average absolute difference between the predicted and actual counts of individuals in a crowd. It provides a straightforward understanding of the average magnitude of the prediction errors. A lower MAE indicates better accuracy in estimating the crowd count. MAE is useful for interpreting the model's overall performance regarding absolute count errors.

$$MAE = \frac{1}{N} \sum_{i=1}^N |C_{li}^{pred} - C_{li}^{gt}| \quad (2.2)$$

Mean squared error (MSE): MSE calculates the average squared difference between the predicted and actual counts of individuals. MSE penalizes larger errors more severely than MAE due to the squaring operation. It provides a measure of the average squared magnitude of the prediction errors. MSE is particularly useful for assessing the model's performance regarding the variance or spread of errors. Lower MSE indicates better precision and less variability in the prediction errors.

$$MSE = \frac{1}{N} \sum_{i=1}^N |C_{li}^{pred} - C_{li}^{gt}|^2 \quad (2.3)$$

$N = \text{Number of test images}$

$C_{li}^{pred} = \text{prediction results}, C_{li}^{gt} = \text{ground truths}$

2.6 Conclusion

Based on the comprehensive analysis, it is evident that researchers have primarily focused on addressing two major challenges in developing crowd-counting models: scale variation and the scarcity of high-quality labelled data. Scale variation refers to the ability of the models to estimate counts in images with varying crowd densities accurately. Additionally, the lack of efficient ground truth labelling methods has posed difficulties in providing reliable supervision for these models. These challenges have driven the exploration of innovative techniques to tackle scale variation and improve the quality and availability of labelled data for more accurate crowd counting.

Chapter 3

Proposed framework

3.1 Crowd counting in the frequency domain

In the Chf-loss [25] paper, the authors proposed a novel approach by utilizing the characteristic function to transform the density map into the frequency domain. This innovative technique offers computational benefits compared to previous state-of-the-art loss functions like OT loss and p2p loss, which require complex algorithms to calculate the distance matrix for determining global relationships in the context of global optimization problems. Furthermore, the authors argue that a compact representation of the ground truth is more conducive to the learning process of deep learning models than a sparse representation.

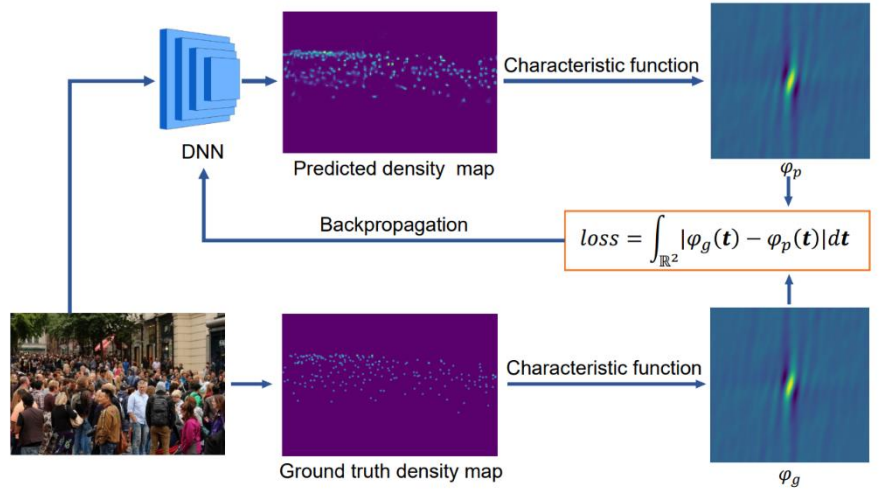


Figure 3.1: Chf-loss [25]

The Fast Fourier Transform (FFT) was employed to compute the frequency response of the predicted and ground truth density maps, inspired by their characteristic function. Subsequently, the L1 loss was utilized to calculate pixel-wise differences between the two,

allowing for accurate quantification of the loss in the predicted density map.

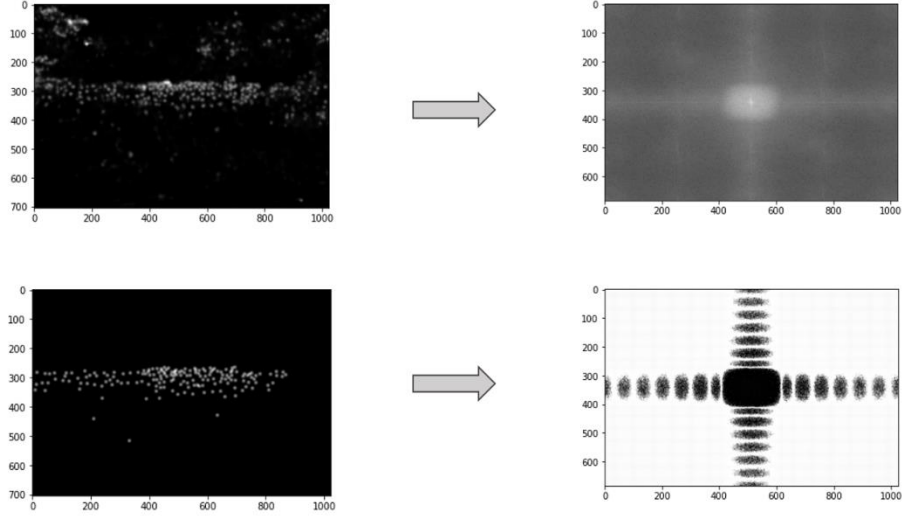


Figure 3.2: Fft-loss

3.1.1 Network architecture

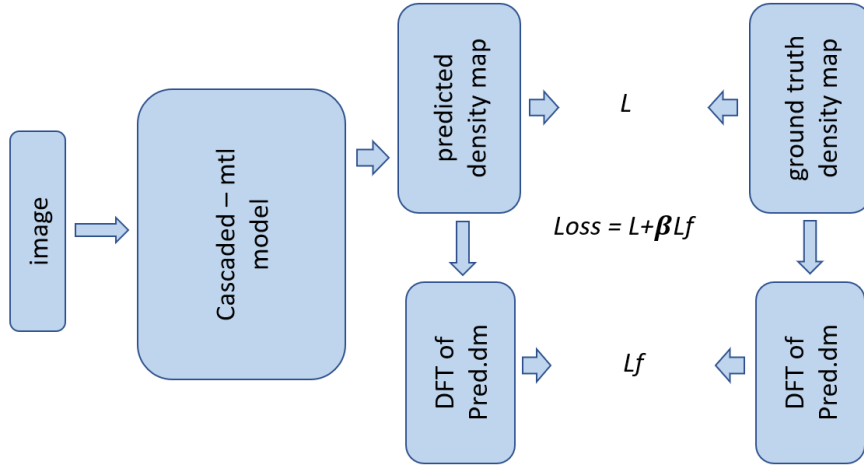


Figure 3.3: FFT-model

To ensure stable results during the experiments, the cascaded-mtl (CMTL) model was used as the base model. Its consistent performance and reliability made it an ideal choice for the research. Further details are given in results and discussions.

3.2 Classification approach

The motivation behind adopting this approach lies in overcoming the scaling problem in crowd counting. The scaling problem refers to the substantial variations in crowd sizes observed within a single scene and across different scenes. The importance of effectively addressing the scaling challenges is acknowledged and understood. The focus is on developing a method or technique to handle these challenges and provide a solution that can accurately estimate crowd counts, irrespective of the encountered scale variations.

3.2.1 Introduction

The scaling problem in crowd counting was addressed by opting for an alternative approach rather than relying on multi-channel bulky models. Instead, a classifier was explored to determine the most suitable model for a given image based on its density, specifically the number of people present. The densities were categorized into three groups: low density (less than 50 people), medium density (between 50 and 500 people), and high density (above 500 people). Initially, three different models were separately pre-trained using this dataset. Subsequently, the entire dataset was divided by passing each image through the corresponding model to determine the best-performing model for each image. So, the data is segregated model-wise. A classifier was then trained on this segregated data to predict a given image's appropriate model (model 1, model 2, or model 3).

3.2.2 Network architecture

The base model utilized in this study is identical to the model employed in the dm-count study, which consists of a pre-trained VGG16 network and a simple three-layer regression head. For

classification tasks, the Resnet-50 network was employed. So, the overall network after combining all these blocks is as follows:

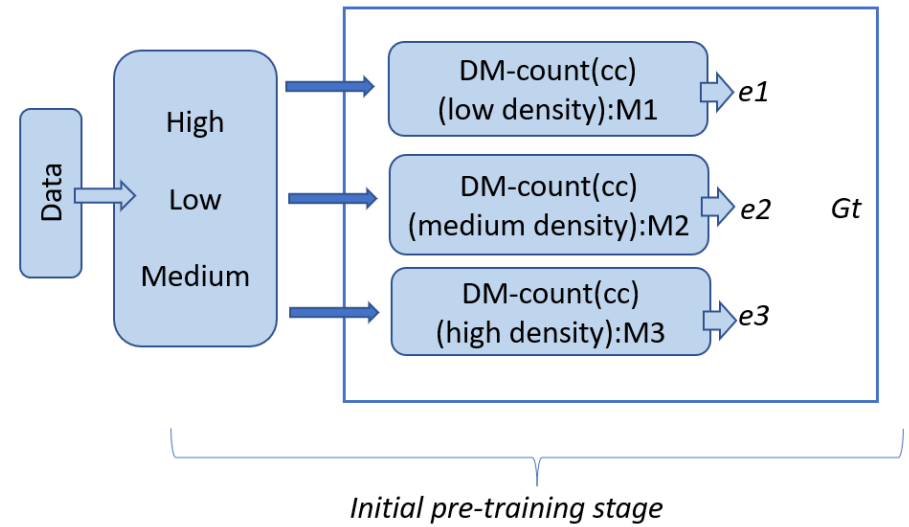


Figure 3.4: Initial pre-training stage

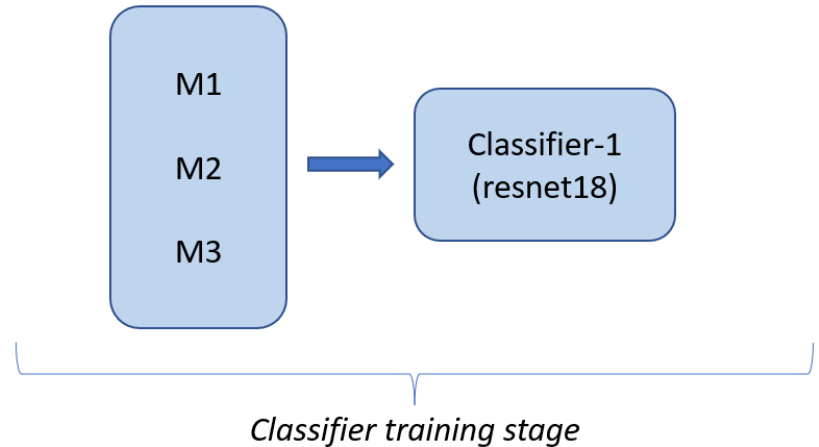


Figure 3.5: Classifier training stage

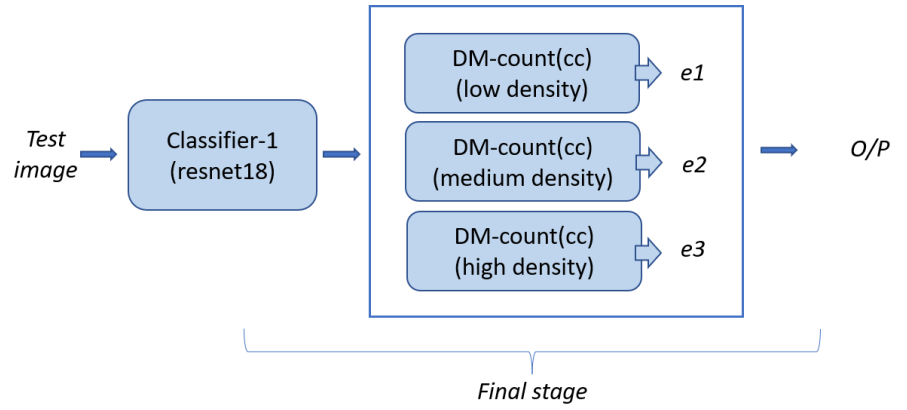


Figure 3.6: Final stage

In the first stage, separate training was conducted for each model according to the density class. In the second stage, a classifier was trained using the labels obtained with the assistance of three pre-trained models named M1, M2, and M3 from stage one. Subsequently, a classifier was trained based on these labels, resulting in three trained models and a classifier. Data is passed through this pipeline during testing, initially going through the classifier and then to the selected model to obtain the final estimation. The classifier acts like a switch here.

3.3 Multiscale model approach

An additional column with a different kernel size (5x5) was introduced to the existing setup to achieve distinct receptive fields. While keeping all other factors unchanged, improved results were observed compared to the base model for both the JHU and SHA datasets. The primary objective was to enhance the performance specifically for high-density data in the JHU dataset. As this dataset provided sufficient data for training and validation, concerns about overfitting were less prominent. However, the Shanghaitech dataset presented a challenge due to its limited availability of training data. To address this, experiments were conducted by freezing specific layers of the VGG16 backbone network to optimize the results in this context.

3.3.1 Network architecture

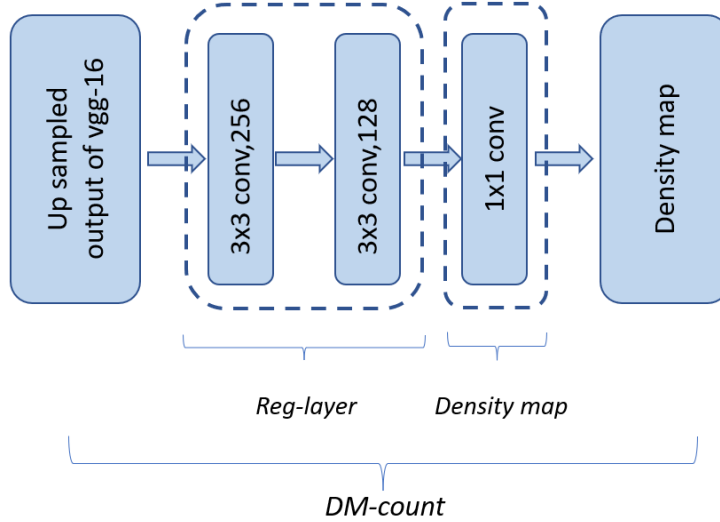


Figure 3.7: Base-model

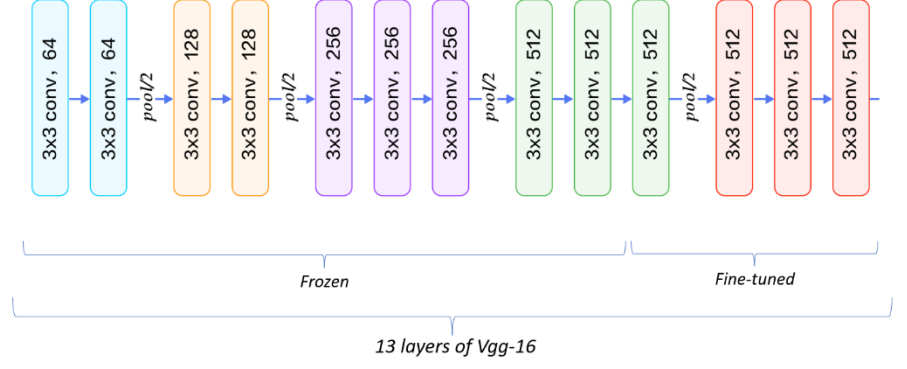


Figure 3.8: Vgg-16

As shown below, an additional layer with a kernel size 5x5 was attached to the network architecture. This layer was designed to capture extra-scale features. The output of this layer was then concatenated and fed into the density map estimation layer, which had a kernel size of 1x1.

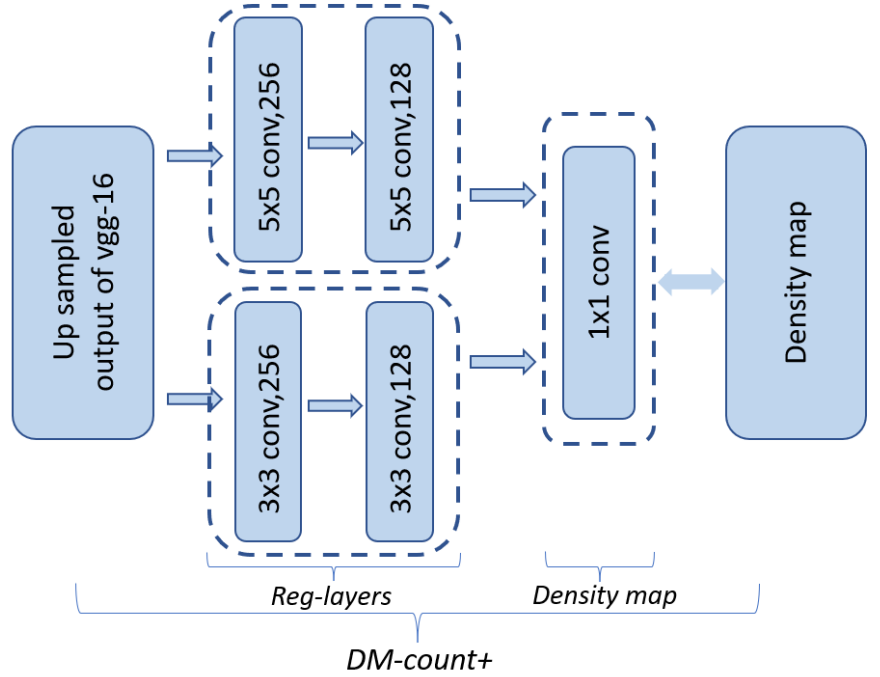


Figure 3.9: multiscale model 1

By incorporating additional layers into the network, its size and the number of parameters grew larger. However, a technique called dilation convolution was employed to mitigate this increase. This approach involved using sparse kernels, effectively expanding the receptive field without adding more parameters or computational

load. To illustrate, by employing a kernel size 3×3 (requiring nine parameters) and a dilation rate of 2, the receptive field achieved was equivalent to that of a standard 5×5 convolutional kernel.

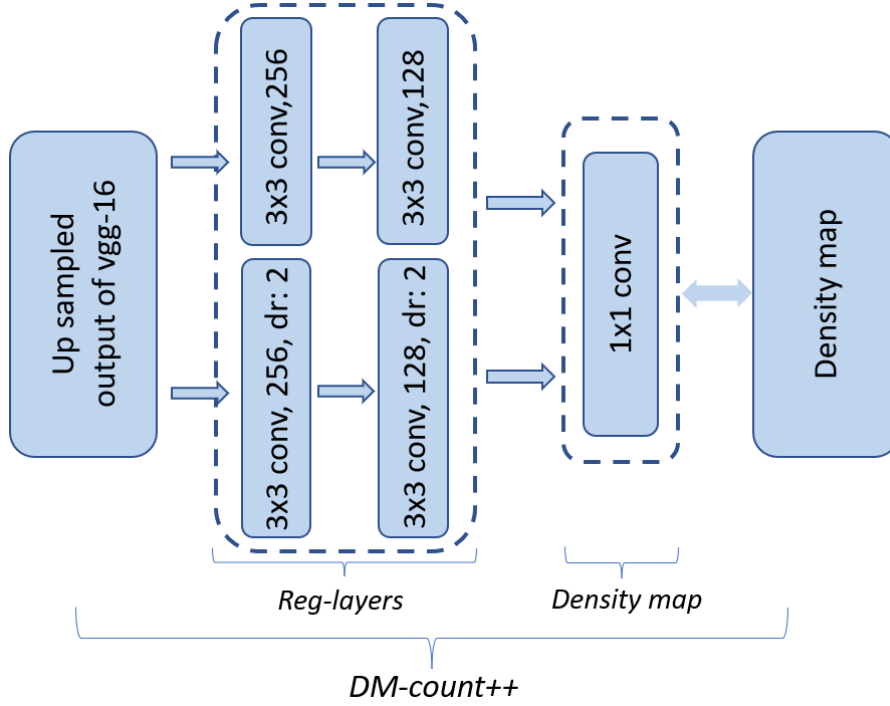


Figure 3.10: multiscale model 2

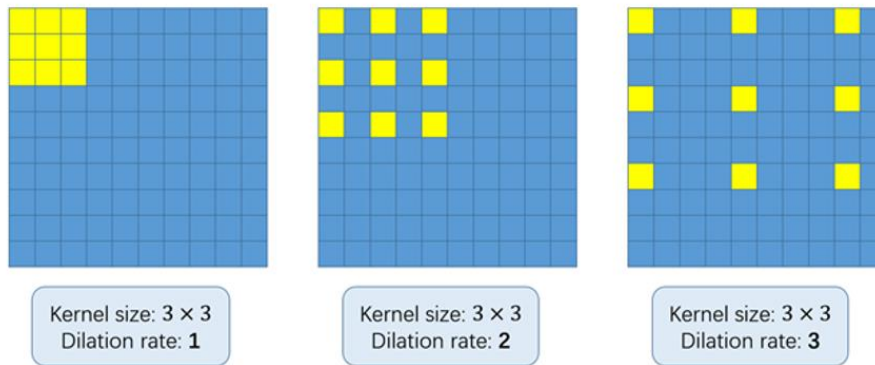


Figure 3.11: Dilated convolution [16]

3.4 Self supervised approach

In transfer learning, a model is typically first trained on a large dataset and a complex task, such as image classification on a vast image dataset. This initial training helps the model learn general features and representations useful for various tasks. The learned knowledge is transferred or adapted to a new, smaller dataset or a different but related task. Transfer learning can be explained in fine-tuning, pretraining, and linear probing.

- **Pretraining:** Pretraining refers to training a model on a large dataset and a specific task that is typically unrelated to the target task. This is commonly done using unsupervised learning or self-supervised learning techniques. For example, in computer vision, a model can be pretrained on a large dataset of images by solving a proxy task such as predicting the relative positions of image patches. The goal of pretraining is to learn general representations and features that capture helpful information from the data.
- **Fine-tuning:** Fine-tuning takes a pre-trained model and further trains it on a smaller, task specific dataset related to the target task. The pre-trained model's weights are used as initial weights for the new model, and the model is then trained on the target task's data. The model's parameters are adjusted or updated during fine-tuning to adapt to the new task. This process allows the model to refine its learned representations and adapt them to the target task.
- **Linear probing:** Linear probing, also known as linear evaluation, is a technique used to evaluate the quality of learned representations in a transfer learning scenario. After the pre-trained model has been fine-tuned on the target task, a linear classifier or an external neural network is added to the pre-trained model's frozen layers. The new classifier is then trained using the target task's labelled data. The

purpose of linear probing is to assess the transferability and quality of the learned representations by measuring the performance of the linear classifier on the target task. It helps evaluate how well the pre-trained model has captured relevant, valuable information for the specific target task.

Transfer learning offers several benefits, including:

- **Reduced training time:** By utilizing pre-existing knowledge, transfer learning can significantly reduce the time and computational resources required for training models on new tasks.
- **Improved performance:** Transfer learning allows models to start with a strong foundation of knowledge, which often leads to better performance on the target task, especially when the new dataset is limited.
- **Effective generalization:** Models trained with transfer learning can better generalize patterns and features learned from the source task to the target task, even when the two tasks are not identical. Pre-training a model on a large dataset and a complex task allows it to learn generic features and representations applicable to various related tasks. These features capture low level patterns and high-level concepts often transferable across different domains.

After gaining a basic understanding of transfer learning, let us discuss the approach to self-supervised learning employed in this study. So instead of relying on human-labelled data, self-supervised learning leverages the abundant unlabelled data that is often easier to obtain. The key idea is to design pretext tasks that require the model to learn valuable representations or predictive patterns from the data. These learned representations can then be transferred to downstream tasks or fine-tuned for specific tasks where labelled data is scarce or expensive.

Self-supervised learning can be implemented in various ways, and some popular methods include:

- **Contrastive learning:** Contrastive learning aims to maximize the similarity between similar examples and minimize the similarity between different examples. It trains the model to distinguish positive pairs (similar examples) from negative pairs (dissimilar examples) in the latent space. By doing so, the model learns representations that capture relevant information about the data.
- **Generative models:** Generative models, such as autoencoders or generative adversarial networks (GANs), are used in self-supervised learning. These models are trained to reconstruct the input data from a compressed representation. By learning to reconstruct the original data, the models implicitly learn meaningful representations that capture the underlying structure of the data.

3.4.1 Network architecture

For this case, a task known as "image inpainting" or "patch prediction" was utilized. This task aims to train the model to fill in the missing patches based on the surrounding context of the image. By removing random patches from an image and asking the model to predict the missing content, information about the image's structure and context is implicitly provided to the model. The model needs to understand the spatial relationships and dependencies between different parts of the image to make accurate predictions. The model, training process, and inspiration were obtained from Kaiming He et al.'s masked autoencoder [29]. The ViT backbone was pre-trained in a self-supervised manner on the crowd-counting database. The data comprises several training datasets, as more data is needed to observe improved or better results. Unlabelled data available on the Internet will be utilized for this purpose.

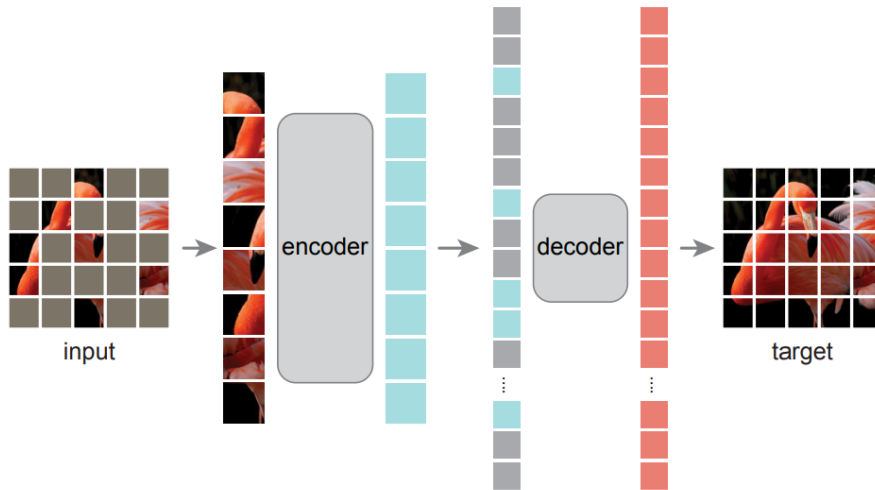


Figure 3.12: Masked autoencoder [29]

Another method being explored is multi-tasking, as it has demonstrated improved results in other fields, as proven by Liang et al.'s Supervised MAE [30].

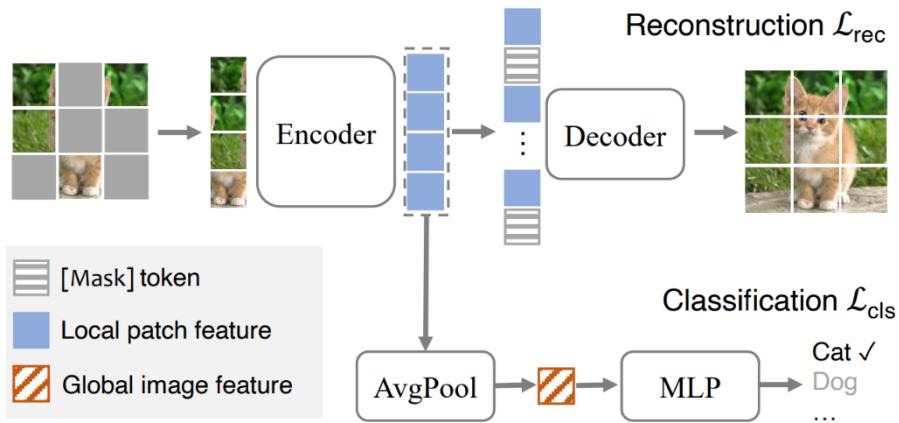


Figure 3.13: supervised MAE [30]

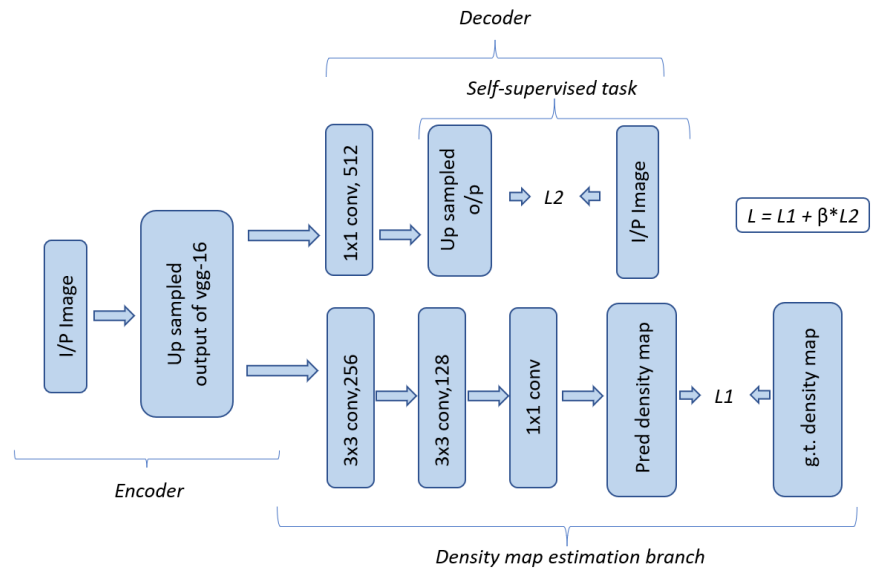


Figure 3.14: Multitask model.

Chapter 4

Results and discussion

4.1 Frequency domain analysis

The Shanghaitech-A dataset was utilized for this study. The MAE and MSE values obtained for the model are nearly the same as the base model. Here the β value indicates the contribution of frequency loss L_f in the model's training.

Before		
Sht A		
MAE	MSE	
86.7	132.0	

After		
Sht A		
β	MAE	MSE
0	85.78	130.92
0.0001	86.02	132.73
0.001	85.00	133.07
0.1	86.81	131.90

Table 4.1

The research aimed to enhance the model's ability to learn specific patterns within density maps by incorporating additional frequency maps for supervision. However, a difference in approach existed between the work conducted and the main paper that was referenced. The main paper utilized a characteristic function with mathematical properties that enable quantitative counting and supervision based on the ground truth and predicted frequency maps. In contrast, the same level of quantitative count supervision

could not be achieved using the frequency map obtained through Fast Fourier Transform (FFT) due to the inability to accurately determine the count from it.

4.2 Classification approach

After		Before	
JHU-Full		JHU-Full	
MAE	MSE	MAE	MSE
72	220	54.03	204.38

Table 4.2

For this analysis, the JHU-crowd dataset was used. Upon examining the Mean Squared Error (MSE) and Mean Absolute Error (MAE) values, it becomes apparent that the model's performance is significantly poorer than the base model. The reasons for these results are as follows. First, this model has so many dependencies that the whole model comprises four sub models: three density estimation models for each class and one classifier model. To get better results, all of them need to perform well. The classifier is not efficient, and the reason for that is it might be tough to find the distribution for the data of each class as some models trained on low density gave good results on some of the high-density images and so it might be difficult for regular single column model to capture this diverse nature. So, Res-Next-50 or other multi-column models should have been used to capture multi-scale information. Additionally, the models trained on high-density images yielded inferior results, possibly due to using the same network architecture for all classes without adjusting their receptive fields. Therefore, experiments were initiated to enhance the individual models, particularly those trained on high-density data.

4.3 Multiscale analysis

Multiscale model 1: Results for JHU-high and JHU-full are given below. Both metrics have shown improvement compared to the base model. Notably, the significant improvement in MSE suggests enhanced generalization, robustness, and stability of the model in dealing with outliers.

Before		After	
JHU-Full		JHU-Full	
MAE	MSE	MAE	MSE
54.03	204.38	58.63	191.01
		52.69	194.65
		52.68	204.69

Table 4.3

Before		After	
JHU-High $n \geq 100$		JHU-High $n \geq 100$	
MAE	MSE	MAE	MSE
110.63	297.34	93.96	282.63
		99.55	284.99

Table 4.4

In the case of the Shanghaitech-A dataset, certain pre-trained layers of VGG16 were frozen to address the overfitting issue.

Before		After		
Sha-A		Sha-A		
MAE	MSE	MAE	MSE	Frozen layers
65.60	103.13	62.11	99.37	9
		65.09	98.38	9
		61.25	102.22	5

Table 4.5

Multiscale model 2: Improved MSE for JHU-full and Shanghaitech-A says that the extra column with dilated convolution improves the stability and generalization capability of the model.

Before		
Sha-A		
MAE	MSE	
65.60	103.13	

After		
Sha-A		
MAE	MSE	Frozen layers
64.58	97.80	9
67.13	96.67	9
67.78	100.06	9

Table 4.6

Before		After	
JHU-Full		JHU-Full	
MAE	MSE	MAE	MSE
54.03	204.38	53.87	190.63

Table 4.7

4.4 Self-supervised

Initially, a task was attempted where the model encodes the input into a lower-dimensional representation. This multitasking approach was applied to JHU-high-density data.

Before		
JHU-High $n \geq 100$		
MAE	MSE	
110.63	297.34	

After		
JHU-High $n \geq 100$		
MAE	MSE	β
116.44	255.35	0.1
112.83	275.26	0.01

Table 4.8

Although the results obtained are satisfactory, there is a need for improved stability. Stability refers to consistently reproducing these results on the same dataset and achieving favourable outcomes on other datasets. The experimentation phase is underway, exploring various adjustments to hyperparameters, incorporating self-learning tasks, and employing other approaches to enhance stability.

Chapter 5

Conclusion and future work

5.1 Conclusions

Crowd counting faces challenges in handling complex scenarios where current models struggle. Developing lightweight models for real-time operation and training is crucial to ensure practical implementation. These efficient and fast models enable real-time crowd analysis. Achieving good generalization across diverse crowd distributions is vital for accurate counting. Overcoming the limitations in data annotation and addressing these challenges will drive the advancement of crowd-counting techniques, enabling more accurate and efficient crowd analysis in real-world applications.

5.2 Future works

There are several exciting directions for advancing crowd counting. Firstly, refining contextual information and exploring more informative features such as spatial layout, crowd dynamics, and environmental factors can enhance accuracy. Additionally, leveraging self-supervised and weakly supervised training techniques with transformers can improve performance. Creating comprehensive datasets, incorporating domain adaptation methods, and exploring hybrid models combining transformers with other techniques are promising future research avenues. These advancements will contribute to more robust and generalized crowd-counting systems with crowd management, security, and urban planning applications.

Bibliography

- [1] Z. Lin and L. S. Davis, “Shape-based human detection and segmentation via hierarchical parttemplate matching,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 32, no. 4, pp. 604–618, 2010.
- [2] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.
- [3] S.-F. Lin, J.-Y. Chen, and H.-X. Chao, “Estimation of the number of people in crowded scenes using perspective transformation,” *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 31, no. 6, pp. 645–654, 2001.
- [4] K. Chen, C. C. Loy, S. Gong, and T. Xiang, “Feature mining for localised crowd counting.,” in *Bmvc*, vol. 1, 2012, p. 3.
- [5] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, Ieee, vol. 1, 2005, pp. 886–893.
- [6] X. Cao, Z. Wang, Y. Zhao, and F. Su, “Scale aggregation network for accurate and efficient crowd counting,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 734–750.
- [7] C. Zhang, H. Li, X. Wang, and X. Yang, “Cross-scene crowd counting via deep convolutional neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 833–841.
- [8] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, “Multi-source multi-scale counting in extremely dense crowd images,” in

- Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2547–2554.
- [9] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-image crowd counting via multicolumn convolutional neural network,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 589–597.
 - [10] V. A. Sindagi and V. M. Patel, “Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting,” in 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE, 2017, pp. 1–6.
 - [11] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, “Decidenet: Counting varying density crowds through attention guided detection and density estimation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5197–5206.
 - [12] R. Girshick, “Fast r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
 - [13] L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang, “Multi-scale convolutional neural networks for crowd counting,” in 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 465–469.
 - [14] C. Szegedy, W. Liu, Y. Jia, et al., “Going deeper with convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
 - [15] X. Cao, Z. Wang, Y. Zhao, and F. Su, “Scale aggregation network for accurate and efficient crowd counting,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 734–750.
 - [16] Y. Li, X. Zhang, and D. Chen, “Csnet: Dilated convolutional neural networks for understanding the highly congested scenes,” in Proceedings of the IEEE conference

- on computer vision and pattern recognition, 2018, pp. 1091–1100.
- [17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.
 - [18] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, “Transcrowd: Weakly-supervised crowd counting with transformers,” *Science China Information Sciences*, vol. 65, no. 6, p. 160 104, 2022.
 - [19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
 - [20] Y. Tian, X. Chu, and H. Wang, “Cctrans: Simplifying and improving crowd counting with transformer,” *arXiv preprint arXiv:2109.14483*, 2021.
 - [21] X. Chu, Z. Tian, Y. Wang, et al., “Twins: Revisiting the design of spatial attention in vision transformers,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 9355–9366, 2021.
 - [22] Z. Ma, X. Wei, X. Hong, and Y. Gong, “Bayesian loss for crowd count estimation with point supervision,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6142–6151.
 - [23] B. Wang, H. Liu, D. Samaras, and M. H. Nguyen, “Distribution matching for crowd counting,” *Advances in neural information processing systems*, vol. 33, pp. 1595–1607, 2020.
 - [24] G. Peyré, M. Cuturi, et al., “Computational optimal transport: With applications to data science,” *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

- [25] W. Shu, J. Wan, K. C. Tan, S. Kwong, and A. B. Chan, "Crowd counting in the frequency domain," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19 618–19 627.
- [26] Q. Wang, J. Gao, W. Lin, and X. Li, "Nwpu-crowd: A large-scale benchmark for crowd counting and localization," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 6, pp. 2141–2149, 2020.
- [27] V. A. Sindagi, R. Yasarla, and V. M. Patel, "Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1221–1231.
- [28] H. Idrees, M. Tayyab, K. Athrey, et al., "Composition loss for counting, density map estimation and localization in dense crowds," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 532–546.
- [29] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16 000–16 009.
- [30] F. Liang, Y. Li, and D. Marculescu, "Supmae: Supervised masked autoencoders are efficient vision learners," arXiv preprint arXiv:2205.14540, 2022.
- [31] Khan, Muhammad Asif, Hamid Menouar, and Ridha Hamila. "Revisiting crowd counting: State-of-the-art, trends, and future perspectives." *Image and Vision Computing* (2022): 104597.