# [VISVESVARAYA TECHNOLOGICAL UNIVERSITY](http://www.vtu.ac.in/)

**“JNANA SANGAMA” BELAGAVI-590018, KARNATAKA**

**A PHASE-I PROJECT REPORT**

**on**

# *“*Detecting Crowd Density Map and Crowd counting using Deep Learning Techniques”

Submitted in the partial fulfillment of the requirement of the Seventh Semester

**Bachelor of Engineering In**

**Computer Science & Engineering**

ALDI ROHITH 1BO19CS004

BIJJULA JAYA SIMHA REDDY 1BO19CS025

CH. MAHESWAR REDDY 1BO19CS027

CHINTAPARTHI RETISH REDDY 1BO19CS031

Under The Guidance of

## Prof. Padmavathi H G

**Asst. Professor, Dept of CSE**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

Brindavan College of Engineering

**DWARAKANAGAR, BAGALUR MAIN ROAD, YELAHANKA, BANGALORE-63, 2022-23**

# ABSTRACT

Counting objects in images is one of the fundamental tasks in deep learning . Currently, deep learning (DL) methods provide the state-of-the-art performance in digital image processing. However, they require collecting a lot of annotated data, which is usually time consuming and prone to labeling errors. Crowd counting is a challenging problem due to the scene complexity and scale variation. Although deep learning has achieved great improvement in crowd counting, scene complexity affects the judgment of these methods and they usually regard some objects as people mistakenly; causing potentially enormous errors in the crowd counting result. To address the problem, we propose a novel end-to end model called Crowd Attention Convolutional Neural Network (CAT-CNN). Our CAT-CNN can adaptively assess the importance of a human head at each pixel location by automatically encoding a confidence map. With the guidance of the confidence map, the position of human head in estimated density map gets more attention to encode the final density map, which can avoid enormous misjudgments effectively. The crowd count can be obtained by integrating the final density map. To encode a highly refined density map, the total crowd count of each image is classified in a designed classification task and we first explicitly map the prior of the population-level category to feature maps. To verify the efficiency of our proposed method, extensive experiments are conducted on three highly challenging datasets. Results establish the superiority of our method over many state-of-the-art methods.

# Chapter 1

# INTRODUCTION

As the phenomenon of crowd congestion is becoming serious, safety- and security-oriented tasks– such as public safety control and traffic safety monitoring– face huge challenges. Manual analysis of the degree of crowd aggregation not only cannot achieve high accuracy but also will perform low efficiently. In contrast, deep-learning-based methods are more applicable at present since their process not only eliminates manual efforts but also can analyze crowd aggregation accurately and quickly. Among them, crowd estimation at the pixel level through the crowd distribution density maps has achieved tremendous progress. A crowd density map is a kind of image label that can reflect the distribution of crowd heads by processing the head coordinate value through Gaussian convolution.

As the convolutional layer and pooling layer of Convolutional Neural Network (CNN) strengthen the relationship between pattern recognition and the context in the image, the density estimation methods of CNN are with strong learning ability. They have achieved high accuracy in dense scenes. The accuracy of crowd counting mainly depends on the quality of the estimated density map which is limited by the image scale. Since the convolution kernel of CNN owns a static size, heads of dynamic scales will worsen the network’s performance, resulting in misjudgments and missing judgments. To solve this problem, the common methods are as follows: (1) introducing a multicolumn structure to estimate the crowd of different scales]; (2) introducing the idea of dilated convolution in the field from image segmentation . This is a special convolution for extracting feature information of different scales, consisting of a 33 convolution kernel and a dilated parameter. By setting the dilated parameter to replace redundant branches of different sizes of convolution kernels, the computational cost of multiscale detection can be reduced; (3) applying different detection methods to regions of different scales in the image. To generate a high-quality density map, spatial continuity should be ensured during the generation process so that the adjacent pixels in the output density graph can transition smoothly.

Crowd counting by computer vision technology plays an important role in safety management, video surveillance, and urban planning. The method of crowd counting can be also extended to other applications, such as cell counting, animal counting, and vehicle counting. However, due to the severe occlusion, scale variation, and high density in the crowd scene, crowd counting is still a challenging task.

To address these problems, a lot of efforts have been done in previous works including detection based methods and regression-based methods. Detection-based methods usually detect the instances of each person with pre-trained detectors. In the sparse crowd scene, they count the crowd accurately, while their accuracies are downgraded in the congested scene. Regression-based methods regress the number of the crowd without detecting people. They implement an implicit mapping between low-level features and crowd counts. However, the location information of the crowd is omitted. So that many CNN-based methods with state-of-the-art results are proposed recently. Most of them map the image to a density map that is more robust than the hand-crafted features. The quantity and location of the crowd at each pixel location are recorded in the density map. The crowd count can be obtained by integrating the density map.

Although CNN-based methods have achieved significant success in crowd counting, we find an important problem that needs to be solved urgently. Due to the complexity of crowd scenes, CNN-based methods usually mistake some objects as the head of people. As shown in Fig. 1, there are no people inside the red box, however, MCNN regards the dense shrubberies as human heads by mistake, which results in enormous errors of crowd counting. To address the above problem, we propose a novel end to-end model called CAT-CNN. An overview of the proposed CAT-CNN, It contains four modules: Multi-information Handling Module, Confidence Module, Density Map Estimation Module, and Fusion Module. The Multi-information Handling Module is utilized to extract robust features for crowd counting. Motivated by [2, 45], we leverage different convolution kernels to encode the input image at the beginning, then we fuse rich hierarchies from different convolutional layers, which is significant for extracting multi-scale features. In addition, the total crowd count of each image is classified in a designed crowd count group classifier. To the best of our knowledge, we first explicitly map the weights of predicted class to feature maps to automatically contribute in encoding a highly refined density map. In the Confidence Module, we classify each pixel to obtain the probability of a human head at each pixel location to encode the confidence map. Unfortunately, the ground-truth confidence map is not provided in present crowd counting datasets. We propose a simple but effective way to obtain the ground-truth confidence map by pasting the ones template on a binary map. The intensive cost of manual labeling is saved. Meanwhile, to address the problem of unbalanced population distribution, we propose the weighted Binary Cross-Entropy Loss (BCELoss) to encode a robust confidence map for population distribution. In the Density Map Estimation Module, the estimated density map is encoded.

### 1.1 Cconvolutional Neural Network (CNN ):

A convolutional neural network (CNN or convnet) is a subset of machine learning. It is one of the various types of artificial neural networks which are used for different applications and data types. A CNN is a kind of network architecture for deep learning algorithms and is specifically used for image recognition and tasks that involve the processing of [pixel](https://www.techtarget.com/whatis/definition/pixel) data.

There are other types of neural networks in deep learning, but for identifying and recognizing objects, CNNs are the network architecture of choice.

**1.2 CNN Architecture**

A CNN's architecture is analogous to the connectivity pattern of the human brain. Just like the brain consists of billions of neurons, CNNs also have neurons arranged in a specific way. In fact, a CNN's neurons are arranged like the brain's frontal lobe, the area responsible for processing visual stimuli. This arrangement ensures that the entire visual field is covered, thus avoiding the piecemeal image processing problem of traditional neural networks, which must be fed images in reduced-resolution pieces. Compared to the older networks, a CNN delivers better performance with image inputs, and also with speech or audio signal inputs.

### 1.3 CNN layers

A deep learning CNN consists of three layers: a convolutional layer, a pooling layer and a fully connected (FC) layer. The convolutional layer is the first layer while the FC layer is the last.

From the convolutional layer to the FC layer, the complexity of the CNN increases. It is this increasing complexity that allows the CNN to successively identify larger portions and more complex features of an image until it finally identifies the object in its entirety.

* **Convolutional layer.**The majority of computations happen in the convolutional layer, which is the core building block of a CNN. A second convolutional layer can follow the initial convolutional layer. The process of convolution involves a [kernel](https://www.techtarget.com/searchdatacenter/definition/kernel) or filter inside this layer moving across the receptive fields of the image, checking if a feature is present in the image.

Over multiple iterations, the kernel sweeps over the entire image. After each iteration a [dot product](https://www.techtarget.com/whatis/definition/dot-product-scalar-product) is calculated between the input pixels and the filter. The final output from the series of dots is known as a feature map or convolved feature. Ultimately, the image is converted into numerical values in this layer, which allows the CNN to interpret the image and extract relevant patterns from it.

* **Pooling layer.**Like the convolutional layer, the pooling layer also sweeps a kernel or filter across the input image. But unlike the convolutional layer, the pooling layer reduces the number of parameters in the input and also results in some information loss. On the positive side, this layer reduces complexity and improves the efficiency of the CNN.
* **Fully connected layer.**The FC layer is where image classification happens in the CNN based on the features extracted in the previous layers. Here, fully connected means that all the inputs or nodes from one layer are connected to every activation unit or node of the next layer.

All the layers in the CNN are not fully connected because it would result in an unnecessarily dense network. It also would increase losses and affect the output quality, and it would be computationally expensive.

### 1.4 Applications of convolutional neural networks

Convolutional neural networks are already used in a variety of CV and image recognition applications. Unlike simple image recognition applications, CV enables computing systems to also extract meaningful information from visual inputs (e.g., digital images) and then take appropriate action based on this information.

The most common applications of CV and CNNs are used in fields such as the following:

* **Healthcare.** CNNs can examine thousands of visual reports to detect any anomalous conditions in patients, such as the presence of malignant cancer cells.
* **Automotive**. CNN technology is powering research into autonomous [vehicles](https://www.techtarget.com/whatis/definition/vehicle-intellgence) and self-driving cars.
* **Social media.** [Social media](https://www.techtarget.com/whatis/definition/social-media) platforms use CNNs to identify people in a user's photograph and help the user tag their friends.
* **Retail.** [E-commerce](https://www.techtarget.com/searchcio/definition/e-commerce) platforms that incorporate visual search allow brands to recommend items that are likely to appeal to a shopper.
* **Facial recognition for law enforcement.** Generative adversarial networks ([GANs](https://www.techtarget.com/searchenterpriseai/definition/generative-adversarial-network-GAN)) are used to produce new images that can then be used to train deep learning models for facial recognition
* **Audio processing for virtual assistants.**CNNs in [virtual assistants](https://www.techtarget.com/searchcustomerexperience/definition/virtual-assistant-AI-assistant) learn and detect user-spoken keywords and process the input to guide their actions and respond to the user.

**CHAPTER 2**

**LITERATURE SURVEY**

**Title:** Crowd Counting and Density Estimation by Trellis Encoder-Decoder Networks

**Year:** 2019

**Publications:** Computer Vision and Pattern Recognition

**Description:**

In this paper, we have presented a new deep learning architecture, called the trellis encoder-decoder network (TEDnet) for crowd counting. It consists of a multi-scale encoder and a multi-path decoder to generate high-quality density estimation maps. It preserves the localization precision in the encoded feature maps, upon which a multi-path decoder with dense skip connections is adopted to achieve thorough aggregation and fusion of multi-scale features.

The TEDnet is trained with the distributed supervision implemented with the proposed combinatorial loss. Experiments on four benchmarks show that the TEDnet achieves new state-of-the-art performance in terms of both density map quality and crowd counting accuracy.

Crowd counting has recently attracted increasing interest in computer vision but remains a challenging problem. In this paper, we propose a trellis encoder-decoder network (TEDnet) for crowd counting, which focuses on generating high-quality density estimation maps. The major contributions are four-fold. First, we develop a new trellis architecture that incorporates multiple decoding paths to hierarchically aggregate features at different encoding stages, which improves the representative capability of convolutional features for large variations in objects. Second, we employ dense skip connections interleaved across paths to facilitate sufficient multi-scale feature fusions, which also helps TEDnet to absorb the supervision information. Third, we propose a new combinatorial loss to enforce similarities in local coherence and spatial correlation between maps. By distributedly imposing this combinatorial loss on intermediate outputs, TEDnet can improve the back-propagation process and alleviate the gradient vanishing problem. Finally, on four widely-used benchmarks, our TEDnet achieves the best overall performance in terms of both density map quality and counting accuracy, with an improvement up to 14% in MAE metric. These results validate the effectiveness of TED net for crowd counting.

**Title:** Active Crowd Counting with Limited Supervision

**Year:** July 2020

**Publications:** Computer Vision and Pattern Recognition

**Description:**

To learn a reliable people counter from crowd images, head center annotations are normally required. Annotating head centers is however a laborious and tedious process in dense crowds. In this paper, we present an active learning framework which enables accurate crowd counting with limited supervision: given a small labelling budget, instead of randomly selecting images to annotate, we first introduce an active labelling strategy to annotate the most informative images in the dataset and learn the counting model upon them. The process is repeated such that in every cycle we select the samples that are diverse in crowd density and dissimilar to previous selections. In the last cycle when the labelling budget is met, the large amount of unlabeled data are also utilized: a distribution classifier is introduced to align the labelled data with unlabelled data; furthermore, we propose to mix up the distribution labels and latent representations of data in the network to particularly improve the distribution alignment in-between training samples. We follow the popular density estimation pipeline for crowd counting. Extensive experiments are conducted on standard benchmarks i.e. ShanghaiTech, UCF CC 50, MAll, TRANCOS, and DCC. By annotating limited number of images (e.g. 10% of the dataset), our method reaches levels of performance not far from the state of the art which utilize full annotations of the dataset.

We present an active learning framework for accurate crowd counting with limited supervision. Given a counting dataset, instead of annotating every image, we introduce a partition-based sample selection with weights to label only a few most informative images and learn a crowd regression network upon them. This process is iterated till the labelling budget is reached. Next, rather than learning from only labelled data, the abundant unlabelled data are also exploited: we introduce a distribution alignment branch with latent Mix Up in the network.

Experiments conducted on standard benchmarks show that labeling only 10% of the entire set, our method already performs close to recent state-of-the-art.

**Title:** Crowd Detection Using Deep Learning

**Year: August 2021**

**Publications:** International Research Journal of Engineering and Technology (IRJET)

**Description:**

This way of life makes life easier for people and increases the use of public services in cities. We present a CNN-MRF-based method for counting people in still images from various scenes. Crowd density is well represented by the features derived from the CNN model trained for other computer vision tasks. The neighboring local counts are strongly correlated when using the overlapping patches separated strategies. The MRF may use this connection to smooth adjacent local counts for a more accurate overall count. We divide the dense crowd visible image into overlapping patches, and then extract features from each patch image using a deep convolutional neural network, followed by a completely connected neural network to regress the local patch crowd count. Since the local patches overlap, there is a strong connection between the crowd counts of neighboring patches. We smooth the counting effects of the local patches using this connection and the Markov random field.

We present a CNN based method for counting people in still images from various scenes. Crowd density is well represented by the features derived from the CNN model trained for other computer vision tasks. The neighboring local counts are strongly correlated when using the overlapping patches separated strategies. The feature extraction may use this connection to smooth adjacent local counts for a more accurate overall count. Experimental findings show that the proposed method outperforms other recent related methods.

**Title:** **Crowd Density Estimation Using Image Processing: A Survey**

**Year:2018**

**Publications:** International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 9 (2018)

**Description:**

People counting is a crucial subject in video surveillance application. Factors such as severe occlusions, scene perspective distortions in real time application make this task a bit more challenging. The use of Infra Red (IR) sensors and Channel State Information (CSI) of the WIFI network, which are the classical methods, give the count but have their own range constraints and its limited applicability to controlled environment. Video-surveillance systems are one of the advanced technologies used to estimate the density of people in a place for security reasons and to obtain the human statistics. The vision based techniques works well when people are in motion and when a high resolution image with clear background are available. This paper presents the state of the art of such image processing algorithms which are used for crowd estimation and their related applications.

This paper presented the crowd density estimation methods which have provided the most satisfactory results. Marana and team has explained about the real time crowd estimation based on texture features and similarly the other authors have provided their ways of dealing the small and tight crowds. The characteristics obtained from the CNN model trained showed a strong capacity to count the crowd, and principles of GrC are used to gestate crowd segmentation issue at different granular levels and other such algorithms of estimation. Further, these methods can be revised to form new algorithms with multiple advantages and can be implemented for a specific application like monitoring the crowd in shopping malls, in uncontrolled environments like bus stands and railway stations thereby preventing congestions and provide comfort.

**Title:** A Survey of Recent Advances in CNN-based Single Image Crowd Counting and Density Estimation

**Year: 2017**

**Publications: Elsevier,** Pattern Recognition Letters

**Description:**

Estimating count and density maps from crowd images has a wide range of applications such as video surveillance, traffic monitoring, public safety and urban planning. In addition, techniques developed for crowd counting can be applied to related tasks in other fields of study such as cell microscopy, vehicle counting and environmental survey. The task of crowd counting and density map estimation is riddled with many challenges such as occlusions, non-uniform density, intra-scene and inter-scene variations in scale and perspective. Nevertheless, over the last few years, crowd count analysis has evolved from earlier methods that are often limited to small variations in crowd density and scales to the current state-of-the-art methods that have developed the ability to perform successfully on a wide range of scenarios. The success of crowd counting methods in the recent years can be largely attributed to deep learning and publications of challenging datasets. In this paper, we provide a comprehensive survey of recent Convolutional Neural Network (CNN) based approaches that have demonstrated significant improvements over earlier methods that rely largely on hand-crafted representations. First, we briefly review the pioneering methods that use hand-crafted representations and then we delve in detail into the deep learning-based approaches and recently published datasets. Furthermore, we discuss the merits and drawbacks of existing CNN-based approaches and identify promising avenues of research in this rapidly evolving field.

This article presented an overview of recent advances in CNN-based methods for crowd counting and density estimation. In particular, we summarized various methods for crowd

counting into traditional approaches (that use hand-crafted features) and CNN-based approaches. The CNN-based approaches are further categorized based on the training process

and the network property. Obviously all the literature on crowd counting cannot be covered, hence, we have chosen a representative subset of the latest approaches for a detailed analysis and review. We also reviewed the results demonstrated by various traditional and CNN-based approaches to conclude that CNN based methods are more adept at handling large density crowds with variations in object scales and scene perspective. Additionally, we observed that incorporating scale and contextual information in the CNN-based methods drastically improves the estimation error. Finally, we identified some of the most compelling challenges and issues that confront research in crowd counting and density estimation using computer vision and machine leaning approaches.

# Title: Near Real-time Crowd Counting using Deep Learning Approach

**Year:** 2020

## Publications: Elsevier, Procedia Computer Science

**Description:**

In the current digital era, at many places crowd counting mechanisms still rely on old-fashioned methods such as maintaining registers, making use of people counters and sensors based counting at entrance. These methods fail in the places where the movement of people is completely random, highly variable and dynamic. These methods are time consuming and tedious. The proposed system is developed for situations where emergency evacuations are required such as fire outbreaks, calamitous events, etc. and making informed decisions on the basis of the number of people such as food, water, detecting congestion, etc. A deep convolution neural network (DCNN) based system can be used for near real-time crowd counting. The system uses NVIDIA GPU processor to exploit the parallel computing framework to achieve swift and agile processing of the video feed taken through a camera. This work contributes towards constructing a model to detect heads captured by CCTV camera. The model is trained extensively by providing several scenarios such as overlapping heads, partial visibility of heads etc. This system provides significant accuracy in estimating the head count in dense population in reasonably less amount of time. This work presents an approach which will be effective for near real time crowd counting using DCNN. The benefits of the applications includes High Performance Computing through the use NVIDIA GPU parallel framework, a swift and agile method for processing of the video feed taken through a camera with an innovative solutions that can be deployed for disaster management, emergency evacuation without having to configure explicit systems for the same. The proposed system performs admirably in situations where manual counting is simply not possible. Deep learning also enables the system to perform in versatile environments and continuously learn from new inputs. The Experimental results reveal that the proposed methodology achieves promising crowd count predictions almost as good as ground truth. Another major advantage of using the end-to-end application is that no external configurations are required for achieving crowd-count except for the video feed of the particular area.

**Title:** Crowd Density Estimation in Spatial and Temporal Distortion Environment Using Parallel Multi-Size Receptive Fields and Stack Ensemble Meta-Learning

**Year: August 2022**

**Publications: Symmetry, MDPI.**

**Description:**

The estimation of crowd density is crucial for applications such as autonomous driving, visual surveillance, crowd control, public space planning, and warning visually distracted drivers prior to an accident. Having strong translational, reflective, and scale symmetry, models for estimating the density of a crowd yield an encouraging result. However, dynamic scenes with perspective distortions and rapidly changing spatial and temporal domains still present obstacles. The main reasons for this are the dynamic nature of a scene and the difficulty of representing and incorporating the feature space of objects of varying sizes into a prediction model. To overcome the aforementioned issues, this paper proposes a parallel multi-size receptive field units framework that leverages the majority of the CNN layer’s features, allowing for the representation and participation in the model prediction of the features of objects of all sizes. The proposed method utilizes features generated from lower to higher layers. As a result, different object scales can be handled at different framework depths, and various environmental densities can be estimated. However, the inclusion of the vast majority of layer features in the prediction model has a number of negative effects on the prediction’s outcome. Asymmetric non-local attention and the channel weighting module of a feature map are proposed to handle noise and background details and re-weight each channel to make it more sensitive to important features while ignoring irrelevant ones, respectively. While the output predictions of some layers have high bias and low variance, those of other layers have low bias and high variance. Using stack ensemble meta-learning, we combine individual predictions made with lower-layer features and higher-layer features to improve prediction while balancing the tradeoff between bias and variance. The UCF CC 50 dataset and the ShanghaiTech dataset have both been subjected to extensive testing. The results of the experiments indicate that the proposed method is effective for dense distributions and objects of various sizes.

**Title:** People Counting in High Density Crowds from Still Images

**Year:** October 16, 2015.

**Publications:** International Journal of Computer and Electrical Engineering

**Description:**

They present a method of estimating the number of people in high density crowds (hundreds to thousands of individuals) from still images. Unlike most existing works our method uses only still images to estimate the count. At this scale, we cannot rely on just one set of features for count estimation. We, therefore, use a fusion of multiple sources, viz. interest points (SIFT), Fourier analysis, wavelet decomposition, GLCM features and head detections, to estimate the counts. Each of these sources gives a separate estimate of the count along with confidences and other statistical measures which are then combined to obtain the final estimate. We tested our method on an existing dataset of fifty images containing over 64000 individuals. Further, we added another fifty annotated images of crowds and tested on the complete dataset of hundred images containing over 87000 individuals.

They considered a method for estimating the number of people in extremely dense crowds from still images. The counting problem at this scale has barely been tackled before. We presented a method that uses information from multiple sources of information (head detections, interest points based counting and texture analysis methods) to estimate the count in an image. Each of these constituent parts gives an independent estimate of the count, along with confidences and other features, which are then fused to give a final estimate. We presented results of extensive tests and experiments we performed. We also introduced a new dataset of still images along with annotations which can complement the existing UCF dataset. The results are very promising and, since the model is extremely simple, it can be applied for real-time counting in critical areas like pilgrimage sites.

**Title:** A Survey on Human detection in Crowd Density Estimation for Video Surveillance

**Year:** December, 2021

**Publications:** International Journal of Mechanical Engineering

**Description:**

Crowds can be seen in numerous day-to-day life situations and it'll be engaging to recognize, dissect and break the challenges involved in crowd density estimation. The density of a crowd is a vital parameter in several operations like operation of crowd for safety and surveillance for law enforcement, development of public transport structure which have been divided using automated or semi-automated computer vision ways. Mortal discovery in a videotape surveillance system has vast operation areas including suspicious event discovery and mortal exertion recognition. In the current terrain of our society suspicious event discovery is a burning issue. For that reason, this paper proposes a frame for detecting humans in different appearances and acts by generating a mortal point vector. Originally, every pixel of a frame is represented as an objectification of several Gaussians and use a probabilistic system to refurbish the representation. These Gaussian representations are also estimated to classify the background pixels from focus pixels. Shadow regions are excluded from focus by exercising a Hue-Intensity difference value between background and current frame. Also morphological operation is used to remove discontinuities in the focus uprooted from the shadow elimination process. Partial occlusion running is employed by color correlogram to marker objects within a group.

**Title:** Utilization of Deep Learning-Based Crowd Analysis for Safety Surveillance and Spread Control of COVID-19 Pandemic

**Year: 2022**

**Publications:** Intelligent Automation & Soft Computing

**Description:**

Crowd monitoring analysis has become an important challenge in academic researches ranging from surveillance equipment to people behavior using different algorithms. The crowd counting schemes can be typically processed in two steps, the images ground truth density maps which are obtained from ground truth density map creation and the deep learning to estimate density map from density map estimation. The pandemic of COVID-19 has changed our world in few months and has put the normal human life to a halt due to its rapid spread and high danger. Therefore, several precautions are taken into account during COVID-19 to slowdown the new cases rate like maintaining social distancing via crowd estimation. This manuscript presents an efficient detection model for the crowd counting and social distancing between visitors in the two holy mosques, Al Masjid Al Haram in Mecca and the Prophet’s Mosque in Medina. Also, the manuscript develops a secure crowd monitoring structure based on the convolutional neural network (CNN) model using real datasets of images for the two holy mosques. The proposed framework is divided into two procedures, crowd counting and crowd recognition using datasets of different densities. To confirm the effectiveness of the proposed model, some metrics are employed for crowd analysis, which proves the monitoring efficiency of the proposed model with superior accuracy. Also, it is very adaptive to different crowd density levels and robust to scale changes in several places.

**Title:** CrowdSurge: A Crowd Density Monitoring Solution Using Smart Video Surveillance with Security Vulnerability Assessment

**Year:2022**

**Publications:** Journal of Advances in Information Technology Vol. 13, No. 2, April 2022

**Description:**

Overcrowding and crowd density monitoring in various places and establishments are being implemented since the pandemic, which helps observe social distancing. This study is about the development of a crowd density solution by utilizing YOLOv4 and Closed-Circuit Television (CCTV) called CrowdSurge. The practice of CCTV has been around for so many years with proven benefits. This has been combined with the state-of-the-art YOLOv4 algorithm that provides high video analytics and object detection performance. With the combination of the said technology and algorithm, it will serve as a smart surveillance system. A system and mobile application have been developed, and the YOLOv4 deep learning detection model was used to detect various set of scenarios considered to assess if the model executes according to the actions assigned in the experimental set-up. The browser-based application was tested using CVSS or Common Vulnerability Scoring system, which shows that the severity level of most vulnerabilities is low and has a minor impact on the system. Based on the overall usability testing and statistical results, the respondents are satisfied with both surveillance system and mobile applications developed in terms of functionality, usefulness, and aesthetics. Therefore, using the developed system in realtime surveillance can aid in crowd density reduction in an area.

# Chapter 3

# System Analysis

# 3.1 Existing System

# In recent years, crowd counting has drawn much attention and various methods have been proposed, especially in deep learning. Next, we will give these methods some introductions.

1. Traditional detection-based algorithms such as Haar wavelets , HOG , and LBP occupy an important position in early works.
2. Regression-based methods learn a mapping between high level features and crowd counts.
3. CNN-based methods
4. Image based Methods

# 3.2 Aim & Objective

## Aim:

The aim of the project is to propose a novel end-to end model called Crowd Attention Convolutional Neural Network (CAT-CNN) and other deep learning algorithms that can adaptively judge the position of a human head at each pixel location by automatically encoding a density map and count the number of Humans.

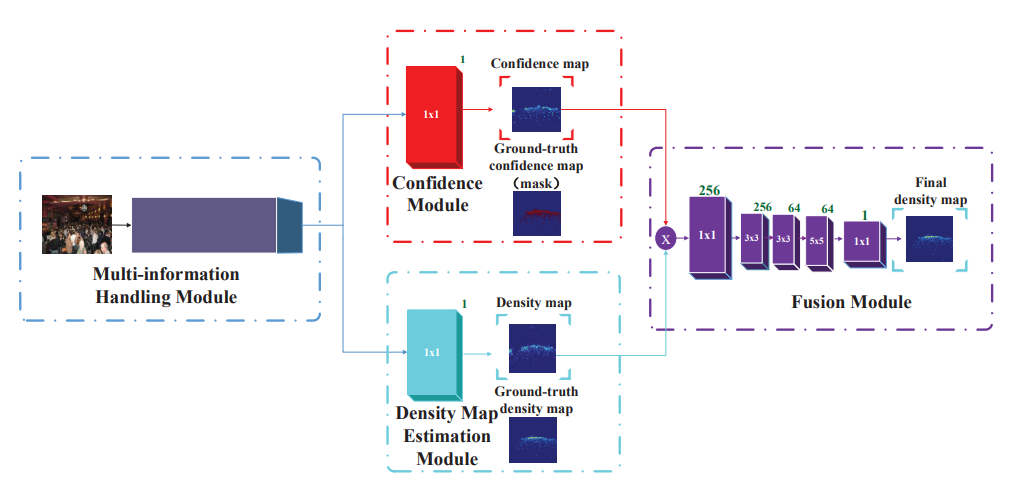
## Ojective:

Crowd counting is a challenging problem due to the scene complexity and scale variation. Although

Deep learning has achieved great improvement in crowd counting, scene complexity affects the judgment of these methods and they usually regard some objects as people mistakenly; causing potentially enormous errors in the crowd counting result.

1. The Crowd Dataset is collected from Machine Learning Repository
2. CAT-CNN and other deep learning techniques that can adaptively assess the importance of a human head at each pixel location to avoid enormous misjudgments in crowd counting.
3. Density Map Estimator to create high-dimensional feature maps
4. Design a novel classification model that can take input of arbitrary size for training in crowd counting.
5. And we first explicitly map the prior information of the population-level category of images to feature maps to automatically contribute in encoding a highly refined density map.
6. Predicting the Human count based on the density map.

# 3.3 Proposed System



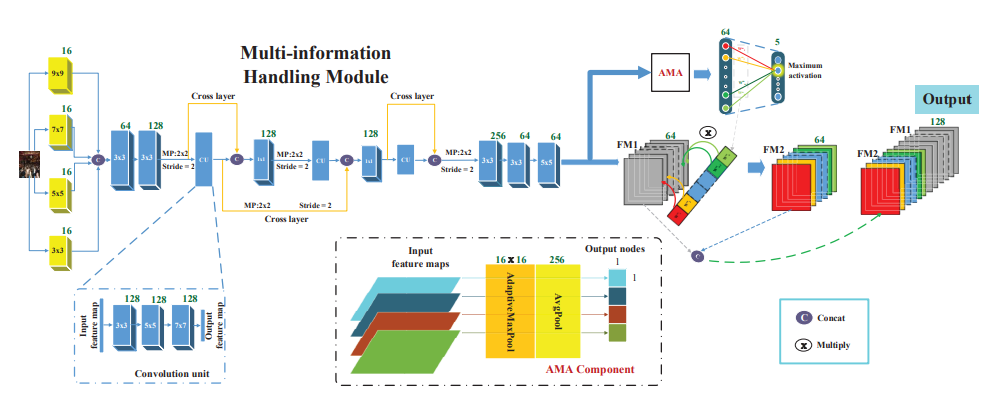


Figure 1: The proposed architecture of our CAT-CNN.

An overview of the proposed CAT-CNN is shown in Fig. 1. Our CAT-CNN is composed of three stages. The first stage contains the first module where the features which can automatically adapt different scales and different crowd count groups are extracted. The second stage consists of two modules in the middle to encode confidence map and estimated density map respectively. The third stage contains the final module. With the guidance of the confidence map, final density map is encoded from the estimated density map in this stage. Next, we will elaborate these modules in each stage.

**Chapter 4**

**SYSTEM REQUIREMENT & SPECIFICATIONS**

Requirements analysis is critical for project development. Requirements must be documented, actionable, measurable, testable and defined to a level of detail sufficient for system design. Requirements can be architectural, structural, behavioral, functional, and functional.

A software requirements specification (SRS) is a comprehensive description of the intended purpose and the environment for software under development.

**4.1 Functional Requirements**

The tools to execute the Python programs can be many, among that we can go with Visual Studio, Anaconda Navigator (Jupyter Notebook) or any IDLE based on Python. The online tool from Google can be an effective solution towards the execution of Python coding.

**4.1.1. Approach 1: Jupyter Notebook (Anaconda Navigator)**

This tool is also known as IPython Notebook, and it is Open-Source Distribution Software and provides the platform for development of web applications, computational interactive and specific environment for the users to create notebook documentations. It support for individual code execution , browser based interoperability, can plot various graphs using python libraries and also support for many open source libraries like Bootstrap, JQuery, Tornado, Matplotlib , Seaborn and others.

The features of Jupyter Notebook can be listed as:

* Flexible Notebook Interface
* Useful tool in Machine learning, Deep learning and Ai based Application and model Design.
* Creating and sharing the computational Documents.

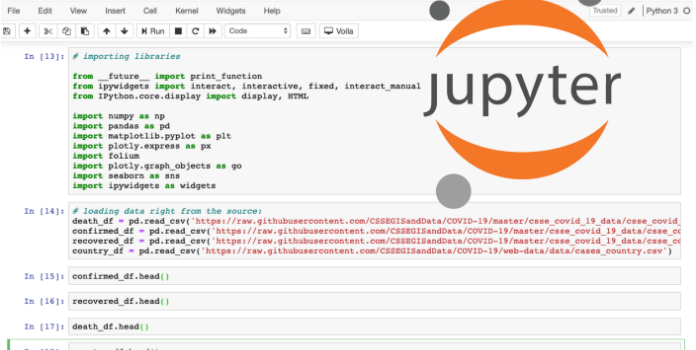


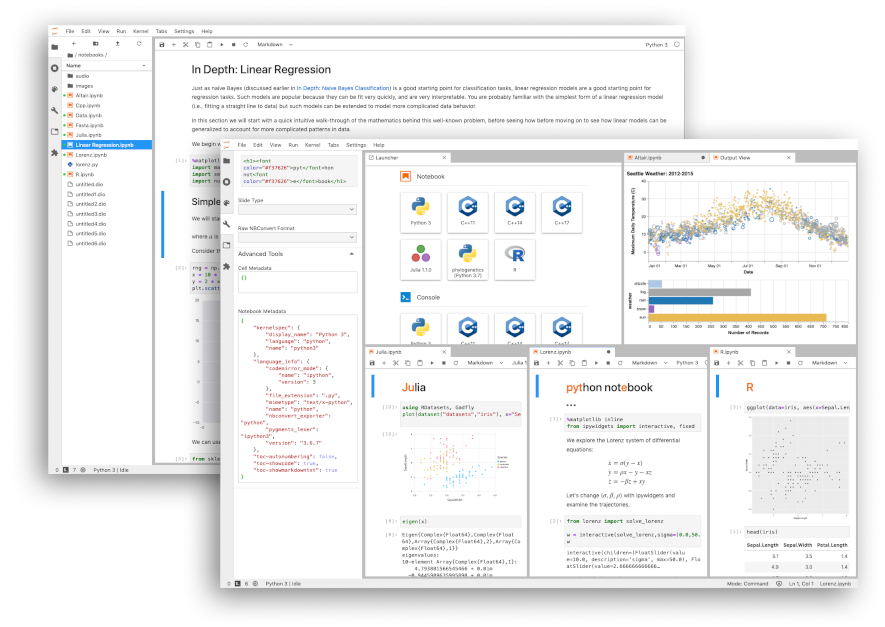
Figure 4.1 Jupyter Notebook Dashboards ****

Figure 4.2: Notebook support for plotting

**4.1.2 Approach 2: Python IDLE**

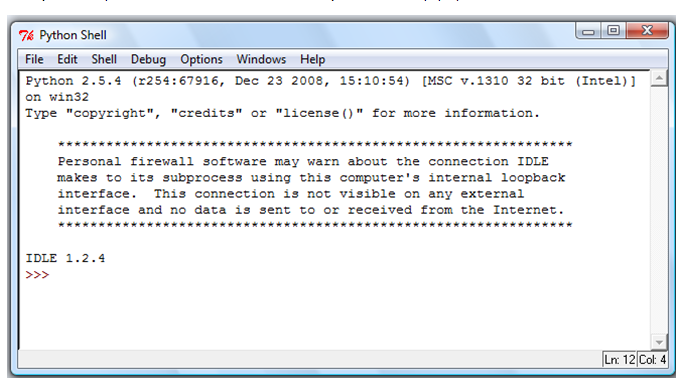
**Python IDLE (Python Integrated Development and Learning Environment) help is writing the code very effectively and efficiently and helpful tool to the Python learning who wants to start from the scratch and beginners can have an advantage to execute the code easily. This is a powerful interpreter and compiler to run the code.**

**It’s an Interactive Interpreter also known as shell, which executes the python written code, reads the input, evaluate the statements and print the output on the standard output screen provided.**

**File Editor Help to edit the code, save the program in text files and store as .py file.**



**Figure 4.3: Python IDLE Download Page**

****

**Figure 4.4: Python IDLE prompt to write and execute code.**

**4.1.3 Approach 3: Google Colab**

**Google Colab, Also called as Colab in short is a powerful Machine Learning, Deep Learning and Data Analysis Tool that allows mixing the Python script along with text document. Rich support for Plotting the graphs, Diagram, Charts, Import Images, HTML Tags Support and LATEX format API conversions. Additional functional is it works on cloud model where document can be accessed and run on any platform independent of framework design and operating system. The runtime support for Virtual Hard Disk space and 12GB of RAM to execute the application is very excited feature of Colab. The uploading of files is very easy in this application so that it connects to the runtime.**

**Some of the important feature is:**

* **Remote Desktop Connection**
* **Runtime Environment**
* **Dataset Upload Features**
* **I/O operations and Operating System API Support**
* **General Processing Unit (GPU) availability**

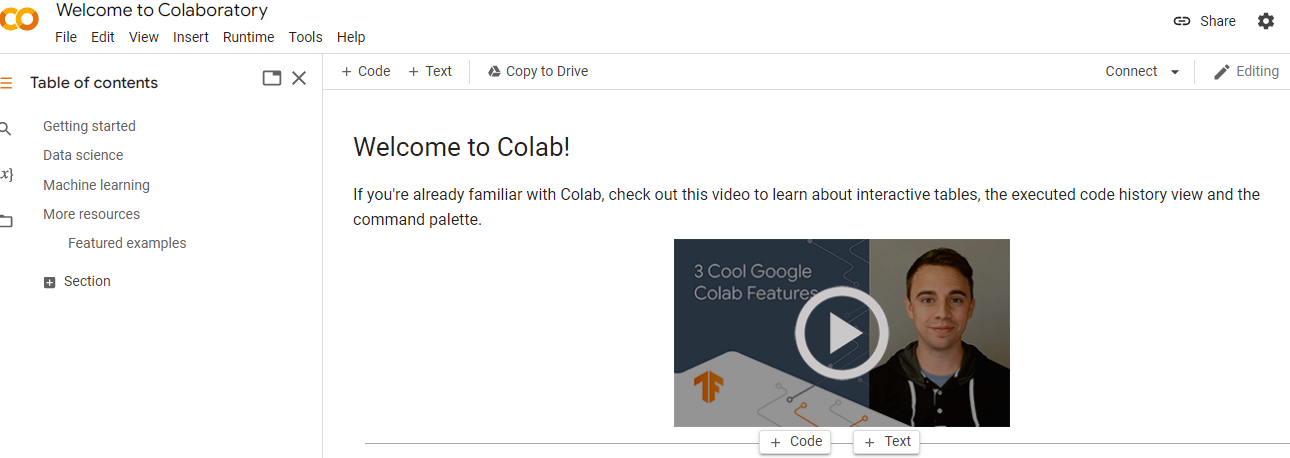


Figure 4.5: Welcome page of Google Colab

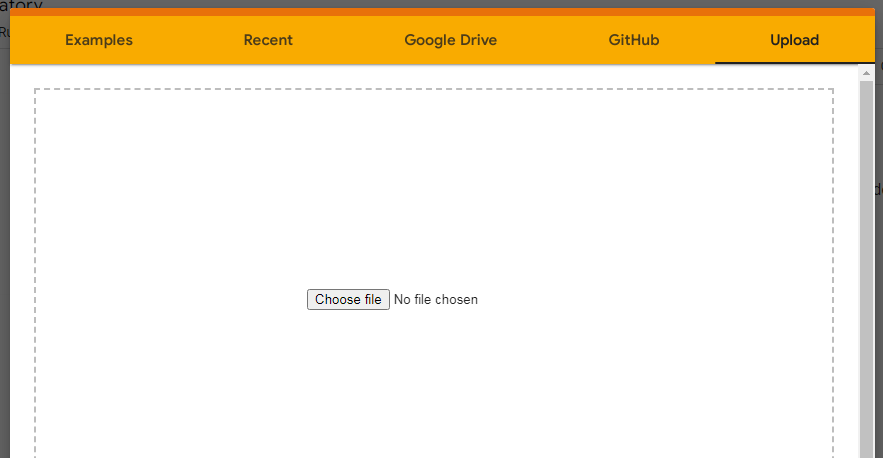


Figure 4.6: Upload the Notebook File

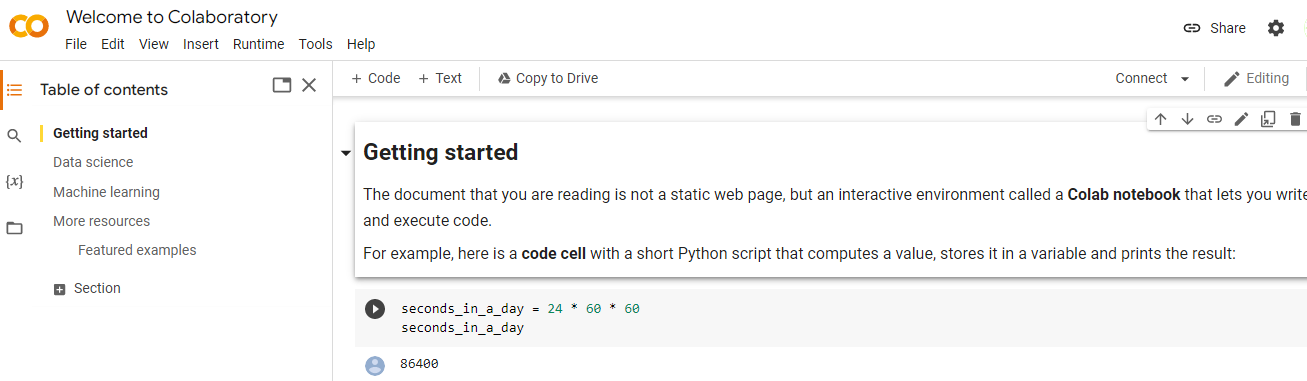


Figure 4.7 Start the Application Page

Requirements analysis is critical for project development. Requirements must be documented, actionable, measurable, testable and defined to a level of detail sufficient for system design. Requirements can be architectural, structural, behavioural, functional, and functional.

A software requirements specification (SRS) is a comprehensive description of the intended purpose and the environment for software under development.

**4.2 Software Requirements**

Scripting language : Python Programming

Scripting Tool : Anaconda Navigator (Jupyter Notebook) or Google Colab

Operating System : Microsoft Windows 8/ 10 or 11

Dataset : Crowd Dataset

Deep Learning Packages : Numpy, Pandas, Matplotlib , Seaborn Packages etc..

**4.3** **Hardware Requirements**

Processor : 3.0 GHz and Above

Output Devices : Monitor (LCD)

Input Devices : Keyboard

Hard Disk : 1 TB

RAM : 16GB or Above

Graphics : 2GB or Higher

**CHAPTER 5**

**SYSTEM DESIGN**

**System design** is the phase that bridges the gap between problem domain and the existing system in a manageable way. This phase focuses on the solution domain, i.e. *“how to implement?”*

It is the phase where the SRS document is converted into a format that can be implemented and decides how the system will operate.

In this phase, the complex activity of system development is divided into several smaller sub-activities, which coordinate with each other to achieve the main objective of system development.

System design is the process of defining the elements of a system such as the architecture, modules and components, the different interfaces of those components and the data that goes through that system. It is meant to satisfy specific needs and requirements of a business or organization through the engineering of a coherent and well-running system.

System design gives the following outputs −

* Infrastructure and organizational changes for the proposed system.
* A data schema, often a relational schema.
* Metadata to define the tables/files and columns/data-items.
* A function hierarchy diagram or web page map that graphically describes the program structure.
* Actual or pseudocode for each module in the program.
* A prototype for the proposed system.

**5.1 Architecture Diagram**

It is also known as high level design that focuses on the design of system architecture. It describes the structure and behavior of the system. It defines the structure and relationship between various modules of system development process.

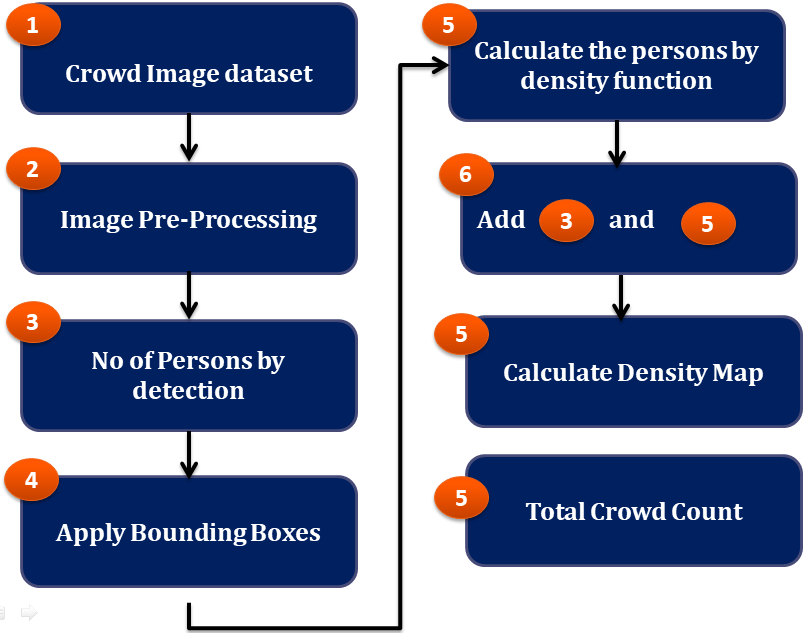


Figure 5.1 Architecture Diagram

5.2 Sequence Diagram

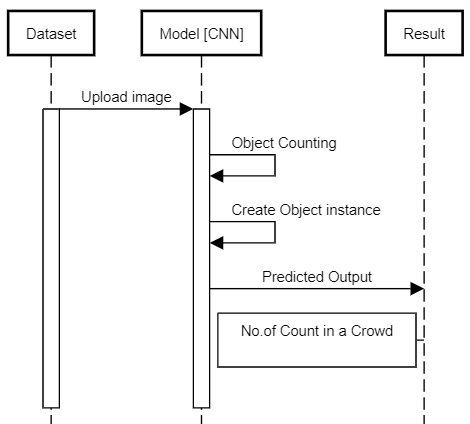
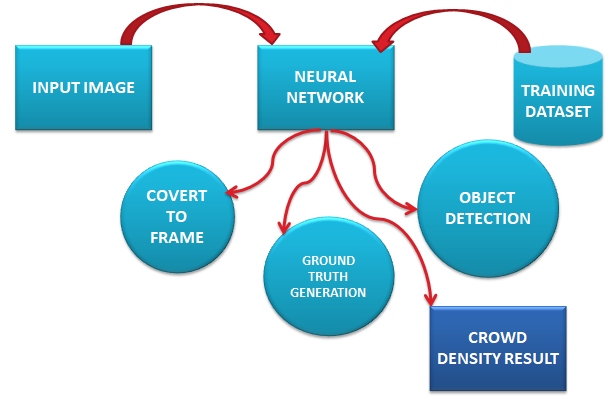
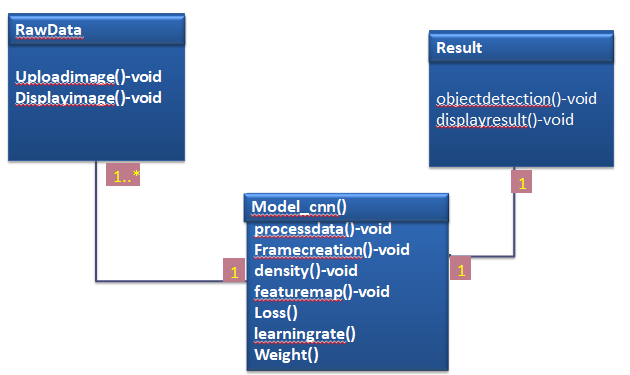


Figure 5.2 Sequence Diagram

5.3 Dataflow Diagram



5.4 Class Diagram



# REFERENCES

# Babu Sam, D., Sajjan, N.N., Venkatesh Babu, R., Srinivasan, M., 2018. Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3618–3626.

# Boominathan, L., Kruthiventi, S.S., Babu, R.V., 2016. Crowdnet: A deep convolutional network for dense crowd counting, in: Proceedings of the 24th ACM international conference on Multimedia, ACM. pp. 640–644.

# Chan, A.B., Liang, Z.S.J., Vasconcelos, N., 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1–7.

# Chan, A.B., Vasconcelos, N., 2009. Bayesian poisson regression for crowd counting, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE. pp. 545–551.

# Collobert, R., Weston, J., 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the 25th international conference on Machine learning, ACM. pp. 160–167.

# Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE. pp. 886– 893.

# Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks, in: Proceedings of the fourteenth international conference on artificial intelligence and statistics, pp. 315–323.

# He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, pp. 1026–1034.

# He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

# Hossain, M., Hosseinzadeh, M., Chanda, O., Wang, Y., 2019. Crowd counting using scale-aware attention networks, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 1280–1288.