Neural Network-Based Crowd Counting Systems: State of the Art, Challenges, and Perspectives

Augustine George, Vinothina V *, and Jasmine Beulah G

Department of Computer Science, Kristu Jayanti College, Bengaluru, India;
Email: augustine@kristujayanti.com (A.G.), jasmine@kristujayanti.com (J.B.G.)
*Correspondence: vinothina.v@kristujayanti.com (V.V.)

***Abstract*—Crowd counting system has gained significant attention in recent years due to its relevance in various domains such as urban planning, public safety, resource allocation and decision-making in crowded environments. Due to differences in crowd densities, occlusions, size changes, and perspective distortions that are frequently seen in real-world scenarios, the system, nevertheless, falls short in terms of its purpose. To address this, it is necessary to create advanced neural network architectures, efficient methods for gathering and annotating data, reliable training procedures, and assessment criteria that accurately reflect the effectiveness of crowd counting systems. Therefore, the purpose of this study is to provide a comprehensive review of the state of the art in neural network-based crowd counting systems. The developments in neural network based crowd counting procedures, along with their features and limitations, most widely datasets and evaluation criteria are explored. The experimental findings of recent crowd counting systems are also examined. Hence, this work serves as an inspiration for additional research and development in this area, ultimately advancing crowd analysis and management systems.**

***Keywords*—deep learning, crowd counting, Convolutional Neural Networks (CNN), scale-aware, transformer, encoder-decoder**

I. Introduction

Overcrowding is the term used to describe when there are too many people for the space and amenities that are available [1]. Public spaces that are filled with people include sports arenas, transportation hubs, retail centers, nightclubs, hotels, places of worship, tourist hotspots, hospitals, and theatres, to name a few. These are but a few instances of places that are typically crowded. Not all congested places are inside buildings; they can also be found in open spaces like parks and pedestrian malls. A high population density also worsens transportation congestion, increases the risk of accidents and illnesses, and increases pollution [2].

Some form of packed public gathering leads to causalities and disaster [3]. More than 100 individuals were murdered and more than 100 others were injured in the stampede in South Korea [4]. The tragedy occurred as a result of the crowding of thousands of people on the small roadway [5]. Crowd counting can be used for a variety of purposes, including security, seating and food arrangements, the distribution of trinkets and presents, and transportation and lodging for various events. Moreover, it is used to organize emergency egress from disasters like fires [6]. Also, by providing signs, it aids in shortening visitor wait times [7].

Video surveillance, traffic monitoring, urban planning, and traffic monitoring are examples of real-world uses for crowd counting [8]. In general, it is very difficult to count the number of attendees at religious events, political rallies, and other public meetings using counters, registration books, or sensor-based technology. Even by hand, it is exceedingly challenging to spot aberrant activity and crowd congestion in busy areas. In Ref. [9], different crowd density estimation approaches and open challenges were discussed. Ahmad *et al.* [10] shown the various person detection approaches for crowd counting in congested areas. The main issues in crowd counting systems are achieving high accuracy and reliability in estimating crowd densities due to environmental factors and scale variations. Reviewing the state-of-the-art methods helps identify the most effective techniques, models, and algorithms that can improve accuracy and reliability, reducing errors and discrepancies in crowd count estimation. Hence, this study contributes the following to the research community to enhance the crowd counting systems.

- Describes the applications and challenges of crowd counting systems.
- Provides the advancements in neural network based crowd counting systems with their features, and limitation
- Assess the performance of existing systems by comparing the experimentation results.

The article is structured into six main sections. Section I is the introduction. Section II is the methodology, which provides the statistics of articles chosen for review. Section III discusses the applications of crowd counting system in various domains. Section IV describes the significant challenges faced by the crowd counting systems. Section V describes the overview of the state-of-the-art of neural network-based crowd counting systems. It includes classification of crowd counting systems based on the structure, along with their features, limitations and

Manuscript received June 1, 2023; revised June 25, 2023; accepted July 10, 2023; published December 26, 2023.

datasets used for experimentation. Section VI provides the evaluation metrics used in the previous study for evaluating the performance of the systems with the comparison of experimental findings. In the final section, we conclude our findings and summarize the main contributions of our work.

II. Methodology

The Web of Science database and google scholar was thoroughly searched for this research starting in 2018. The significant keywords used for this search are "Neural Network based Crowd Counting" that describes the objective of this study. When gathering data for the research, we only included papers that dealt with crowd counting and were written in English. To discover papers relevant to the search, we use the words "Crowd", "Neural Network", and "Crowd Count" in combination. Between the phrases are logical operators. Only articles published in journals and conferences were approved.

Many articles were returned as results. Then, these articles were carefully examined and related articles were chosen for further review. Fig. 1 shows the number of articles and publication year of chosen articles. The articles were categorized based on the neural network structure adopted to achieve the goal.

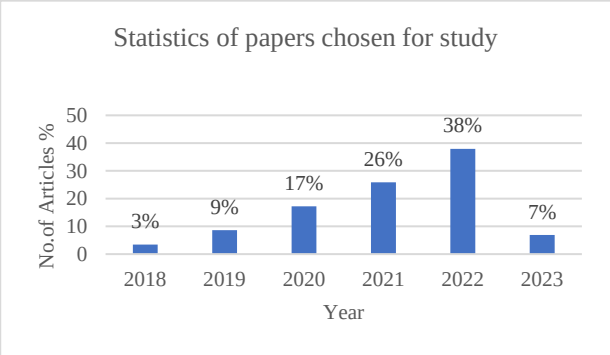


Figure 1. Statistics of articles chosen for study.

III. Crowd Counting Applications

Crowd counting techniques have wide range of applications ranges from pedestrian detection from UAV for crowd flow detection [11, 12], passenger flow detection in exhibition center [13] and bus [14], surveillance System to detect suspicious activities, Security System, crowd analysis to avoid any disaster in public event and traffic management, military applications and health-care applications as in Fig. 2.

Crowd Analysis in Public Places: Crowd-counting methods can be used to collect data for logical analysis and inference. For instance, the length of the line outside a billing reception centre such as electric, gas, and water bills, could be observed and evaluated to appoint the most staff members necessary. In regard to crowd flow, traffic signal wait times could be reduced, especially after work [15, 16].

Health-Care: In Healthcare systems particularly for patients with cancer and other disorders, Convolutional Neural Networks (CNN) based crowd counting is used to count a certain amount of malignant cells for early-stage diagnosis [17]. It can also detects skin cancer from skin images [18] and forecasts liver diseases [19].

Military Applications: Counting drones, fighter jets, enemy personnel, and their weaponry are just a few examples of how CNN-CC techniques can be employed in the military. In order to fight a surge, it was possible to estimate the size of the enemy's armed forces [20, 21].

Public Gathering: In order to count the number of attendees at sporting events, political rallies, religious events, college/school related events, and concerts, crowd counting techniques can be utilized. In order to prevent disastrous scenarios, these events can be handled by assessing and counting the population [22, 23]. Also, this would help in effectively managing resources, such as crowd flow and spatial capacity.

Disaster Management: Huge numbers of people have recently perished from asphyxia in crowded places during various public gatherings. Analyzing the crowd assembling enables earlier overcrowding identification and improved crowd control during political rallies, sporting events, and in other public gatherings [12, 22].

Surveillance: By applying approaches for detecting violent crowd behavior and crowd analysis, terrorist attacks in public areas can be reduced. Conventional handcrafted approaches could be replaced with CNN-based facial identification and detection techniques in video surveillance for better crowd analysis since they perform better in severe and densely populated situations. Safety of people is ensured with the help of crowd analysis [24–27].

Safety Monitoring: At airports, buses, houses of worship, and other public sites, to protect the public's safety and security, crowd counting could be further examined to find patterns of behavior and congestion-time slots [28].

Figure 2. Applications of crowd counting.

IV. Crowd Counting Challenges

The traditional methods are only effective in low density crowds. By comparing the inaccuracy of various classical and neural network based techniques, it can be determined that the latter are more suited to handle crowds with high densities and varying object scales and scene perspectives [29]. Several obstacles limit the accuracy of these networks' counting and the density map's resolution

when using crowd counting. The challenges are described below:

Occlusion: Occlusion happens when two or more items combine when they are very close to one another, making it difficult to distinguish between the distinct objects.

Clutter: An unorganized collection of items that are placed near to one another is known as clutter. It is also connected to picture noise, which makes activities like counting and recognizing more difficult [14].

Object Overlapping: In highly dense image, objects are overlapped with each other. In such case crowd counting would be difficult [30].

Non Uniform Object Scale: Due to difference perspective, the objects in an image can be in difference scale [31]. When viewed from a distance, objects appear smaller, while when viewed up close, they appear larger. Crowd counting becomes difficult in such a situation.

V. Crowd Counting—An Overview

Neural Networks are useful in numerous applications, such as signal processing, image processing, and computer vision. In this regard, various CNN based crowd counting methods were proposed to cope with major issues like occlusion, low visibility, inter- and intra-object variation, and scale variation due to different perspectives [29]. Complex network design, an increase in parameters and high computational cost, and real-time deployment are particular difficulties faced by CNN based crowd counting [32]. But it performs better in terms of prediction accuracy and resolution than traditional methods.

However, having limited receptive fields for context modeling is an intrinsic limitation of these weakly-supervised CNN-based methods. These methods thus cannot achieve satisfactory performance, with limited applications in the real world [33]. CNN inherently fails in modeling the global context due to the limited receptive fields. However, the transformer can model the global context easily [34]. Crowd counting is becoming increasingly popular in computer vision. As a result, we reviewed and analyzed the most current and well-known crowd counting research initiatives on the most difficult datasets. The rising use of deep learning, encoders and decoders, transformers, attention mechanisms and weakly supervised learning in deep learning was the driving force behind this. Table I lists crowd counting models with classification in brief. The classification is based on the neural network structure as follows:

- Deep Neural Network
- Attention-guided Neural network
- Multicolumn Neural Network
- Scale-aware Neural Network
- Transformer Neural Network
- Encoder-Decoder Neural Network

TABLE I. Overview of the Crowd Counting Mode

Category	Model	Limitation	Feature	Dataset
Deep CNN	LR-CNN [13]	heavy computation demands call for quicker hardware support	addresses the issues of crowd distribution, scale feature, and crowd feature extraction difficulty in show centers	UCF_CC_50, ShanghaiTech dataset
	Compact CNN [14]	significant occlusion evident in experimental results	solves the difficulties of counting people from low-resolution images with cluttered background and hazy foreground in a bus	Bus
	CNN-Multiview [36]	-	counts the crowd in large and wide scenes which used multiple cameras with overlapping fields of view	PETS2009, DukeMTMC, crowded street Intersection
	MLP-CNN [37]	Accuracy lacks in very congested locations	Built on VGG16 improves the low recall rate and multiscale headcount	Shanghai Tech and UCF_ CC_ 50 and UCF-QNRF
Attention Guided	CATCNN [40]	ground-truth density map and the calculated density map do not line up	accurately recognizes people in challenging scenarios by paying close attention to the estimated density map's head region	ShanghaiTech Part_A & B, Part_B, UCF_CC_50, WorldExpo'10
	JANet [41]	-	increases the precision of crowd estimation in natural settings with scale variations	ShanghaiTech, UCF-QNRF, JHU-CROWD++, NWPU-Crowd
	AMS-Net [42]	-	enhances crowd count accuracy by paying attention to dense areas and taking into account people viewed from different perspectives	ShanghaiTech, UCF_CC_50, UCF-QNRF, JHU-Crowd++, NWPU Crowd
	MACC Net [43]	-	addresses the problem of crowd density estimation in both sparse and crowded situations	ShanghaiTech, UCF-CC-50, UCF-QRNF, and a recently launched dataset HaCrowd
	DRL [44]	-	utilises a context-aware attention module to estimate the counting using recurrent learning	UCF_QNRF, UCF_CC_50, ShangaiTech_(A, B)
Multicolumn	M-MCNN [45]	-	determine the number of people per square meter, using the population density map's and performs population count	ShanghaiTech Part_A & B, UCSD, WorldExpo'10, GCC
	PRM based [46]	-	obtains more counting accuracy in crowded backgrounds, big perspectives, extreme occlusions, and crowd variation	ShanghaiTech Part_A & B, WorldExpo'10, UCF_QNRF
Scale-aware	L2S [47]	Incorrect predictions on the noisy, very dense, and more than one dense region	investigates long-tailed distribution problems at the pixel level	NWPU-Crowd, JHU-Crowd++, ShanghaiTech Part A, UCF-QNRF

	LSANet [48]	-	achieves high accuracy in real-time crowd counting in complex scenes and enables fast inference in intelligent edge devices	ShanghaiTech, Mall, WorldExpo'10, and UCF_QNRF
	PDD-CNN [49]	-	considers the significant occlusions and appearance differences caused by camera points	ShanghaiTech, UCF_CC_50, and UCF-QNRF
	Multi task framework [50]	Rely on crowd head annotations	address simultaneous crowd-scale changes and background interference issues in still image crowd counting	ShanghaiTech Part_A & B, UCF_CC_50, and UCF-QNRF
	ASANet [51]	-	creates maps of the expected densities by utilizing shared knowledge, can do away with the need for a specialised object detection module	ShanghaiTech, UCF_CC_50, and UCF_QNRF
	MSSRM [52]	-	Plug-in-Plug-Out deals with Low Resolution problems in image or video	ShanghaiTech Part-A, UCF-QNRF
	JCCL [53]	Suffers undercounting in extremely dense crowd scenes	increases counting accuracy and doesn't need separate network components to be pretrained or fine-tuned	ShanghaiTech , UCF-QNRF and JHU-CROWD
Transformer	JCTNet [54]	-	WS transformer represents the overall environment and teaches contrast characteristics for crowd counting	UCF CC 50, ShanghaiTech Part_A & B, UCF-QNRF and NWPU-Crowd datasets
	TransCrowd [33]	video-based counting is not supported	WS model effectively extracts the semantic crowd information using count-level annotations instead of pixel-based	ShanghaiTech Part A and B, UCF-QNRF
	CCTrans [34]	-	get accurate regression results using a transformer, feature aggregation, and regression head that supports both supervised and unsupervised supervision	UCF CC 50, ShanghaiTech Part A and B, UCF-QNRF and NWPU-Crowd datasets
	WSITrans [55]	-	transformer based on WS model obtains head-position for crowd localization and crowd courting	ShanghaiTech Part-A, & B, UCF-QNRF, and NWPU-Crowd
Encoder Decoder	HRANet [56] and AEDNet [57]	-	aims to address the most difficult scale variation and complicated backgrounds in crowd counting issues and crowd localization	ShanghaiTech, UCF-CC-50, UCF-QNRF, WorldExpo'10
	AU-CNN [58]	exhibits a small shortcoming in experimental results	attempts to use the average up-sample module to produce high-quality density maps and precise counting estimation	ShanghaiTech, UCF_CC_50, and UCF_QNRF
	MANet [59]	background elements in images are wrongly identified as pedestrians in a crowd	tackles multiscale and contextual loss issues in crowd counting with a fully convolutional network	ShanghaiTech, WorldExpo'10, UCF_CC_50, and SmartCity

A. Deep Neural Network

At a high level, a CNN consists of an input layer, several convolutional layers, pooling layers, and fully connected layers whereas Deep CNN consists of many layers of convolutional, pooling, and fully connected layers. These networks are designed to learn complex and abstract representations of the input data, enabling them to achieve state-of-the-art performance on a wide range of computer vision tasks, such as image classification, object detection, and segmentation. The depth of the network refers to the number of layers it has, and deep CNNs typically have tens to hundreds of layers. The deeper the network, the more complex the learned representations become, but also the more difficult they are to train due to issues such as vanishing gradients.

Tian and Chu *et al.* [35] proposed a Deep CNN model to enhances crowd-counting accuracy while attempting to address the need for high-performance GPUs for training, subsequent usage, and inference. The research provided a new way to overcome the need of high computing resources based on three important components: feature fusion, Bayesian Loss and datasets utilizing bounding-box annotations to boost the efficiency of the crowd-counting assignment. The Complete Local Binary Pattern (CLBP) is used in this study [13] to derive the properties of crowd aggregation. On the basis of this, the deep learning model is built to identify crowd gathering. Crowd group recognition is done using CNN, and the CLBP feature is learned using techniques like convolution and pooling. The prediction outcome of crowd gathering is obtained after extracting the fundamental features.

By adding up the pixel values from the density map calculated using a compact convolutional neural network [14], which is resistant to scale fluctuations since it uses skip connections, the counting is done. For the purpose of handling cluttered backgrounds and blurry foregrounds, a weighted Euclidean loss is suggested. The loss can reduce the activations in background regions while increasing them in dense regions. Smoothing, which makes use of limitations between subsequent frames, significantly enhances the outcomes of counting.

Zhang and Chan [36] offered a deep neural network architecture for multi-view crowd counting that combines data from several camera angles to forecast a scene-level density map on the 3D world's ground plane. We look at three different implementations of the fusion framework: the late fusion model, which combines camera-view density maps; the nave early fusion model, which combines camera-view feature maps; and the multi-view multi-scale early fusion model, which guarantees that features aligned to the same ground-plane point have uniform scales.

Ren and Lu *et al.* [37] suggested an MLP-CNN model that, when used in conjunction with an FPN feature pyramid, can effectively resolve the issue of an erroneous

head count of multiscale persons by fusing the feature map of low-resolution and high-resolution semantic information. Effective feature fusion of the RGB head image and RGB-Mask picture is possible with the MLP-CNN fusion model. The low recall of head detection can be effectively fixed by creating an upgraded density map using head RGB-Mask annotation and adaptive Gaussian kernel regression.

Alotibi and Jarraya *et al.* [7] designed a DCNN Mobile-based model that reduces customers waiting time and ensures safety and comfort in Low and high crowded Public places in Saudi. It is built on VGG net and takes image of arbitrary size as input. The CNN can accurately measure people anywhere and in a variety of crowd densities. The suggested technique must function properly and effectively both inside and outside of defined environments operates in a smartphone environment with typical levels of processing power, storage, and capacity. Abdou and Erradi [31], Khadka and Argyriou *et al.* [32] also proposed a Deep CNN model that addresses different scale perspectives in complex scenes from video acquisition devices by extracting multiscale features and by providing long contextual information.

The crowd counting proposed in [38] can gradually create density estimation maps of excellent quality via distributed supervision. EDENet is made up of the Feature Extraction Network, Feature Fusion Network, Double-Head Network, and Adaptive Density Fusion Network, specifically. The FEN uses Spatial Adaptive Pooling to extract coarse-grained features and VGG as the network's backbone. The FFN can successfully combine contextual and localization data to improve the ability of fine-grained characteristics to describe space. In the DHN, the Density Attention Module can offer foreground-background attention masks, encouraging the Density Regression Module to pay attention to the pixels around the head annotations when regressing density maps with various resolutions.

He and Xia *et al.* [39] proposed a tiny CNN model called switchable speed CNN to achieve the crowd counting in embedded devices. It is simple to swap between SsCNN_A and the other two modes, SsCNN_B and SsCNN_C, which represent various trade-offs between speed and accuracy. A switching system can function without retraining and exchanges parameters across various modes. The basic idea behind Switch-CNNs is to use multiple convolutional filters with different kernel sizes or shapes and then dynamically switch between these filters based on the input data. This allows the network to effectively capture features of different sizes and shapes, without having to rely on fixed filter sizes or complicated pooling operations.

B. Attention-Guided Neural Network

An attention-guided neural network is a type of neural network architecture that incorporates an attention mechanism. The attention mechanism is a technique used in deep learning to allow the network to focus on specific parts of an input sequence that are relevant to the task at hand. It can be thought of as a way to selectively weight different parts of the input when making a prediction, rather than treating all parts equally. This can be particularly useful in tasks such as machine translation or image captioning, where the relevant information may be spread out over a large input sequence.

In an attention-guided neural network, the attention mechanism is typically incorporated as a separate module within the network, which takes as input the hidden states of the network and computes attention weights that are used to weight the input features. These weighted features are then used to make predictions. Attention Modules proposed in [36, 37, 40] allows the model to selectively focus on the most relevant parts of the input data when making predictions. The main function of an attention component is to enhance the performance of the model by dynamically weighting the importance of different parts of the input data. When compared to complete body representations, head regions were employed by Zhang and Wang *et al.* [40] to separate people from one another in congested surroundings, even in low quality and low resolution conditions.

Jointly Attention Network (JANet) proposed by Aldhaheri *et al.* [41] contains attention module called module scale attention investigates significant high-order statistics and aids the backbone network in explicitly obtaining more discriminative features with rich scale information.

Gong and Bourennane *et al.* [43] proposed Multi-task Attention-based Crowd Counting Network (MACCNet) that comprises of Density level classification, which provides the density estimation network with global contextual information, Density map estimation and segmentation guided attention to separate the foreground features from the background noise. By automatically encoding a confidence map, CAT-CNN [40] can adaptively evaluate the significance of a human head at each pixel position. The position of the human head in the estimated density map is given more consideration when encoding the final density map under the direction of the confidence map, which can effectively prevent enormous misjudgments. The finished density map can be integrated to determine the population density.

Context-Aware pyramid attention module [60] is designed to learn contextual information by incorporating both local and global features of the image. The multi-label classification module uses the learned features to predict the labels for the input image. The counting performance is enhanced by an attention module created in CAPAN that deals with the interdependence on feature information in the spatial dimension and the channel dimension. Due to extreme occlusions, perspective distortion, and wildly varying crowd densities, the photos provide a number of difficulties in estimating the crowd size [31]. As the network is trained based on the dataset, the results may be ineffective for another dataset.

C. Multicolumn Neural Network

A multicolumn neural network consists of multiple parallel columns of neural networks, each column processing a different type of input or feature representation. The output of each column is then combined in some way to produce a final output. The idea behind multicolumn neural networks is to exploit the fact

that different types of input data may be best represented using different feature representations. By processing each type of input data separately in its own column, the network can learn to extract the most informative features from each type of data.

Gong *et al.* [45] proposed a Multi-feature Multi-column CNN that first extracts features, learn the importance of features and then the 3-column sub network in M-MCNN captures the features of different sizes of human heads. Finally, the population density map is obtained by assigning the weight linearly to the feature map. Three deep-layered branches make up the proposed PRM-based model [46], each of which produces feature maps with a different resolution. These branches combine their features at the feature level to create the necessary collective knowledge for the final crowd estimate, which improves comprehension of the foreground area.

D. Scale-Aware Neural Network

Scale-aware neural networks are a type of deep neural network architecture that is designed to be sensitive to the scale of the input data. These networks are particularly useful when working with images that have objects at different These networks use multiple filters with different sizes in each layer, which enables them to detect objects at different scales. One common type of scale-aware neural network is the pyramid network, which consists of multiple branches that process the input image at different resolutions. The outputs of these branches are then combined to generate a final prediction. Another type is the feature pyramid network, which uses a similar approach to generate a multi-scale feature map that is used for object detection and recognition.

The Learning to Scale Module proposed in [47] scales dense regions into acceptable closeness levels automatically. In order to dynamically separate the overlapped blobs and decompose the accumulated values in the ground-truth density map, L2S directly normalizes the closeness in various patches. This reduces pattern shifts and the long-tailed distribution of density values. This makes it easier for the model to learn the density map.

By effectively leveraging scale information, LSANet [48] attempted to achieve a better lightweight balance between accuracy and efficiency for things like network characteristics and processing resources. The Scale feature extraction modules are precisely engineered to retrieve multi-scale information at varied dilation rates. The Scale feature fusion modules are made up of three efficient attention-based fusion blocks, aggregates the feature map of the hybrid dilated convolution block. In order to produce precise density maps, the density map regressor will receive the output feature maps of the SFFM.

Pyramid Diluted Deep (PDD)-CNN [49] is able to create dense attribute feature maps from images of any size or resolution. Then, adopted two pyramid dilated modules, each of which consists of four parallel, dilated convolutional layers running at a separate rate and a parallel average pooling layer to capture the multiscale data. In order to accurately estimate the count, three cascade dilated convolutions are employed to regress the density map. The Pyramid Dilated Module uses dilated convolution at various rates to combine multiscale convolution features and manage the significant appearance fluctuations in a single image.

To differentiate crowds from complex backgrounds, the crowd head edge regression module [50] produces discrete crowd head edge features. Relative depth map regression task detects differences in crowd scale and produces multi-scale crowd characteristics and density map regression moule generates high quality density map. Jiang and Lin *et al.* [51] proposed an adversarial scale-adaptive neural network for crowd counting comprising of three branches. First branch focuses on generating high-quality density maps, second branch detect and recognize objects and third acquire the similar attention area and support crowd counting.

To overcome the overlapping head regions and lost details in the low resolution images, Wang and Liu *et al.* [52] proposed a method called Multi-Scale Super-Resolution Module which directs the network to estimate the lost features without inference cost. To evaluate the method, SR-crowd dataset is introduced that contains both Low Resolution (LR) and High Resolution (HR) labeled images. When the network is being trained, the MSSRM directs the network to predict accurate details even when given LR images. The detail loss brought on by hazy images might be made up for during the inference stage, without altering the original network topology. The authors used MSSRM to propose a Multi-Scale Super-Resolution Guided Network for LR circumstances.

Gao and Zhao *et al.* [53] proposed a framework for joint crowd counting and localization to increases counting accuracy. The scale adaptive module in JCCL addresses the large scale variation. This module used to control the pooling and upsampling in the network which is based on the adaptive dilated convolution (ADC). The ADC enables learning of location-specific receptive fields and adjustable dilation rate. JCCL doesn't need separate network components to be pretrained or fine-tuned.

E. Encoder-Decoder Based Neural Network

An encoder-decoder neural network consists of two parts: an encoder and a decoder. The encoder is a series of convolutional or recurrent layers that process the input data and create a condensed representation of the data and extract meaningful features from the input data and map it to a lower dimensional representation. The decoder, on the other hand, takes the latent space representation created by the encoder and generates the output. The decoder's goal is to reconstruct the input data or create a new output that is related to the input data.

Xie and Gu *et al.* [54] proposed Hierarchical Region Aware net can more effectively concentrate on crowded areas to predict crowd density. The Region Aware Module in HRANet allows adaptive extraction of contextual information inside various regions. Region Recalibration Module recalibrates the feature weights of various areas using a brand-new region-aware attention mechanism. The influence of background regions can be successfully suppressed through the integration of the two modules mentioned above and improved the counting accuracy.

Xia and He *et al.* [59] proposed an innovative encoder-decoder structure for crowd counting. The Feature Extraction Encoder in the network transforms high-dimensional input data into a lower-dimensional representation that captures the most relevant information or features of the input data. Generally, an encoder in neural network is a component that transforms input data into a compressed representation that is easier to process. Density Map Decoder transforms a low-dimensional input into a high-dimensional density map output. The main function of a density map decoder is to reconstruct a high-resolution representation of the input data from a lower-resolution or compressed representation.

For high-quality density maps and precise counting estimation, Huang and Sinnott *et al.* [58] proposed an encoder-decoder structure network called Average Up-sample Convolution Neural Network (AU-CNN). The decoder gradually restores the size of the feature map to the original size of the input picture while the encoder extracts the features from the input image using a straightforward but efficient average up-sample module.

Attentive Feature Refinement [57] block in the encoder to adaptively extract multi-scale features. The decoder's Non-local Fusion block aggregates multi-scale information from several layers at lower computation costs using self-attention mechanism. Spatial Attention Module [61] learns to selectively focus on specific spatial regions of an image by assigning different weights to different spatial locations. The goal of SAM is to enhance the representation of informative regions of the image while suppressing the influence of irrelevant or noisy regions. The attention mechanism in SAM learns to assign weights to the features of the input image based on their relevance to the task at hand.

F. Transformer Based Neural Network

The transformer architecture consists of an encoder and a decoder, both of which use self-attention layers. In the encoder, the input sequence is first embedded into a high-dimensional space, and then multiple self-attention layers are applied to the embedded sequence. The output of the encoder is then passed to the decoder, which uses a similar architecture to generate the output sequence. One key feature of the transformer architecture is that it can process the entire input sequence in parallel, making it more efficient than RNNs. Additionally, it can capture long-term dependencies in the input sequence without being affected by the vanishing gradient problem, which often occurs in RNNs.

Crowd counting using weakly supervised learning via CNN typically cannot demonstrate good performance because CNN is not adequate for modelling the global context and the interactions between image patches. To model the overall environment and teach contrast characteristics, the weakly supervised model via Transformer was proposed in [33, 54]. The network's parameter number is rather huge, and the transformer directly divides the crowd photos into a collection of tokens, which may not be the best option given that each pedestrian is a separate individual. Transformer [54] can process input sequences in parallel, allowing it to handle longer sequences more efficiently. The self-attention mechanism of the Transformer allows it to weigh the importance of different parts of the input sequence when generating outputs. This attention mechanism makes it possible for the Transformer to capture long-range dependencies in the input data, which can be important for many NLP tasks.

Liang and Chen *et al.* [34] proposed a straightforward pathway for crowd counting in contexts that are both weakly and fully supervised. The Pyramid Vision Transformer architecture also introduces multi-scale feature fusion and attention-based spatial sampling, to better capture both local and global information in the input image.

One of the main differences is that the transformer architecture uses self-attention layers, while the encoder-decoder architecture typically uses recurrent or convolutional layers. Self-attention allows the transformer to attend to different parts of the input sequence when generating each output, while recurrent or convolutional layers process the input sequence sequentially. Another key difference is that the transformer can process the entire input sequence in parallel, while the encoder-decoder architecture processes the input sequence sequentially. This makes the transformer faster and more efficient than the encoder-decoder architecture.

G. Significant Modules in Crowd Counting System

This section describes the some of the other significant modules proposed in the crowd counting system to improve the counting performance. The Channel Attention Module (CAM) designed in [61] operates on the channel dimension of the feature map, generating a vector of weights that are applied to each channel. These weights indicate the importance of each channel in the feature map, and higher weights indicate more informative channels. CAM learns to selectively amplify informative channels of a feature map.

Guo and Gao *et al.* [62] attempted to apply sequential crowd counting as a physical process in reality in the form of LibraNet. The scale-weighing agent learns to set the right weights to match the count by placing a crowd image on the scale. At each step, a weight is selected from the weight box based on the image attributes and the weights placed, and this process continues until the pointer indicates balance. Different learning paradigms, such as Deep Q-Network (DQN), Actor-Critic (AC), Imitation Learning (IL), and combined AC+IL, are used to create LibraNet.

In order to enhance the head/non-head classification, a Dual-Path guided Detection Network is introduced in [63], which makes use of a density map. The density suggests the likelihood that a pixel is a head, and a depth-adaptive kernel that creates a high-fidelity density map takes into account the variations in head sizes. A density map is used for post-processing of head detection in order to prevent dense heads from being filtered out. Additionally, a depth-aware anchor is built for improved initialization of anchor sizes in the detection framework, along with a depth-guided detection module that creates a dynamic dilated convolution to extract characteristics of heads of various scales.

Considering that the lack of diverse training samples and imbalanced distribution across different classes in crowd scenes inevitably result in large prediction deviation caused by the DL model Liu et.al proposed a method called Pseudo-label Growth Dictionary Pair Learning [64] to improve the counting accuracy by applying pseudo label growth and adaptive dictionary size.

Liu and Wang *et al.* [65] proposed an online knowledge learning method for crowd counting. By creating a new inter-layer relationship matrix, the feature relation distillation method helps the student branch better understand how inter-layer characteristics have evolved. In order to improve the transfer of knowledge that is mutually beneficial from the instructor branch to the student branch, it is combined with response distillation and feature internal distillation. This method reduces the size and computational complexity of crowd counting using a knowledge distillation process. Knowledge Distillation Module transfers the knowledge from a large and complex model such as teacher model to a smaller and more efficient model known as the student model. It can also be used to transfer knowledge from an ensemble of models to a single model, or from a model trained on one task to a model trained on a different but related task.

VI. Performance Evaluation

The performance of the proposed crowd counting systems was evaluated on the benchmark dataset with evaluation metrics. Datasets are a crucial necessity for evaluating the proposed design of vision processing systems. The datasets utilized by earlier publications to evaluate the performance of the proposed models are shown in Fig. 3.

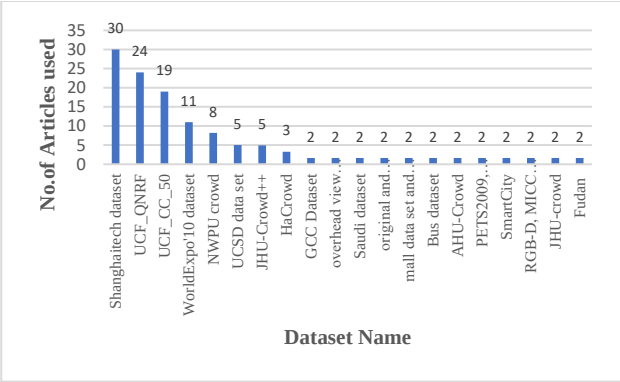


Figure 3. Datasets used by the articles chosen for overview.

Based on the review, Mean Absolute Error (MAE) and Root Mean Squared (RMSE) error were mostly used to evaluate the performance of the model whose definitions are as follows [65] :

$$\text{MAE} = \frac{1}{M}\sum_{i=1}^{M}|c_i - \hat{c}_i| \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}|C_i - \hat{C}_i|^2} \quad (2)$$

In addition to MAE and RMSE, Normalized Absolute Error (NAE) is also used to evaluate the model [41]. It is defined as follows:

$$\text{NAE} = \frac{1}{M}\sum_{i=1}^{M}\frac{|c_i - \hat{c}_i|}{c_i} \quad (3)$$

The number of test image is represented by M and predicted crowd counts and ground truth crowd counts are represented by the variable C_i and $\hat{C}_i$ respectively. These metrics evaluates the accuracy and robustness of the model.

Another metric mean error is proposed in [57] to measure the variance of ground truths and it is defined as follows:

$$\text{ME} = \frac{1}{N}\sum_{i=1}^{N}\left(C_i^{GT} - C_i^{GT'}\right) \quad (4)$$

where *N* denotes the number of images, and represents the sum of the ground truth generated and the sum of down sampled ground truth respectively.

But few research used different criteria based on the model designed and task. The parameters are Precision, Recall, mean Average Precision [25], Spatial Computational Complexity [19], scaling branches [23], absolute counting errors based on LoS Blockage Detection Threshold and counting window length [29, 30] and inference speed to name a few.

In the recent research [45, 47], Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity in Image (SSIM) [48] were used as evaluation criteria. PSNR is used to calculate the error between the corresponding pixels, SSIM is used to measure the structural similarity between the ground-truth and the estimated density map. The following Tables II–IV compare the results of the related works according to the used datasets and the used metrics. The corresponding graph shown in Figs. 4–6.

TABLE II. Comparisons of Found Experimental Results on Shanghai Tech Dataset

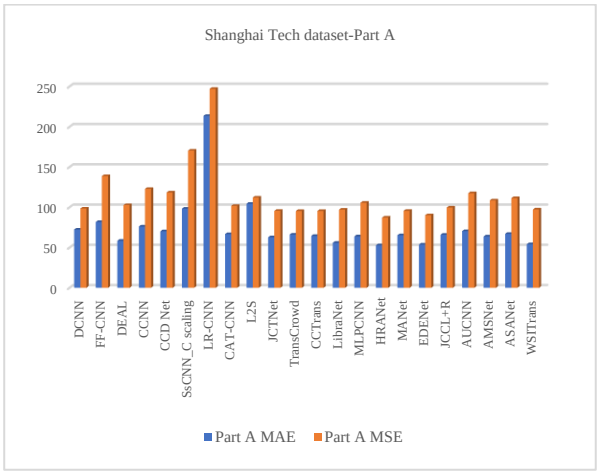
Methods	Part A		Part B	
	MAE	MSE	MAE	MSE
DCNN [7]	72.3	98.4	11.1	17.3
FF-CNN [66]	81.75	138.8	16.45	26.19
DEAL [50]	58.4	102.7	6.8	11.1
CCNN [14]	76.1	122.8	20	31.6
CCD Net [65]	70.2	118.38	-	-
SsCNN_C scaling [39]	98.3	170.4	18.4	30.8
Switch CNN [67]	90.4	135	21.6	33.4
LR-CNN [13]	213.5	247.1	85.3	99.7
CAT-CNN [40]	66.7	101.7	11.2	20
L2S [47]	104.4	112.1	8.6	13.9
JCTNet [54]	62.8	95.6	7.2	11.5
TransCrowd [33]	66.1	95.4	9.3	16.1
CCTrans [34]	64.4	95.4	7	11.5
LibraNet [62]	55.9	97.1	7.3	11.3
MLPCNN [37]	63.9	105.6	9.58	12.59
HRANet [56]	52.8	87.2	6.2	9.7
MANet [59]	65.31	95.54	10.2	16.5
EDENet [38]	53.7	90.1	6.6	11.3
JCCL+R [53]	65.9	99.7	8.2	14.8
AUCNN [58]	70.4	117.5	8.6	13
AMSNet [42]	63.8	108.5	7.3	11.8
ASANet [51]	67	111.4	7.6	11.7
WSITrans [55]	54.1	97.3	7.1	9.9

TABLE III. COMPARISON OF FOUND EXPERIMENTAL RESULTS ON UCF_CC_50

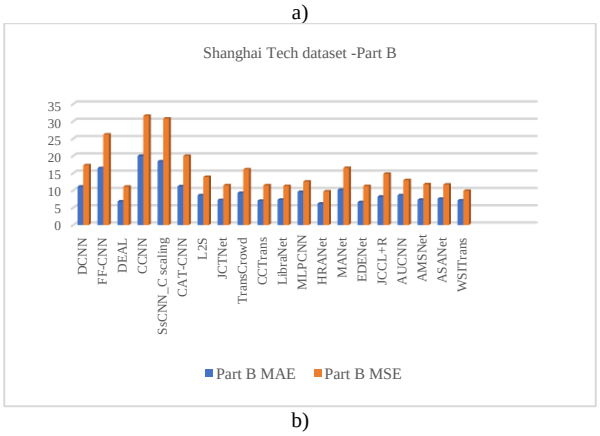
Methods	MAE	MSE
DEAL [50]	194.1	297.8
Switch CNN [67]	318.1	439.2
LR-CNN [13]	325.6	369.4
CAT-CNN [40]	235.5	324.8
CCTrans [34]	245.0	343.6
JCTNet [54]	222.9	306.5
LibraNet [62]	181.2	262.2
MLPCNN [37]	238.63	317.28
HRANet [56]	160.9	235.8
MANet [59]	240.8	311.5
AUCNN [58]	231.8	312.4
AMSNet [42]	236.5	319.2
ASANet [51]	185.5	268.3

TABLE IV. COMPARISON OF FOUND EXPERIMENTAL RESULTS ON UCF_QNRF

Methods	MAE	MSE
CCD Net[65]	136.26	240.83
DEAL [50]	100.1	166.5
JCTNet [37]	90	161
TransCrowd [33]	97	168
CCTrans [34]	92	158
LibraNet [62]	88.1	143.7
MLPCNN [37]	103.61	168.9
HRANet [56]	84.6	146.2
EDENet [38]	86.6	158.5
JCCL+R [53]	100	169
AUCNN [58]	112.3	195.6
AMSNet [42]	86.5	167.2
WSITrans [55]	86.5	140.3



a)



b)

Figure 4. Comparison of found experimental results on Shanghai Tech dataset. a) Part A; b) Part B.

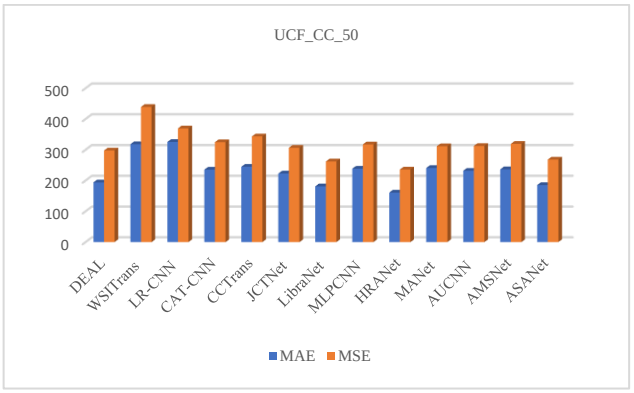


Figure 5. Comparison of found experimental results on UCF_CC_50.

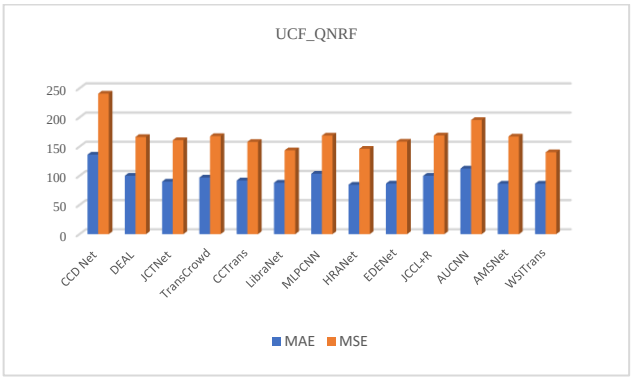


Figure 6. Comparison of found experimental results on UCF_QNRF.

From Tables II–IV and its corresponding graph Figs. 2–4, it is evident that transformer based models are outperforming than other models on most widely datasets like shanghai Tech dataset, UCF_CC_50 and UCF_QNRF. These datasets have dense and diverse images of crowd. Compared to traditional methods, transformer based model is more effective in terms of computing time as it could process the input sequence in parallel.

VII. CONCLUSION

The demand for crowd counting in numerous fields has significantly increased research in crowd counting in recent years. The performance of crowd counting models has significantly improved with the development of deep learning, and the possibilities for real-world applications have increased. This paper provided an overview of current developments in crowd counting from the perspective of network design, challenges, applications, performance evaluation criteria and most often used datasets. Since the application of crowd counting growing in number, lot of researchers focuses on this domain. This review would be very useful for the further research as crowd counting still requires enhancement to overcome the challenges of large scale variations, real-time constraints, wide-scenes, overlapped human heads, occlusions and generalization of weakly supervised labeled dataset. Moreover, as per advanced search, there is 10% increase in the number of articles published in 2022 compared to 2021. This shows not only the interest of the domain but also the challenges continued in crowd counting.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Vinothina designed the research plan and organized the study; coordinated the data analyzed and contributed to the written of the manuscript. Augustine George and Jasmine Beulah provided assistance and support at every stage of the manuscript preparation process. All authors had approved the final version.

ACKNOWLEDGMENT

We appreciate Kristu Jayanti College Autonomous' cooperation and assistance in helping us conduct our research.

REFERENCES

[1] R. Brynen. (2013). Social and psychological effects of overcrowding. *Palestinian Refugee ResearchNet*. [Online]. Available: https://prrn.mcgill.ca/research/papers/marshy.htm

[2] A. Johansson, M. Batty, K. Hayashi, O. Al Bar, D. Marcozzi, and Z. A. Memish, "Crowd and environmental management during mass gatherings," *The Lancet Infectious Diseases*, vol. 12, no. 2, pp. 150–156, Feb. 2012. doi: 10.1016/S1473-3099(11)70287-0

[3] K. Still, "Crowd dynamics," PhD thesis, University of Warwick, 2000.

[4] South Korea Halloween stampede | Chocolates, flowers and soju as Seoul mourns its dead. (2022). The Hindu, Seoul. [Online]. Available: https://www.thehindu.com/news/international/south-korea-halloween-stampede-chocolates-flowers-and-soju-as-seoul-mourns-its-dead/article66107415.ece

[5] Halloween Horror in South Korea: What Led to The Death of More Than 120 in Seoul | Explained. [Online]. Available: https://www.india.com/explainer/halloween-horror-in-south-korea-what-led-to-the-death-of-more-than-120-in-seoul-explained-5712648/

[6] U. Bhangale, S. Patil, V. Vishwanath, P. Thakker, A. Bansode, and D. Navandhar, "Near real-time crowd counting using deep learning approach," *Procedia Computer Science*, vol. 171, pp. 770–779, 2020. doi: 10.1016/j.procs.2020.04.084

[7] M. H. Alotibi, S. K. Jarraya, M. S. Ali, and K. Moria, "CNN-based crowd counting through IoT: Application for Saudi public places," *Procedia Computer Science*, vol. 163, pp. 134–144, Jan. 2019. doi: 10.1016/j.procs.2019.12.095

[8] R. Gouiaa, M. A. Akhloufi, and M. Shahbazi, "Advances in convolution neural networks based crowd counting and density estimation," *Big Data and Cognitive Computing*, vol. 5, no. 4, Dec. 2021. doi: 10.3390/bdcc5040050

[9] S. A. Alshaya, "Open challenges for crowd density estimation," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 1, 2020. doi: 10.14569/IJACSA.2020.0110123

[10] M. Ahmad, I. Ahmed, K. Ullah, I. khan, A. Khattak, and A. Adnan, "Person detection from overhead view: A survey," *International Journal of Advanced Computer Science and Applications* (*IJACSA)*, vol. 10, no. 4, 2019. doi: 10.14569/IJACSA.2019.0100470

[11] Z. Hu, W. Cao, F. Wu, Z. Zhang, C. Dong, and Y. Qu, "A real-time UAV crowd counting system based on edge computing," in *Proc. 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP)*, Oct. 2021, pp. 1–5. doi: 10.1109/WCSP52459.2021.9613445

[12] C. Martani, S. Stent, S. Acikgoz, K. Soga, D. Bain, and Y. Jin, "Pedestrian monitoring techniques for crowd-flow prediction," in *Proc. the Institution of Civil Engineers, Smart Infrastructure and Construction*, Jun. 2017, vol. 170, pp. 1–11. doi: 10.1680/jsmic.17.00001

[13] J. Xiang and N. Liu, "Crowd density estimation method using deep learning for passenger flow detection system in exhibition center," *Scientific Programming*, vol. 2022, e1990951, Feb. 2022. doi: 10.1155/2022/1990951

[14] B. Yang, J. Cao, X. Liu, N. Wang, and J. Lv, "Edge computing-based real-time passenger counting using a compact convolutional neural network," *Neural Comput & Applic*, vol. 32, no. 9, pp. 4919–4931, May 2020. doi: 10.1007/s00521-018-3894-2

[15] X. Wu, G. Liang, K. K. Lee, and Y. Xu, "Crowd density estimation using texture analysis and learning," in *Proc. 2006 IEEE International Conference on Robotics and Biomimetics*, 2006, pp. 214–219.

[16] M. C. Mongeon, R. P. Loce, and M. A. Shreve. (2017). Busyness defection and notification method and system. Patent US9576371B2. [Online]. Available: https://patents.google.com/patent/US9576371/en

[17] B. Dong, L. Shao, M. Da Costa, O. Bandmann, and A. F. Frangi, "Deep learning for automatic cell detection in wide-field microscopy zebrafish images," in *Proc. 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2015, pp. 772–776. doi: 10.1109/ISBI.2015.7163986

[18] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017. doi: 10.1038/nature21056

[19] P. V. Nayantara, S. Kamath, K. N. Manjunath, and K. V. Rajagopal, "Computer-aided diagnosis of liver lesions using CT images: A systematic review," *Comput. Biol. Med.*, vol. 127, 104035, Dec. 2020. doi: 10.1016/j.compbiomed.2020.104035

[20] A. Albert, J. Kaur, and M. Gonzalez, "Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale." arXiv prerint, arXiv 1704.02965, 2017.

[21] B. Kellenberger, D. Marcos, and D. Tuia, "Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning," *Remote Sensing of Environment*, vol. 216, pp. 139–153, Oct. 2018. doi: 10.1016/j.rse.2018.06.028

[22] M. N. Kamel Boulos *et al.*, "Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples," *International Journal of Health Geographics*, vol. 10, no. 1, p. 67, Dec. 2011. doi: 10.1186/1476-072X-10-67

[23] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, Apr. 2015. doi: 10.1109/TITS.2014.2345663

[24] M. Khouj, C. López, S. Sarkaria, and J. Marti, "Disaster management in real time simulation using machine learning," in *Proc. 2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE)*, May 2011, pp. 001507–001510. doi: 10.1109/CCECE.2011.6030716

[25] J. R. Barr, K. W. Bowyer, and P. J. Flynn, "The effectiveness of face detection algorithms in unconstrained crowd scenes," in *Proc. IEEE Winter Conference on Applications of Computer Vision*, Mar. 2014, pp. 1020–1027. doi: 10.1109/WACV.2014.6835992

[26] S. Chackravarthy, S. Schmitt, and L. Yang, "Intelligent crime anomaly detection in smart cities using deep learning," in *Proc. 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, Oct. 2018, pp. 399–404. doi: 10.1109/CIC.2018.00060

[27] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 443–449. doi: 10.1145/2818346.2830593

[28] H. Song, X. Liu, X. Zhang, and J. Hu, "Real-time monitoring for crowd counting using video surveillance and GIS," in *Proc. 2012 2nd International Conference on Remote Sensing, Environment and Transportation Engineering*, Jun. 2012, pp. 1–4. doi: 10.1109/RSETE.2012.6260673

[29] B. Zhou, X. Tang, and X. Wang, "Learning collective crowd behaviors with dynamic pedestrian-agents," *International Journal of Computer Vision*, vol. 111, pp. 50–68, Jan. 2014. doi: 10.1007/s11263-014-0735-3

[30] S. Thasveen and L. Mredhula, "Real time crowd counting: A review," in *Proc. 2020 International Conference on Futuristic Technologies in Control Systems & Renewable Energy (ICFCR)*, Sep. 2020, pp. 1–5. doi: 10.1109/ICFCR50903.2020.9249984

[31] M. Abdou and A. Erradi, "Crowd counting: A survey of machine learning approaches," in *Proc. 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, Feb. 2020, pp. 48–54. doi: 10.1109/ICIoT48696.2020.9089594

[32] A. Khadka, V. Argyriou, and P. Remagnino, "Accurate deep net crowd counting for smart IoT video acquisition devices," in *Proc. 2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, May 2020, pp. 260–264. doi: 10.1109/DCOSS49796.2020.00049

[33] N. Ilyas, A. Shahzad, and K. Kim, "Convolutional-neural network-based image crowd counting: Review, Categorization, analysis, and performance evaluation," *Sensors*, vol. 20, no. 43, 2019. doi: 10.3390/s20010043

[34] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, "TransCrowd: Weakly-supervised crowd counting with transformers," *Sci. China Inf. Sci.*, vol. 65, no. 6, 160104, Apr. 2022. doi: 10.1007/s11432-021-3445-y

[35] Y. Tian, X. Chu, and H. Wang, "CCTrans: Simplifying and improving crowd counting with transformer," arXiv preprint, arXiv 2109.14483, 2021.

[36] Q. Zhang and A. B. Chan, "Wide-area crowd counting: Multi-view fusion networks for counting in large scenes," *Int J Comput Vis*, vol. 130, no. 8, pp. 1938–1960, Aug. 2022. doi: 10.1007/s11263-022-01626-4

[37] G. Ren, X. Lu, J. Wang, and Y. Li, "Enhancement of local crowd location and count: Multiscale counting guided by head RGB-mask," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–19, Aug. 2022. doi: 10.1155/2022/5708807

[38] J. Chen, W. Su, and Z. Wang, "Crowd counting with crowd attention convolutional neural network," *Neurocomputing*, vol. 382, pp. 210–220, Mar. 2020. doi: 10.1016/j.neucom.2019.11.064

[39] Y. He, Y. Xia, Y. Wang, and B. Yin, "Jointly attention network for crowd counting," *Neurocomputing*, vol. 487, pp. 157–171, May 2022. doi: 10.1016/j.neucom.2022.02.060

[40] B. Zhang, N. Wang, Z. Zhao, A. Abraham, and H. Liu, "Crowd counting based on attention-guided multi-scale fusion networks," *Neurocomputing*, vol. 451, pp. 12–24, Sep. 2021. doi: 10.1016/j.neucom.2021.04.045

[41] S. Aldhaheri *et al.*, "MACC Net: Multi-task attention crowd counting network," *Appl. Intell.*, Aug. 2022. doi: 10.1007/s10489-022-03954-x

[42] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, K. Abualsaud, A. Mohamed, and T. Khattab, "Crowd counting using DRL-based segmentation and RL-based density estimation," in *Proc. 2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Nov. 2022, pp. 1–8. doi: 10.1109/AVSS56176.2022.9959690

[43] S. Gong, E.-B. Bourennane, and J. Gao, "Multi-feature counting of dense crowd image based on multi-column convolutional neural network," in *Proc. 2020 5th International Conference on Computer and Communication Systems (ICCCS)*, May 2020, pp. 215–219. doi: 10.1109/ICCCS49078.2020.9118564

[44] U. Sajid and G. Wang, "Towards more effective PRM-based crowd counting via a multi-resolution fusion and attention network," *Neurocomputing*, vol. 474, pp. 13–24, Feb. 2022. doi: 10.1016/j.neucom.2021.12.027

[45] C. Xu *et al.*, "AutoScale: Learning to scale for crowd counting," *Int J. Comput. Vis.*, vol. 130, no. 2, pp. 405–434, Feb. 2022. doi: 10.1007/s11263-021-01542-z

[46] F. Zhu, H. Yan, X. Chen, and T. Li, "Real-time crowd counting via lightweight scale-aware network," *Neurocomputing*, vol. 472, pp. 54–67, Feb. 2022. doi: 10.1016/j.neucom.2021.11.099

[47] W. Wang, Q. Liu, and W. Wang, "Pyramid-dilated deep convolutional neural network for crowd counting," *Appl. Intell.*, vol. 52, no. 2, pp. 1825–1837, Jan. 2022. doi: 10.1007/s10489-021-02537-6

[48] S. Peng, B. Yin, X. Hao, Q. Yang, A. Kumar, and L. Wang, "Depth and edge auxiliary learning for still image crowd density estimation," *Pattern Anal. Applic.*, vol. 24, no. 4, pp. 1777–1792, Nov. 2021. doi: 10.1007/s10044-021-01017-4

[49] X. Chen, H. Yan, T. Li, J. Xu, and F. Zhu, "Adversarial scale-adaptive neural network for crowd counting," *Neurocomputing*, vol. 450, pp. 14–24, Aug. 2021. doi: 10.1016/j.neucom.2021.03.128

[50] J. Xie *et al.*, "Super-resolution information enhancement for crowd counting," arXiv preprint, arXiv 2303.06925, 2023.

[51] M. Jiang, J. Lin, and Z. J. Wang, "A smartly simple way for joint crowd counting and localization," *Neurocomputing*, vol. 459, pp. 35–43, Oct. 2021. doi: 10.1016/j.neucom.2021.06.055

[52] F. Wang, K. Liu, F. Long, N. Sang, X. Xia, and J. Sang, "Joint CNN and transformer network via weakly supervised learning for efficient crowd counting," arXiv preprint, arXiv 2203.06388, 2022.

[53] H. Gao, W. Zhao, D. Zhang, and M. Deng, "Application of improved transformer based on weakly supervised in crowd localization and crowd counting," *Sci. Rep.*, vol. 13, no. 1, Jan. 2023. doi: 10.1038/s41598-022-27299-0

[54] J. Xie, L. Gu, Z. Li, and L. Lyu, "HRANet: Hierarchical region-aware network for crowd counting," *Appl. Intell.*, vol. 52, no. 11, pp. 12191–12205, Sep. 2022. doi: 10.1007/s10489-021-03030-w

[55] X. Liu, Y. Hu, B. Zhang, X. Zhen, X. Luo, and X. Cao, "Attentive encoder-decoder networks for crowd counting," *Neurocomputing*, vol. 490, pp. 246–257, Jun. 2022. doi: 10.1016/j.neucom.2021.11.087

[56] D. Wu, Z. Fan, and M. Cui, "Average up-sample network for crowd counting," *Appl. Intell.*, vol. 52, no. 2, pp. 1376–1388, Jan. 2022. doi: 10.1007/s10489-021-02470-8

[57] P. Li, M. Zhang, J. Wan, and M. Jiang, "Multiscale aggregate networks with dense connections for crowd counting," *Computational Intelligence and Neuroscience*, vol. 2021, e9996232, Nov. 2021. doi: 10.1155/2021/9996232

[58] Z. Huang, R. Sinnott, and Q. Ke, "Crowd counting using deep learning in edge devices," in *Proc. 2021 IEEE/ACM 8th International Conference on Big Data Computing, Applications and Technologies (BDCAT '21)*, New York, NY, USA, Jan. 2022, pp. 28–37. doi: 10.1145/3492324.3494161

[59] Y. Xia, Y. He, S. Peng, X. Hao, Q. Yang, and B. Yin, "EDENet: Elaborate density estimation network for crowd counting," *Neurocomputing*, vol. 459, pp. 108–121, Oct. 2021. doi: 10.1016/j.neucom.2021.06.086

[60] J. Chen, Q. Zhang, W.-S. Zheng, and X. Xie, "Efficient and switchable CNN for crowd counting based on embedded terminal," *IEEE Access*, vol. 7, pp. 51533–51541, 2019. doi: 10.1109/ACCESS.2019.2910458

[61] L. Gu, C. Pang, Y. Zheng, C. Lyu, and L. Lyu, "Context-aware pyramid attention network for crowd counting," *Appl. Intell.*, vol. 52, no. 6, pp. 6164–6180, Apr. 2022. doi: 10.1007/s10489-021-02639-1

[62] X. Guo, M. Gao, W. Zhai, J. Shang, and Q. Li, "Spatial-frequency attention network for crowd counting," *Big Data*, vol. 10, no. 5, pp. 453–465, Oct. 2022. doi: 10.1089/big.2022.0039

[63] L. Liu, H. Lu, H. Zou, H. Xiong, Z. Cao, and C. Shen, "Weighing counts: Sequential crowd counting by reinforcement learning," in *Proc. European Conference on Computer Vision (ECCV 2020)*, 2020, pp. 164–181. doi: 10.1007/978-3-030-58607-2_10

[64] D. Lian, X. Chen, J. Li, W. Luo, and S. Gao, "Locating and counting heads in crowds with a depth prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9056–9072, Dec. 2022. doi: 10.1109/TPAMI.2021.3124956

[65] W. Liu, H. Wang, H. Luo, K. Zhang, J. Lu, and Z. Xiong, "Pseudo-label growth dictionary pair learning for crowd counting," *Appl. Intell.*, vol. 51, no. 12, pp. 8913–8927, Dec. 2021. doi: 10.1007/s10489-021-02274-w

[66] S. Jiang, B. Li, F. Cheng, and Q. Liu, "Crowd Counting with online knowledge learning," arXiv preprint, arXiv 2303.10318, 2023.

[67] H. Luo *et al.*, "A high-density crowd counting method based on convolutional feature fusion," *Applied Sciences*, vol. 8, no. 12, Dec. 2018. doi: 10.3390/app8122367

[68] A. Olugboja, Z. Wang, and Y. Sun, "Parallel convolutional neural networks for object detection," *Journal of Advances in Information Technology*, vol. 12, no. 4, pp. 279–286, November 2021. doi: 10.12720/jait.12.4.279-286