Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.058637

REVIEW

A Survey on Supervised, Unsupervised, and Semi-Supervised Approaches in Crowd Counting

Jianyong Wang[1], Mingliang Gao[1], Qilei Li[2], Hyunbum Kim[3] and Gwanggil Jeon[3,*]

[1]School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, China

[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, UK

[3]Department of Embedded Systems Engineering, Incheon National University, Incheon, 22012, Republic of Korea

*Corresponding Author: Gwanggil Jeon. Email: ggjeon@gmail.com

Received: 17 September 2024 Accepted: 19 November 2024 Published: 19 December 2024

ABSTRACT

Quantifying the number of individuals in images or videos to estimate crowd density is a challenging yet crucial task with significant implications for fields such as urban planning and public safety. Crowd counting has attracted considerable attention in the field of computer vision, leading to the development of numerous advanced models and methodologies. These approaches vary in terms of supervision techniques, network architectures, and model complexity. Currently, most crowd counting methods rely on fully supervised learning, which has proven to be effective. However, this approach presents challenges in real-world scenarios, where labeled data and ground-truth annotations are often scarce. As a result, there is an increasing need to explore unsupervised and semi-supervised methods to effectively address crowd counting tasks in practical applications. This paper offers a comprehensive review of crowd counting models, with a particular focus on semi-supervised and unsupervised approaches based on their supervision paradigms. We summarize and critically analyze the key methods in these two categories, highlighting their strengths and limitations. Furthermore, we provide a comparative analysis of prominent crowd counting methods using widely adopted benchmark datasets. We believe that this survey will offer valuable insights and guide future advancements in crowd counting technology.

KEYWORDS

Crowd counting; density estimation; convolutional neural network (CNN); un/semi-supervised learning

1 Introduction

In recent decades, the problem of object counting has garnered increasing attention within research communities, becoming a central research direction. Consequently, extensive studies have been conducted to count objects in image processing across a broad spectrum of fields, including crowd counting [1–5], cell microscopy [6–9], animal populations [10], vehicle counts [11–15], foliage [16,17], and environmental monitoring [18,19]. Crowd counting is particularly important in these fields, as it plays an extremely important role in tasks such as crowd analysis [5,20,21] and video surveillance [22]. The rapid growth of the global population and urbanization has led to more frequent gatherings of people at certain events. In such scenarios, crowd counting is essential for maintaining social safety and

effective crowd management. Additionally, by reviewing recent crowd counting studies, we noted that Khan et al. [23] summarized many significant contributions in the field. By filtering out less innovative or underperforming approaches, they refined the recent advancements in crowd counting. Following this, Khan et al. [24] further explored the domain of visual crowd analysis, covering six key areas of crowd vision analysis and providing an overview of the current state-of-the-art methods in crowd analysis.

Crowd counting approaches are typically categorized into four types: detection method [25–27], regression method [28–30], density estimation method [31,32], and the emerging Convolutional Neural Network (CNN)-based density estimation techniques [33,34]. This paper centers on the analysis of CNN-based density estimation methods for crowd counting, which are gaining increasing recognition for their effectiveness.

Early approaches to crowd counting [35–38] relied on detection methods, often using a sliding window technique to locate individuals or heads within images. With the introduction of advanced object methods [39–43], detection accuracy has improved significantly in sparse scenarios. Nevertheless, these methods perform poorly in addressing occlusion issues in high-density crowd scenarios. Detection-based approaches also face challenges, including high computational complexity, reliance on complex post-processing steps, and difficulty handling non-rectangular targets.

To address the aforementioned issues, several studies [22,44,45] have introduced regression approaches that directly use the mapping from image to crowd counting. These methods generally begin by extracting global coarse-grained features [46], or local fine-grained features [47] like Scale Invariant Feature Transform (SIFT) [48], Local Binary Patterns (LBP) [49], Histograms of Oriented Gradient (HOG) [50]. Regression models, including linear regression [51] and Gaussian mixture regression [52], were applied to map these features to crowd counting.

While these methods manage occlusion and background clutter effectively, they often fail to account for spatial information. To solve these issues, Lemptisky et al. [8] introduced a framework for supervised learning, which enhances counting accuracy by learning the correlation between image features and density maps. To further optimize the framework, [53] proposed a random forest regression method. This technique, which incorporates a crowdedness prior, allows for the training of two separate forests with reduced memory consumption for forest storage. Despite these advantages, these methods still depend on traditional hand-crafted features, which fall short of capturing the low-level details necessary for creating high-quality density maps, thereby limiting their effectiveness of crowd counting.

CNNs have attracted significant attention in computer vision due to their powerful feature extraction capabilities. For instance, in intelligent transportation systems, Ren et al. [54] utilized CNNs along with Interactive Multiple Model (IMM) filters to accomplish multi-object tracking. The IMM filter, as a multi-modal filtering method, uses multiple filter models simultaneously to represent possible target movement patterns and improves tracking accuracy through interactions among these models. In the field of crowd counting, CNNs have led researchers to adopt them for enhancing density estimation. The original methods in this section of the study primarily used basic CNN models for population density estimation [7,55–57], resulting in a significant improvement compared to traditional hand-crafted features. These models differ in their supervision levels and learning paradigms, with some specifically designed for cross-scene and multi-domain use. Fig. 1 provides an overview of the major developments and milestones in crowd counting technology.

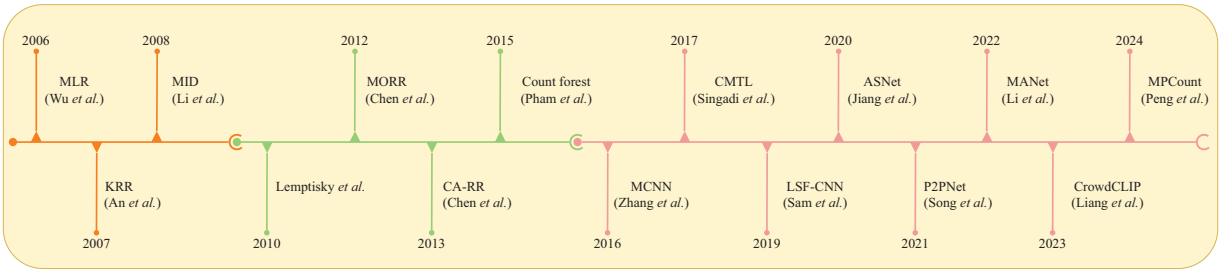

Figure 1: A brief history of crowd counting reveals that the dominant research directions have been crowd counting models based on detection (orange), regression (green), and density estimation (red). These directions have shaped both current research and potential future developments. Milestone models in this figure: MLR [58], KRR [59], MID [36], Lemptisky et al. [8], MORR [46], CA-RR [60], Count forest [53], MCNN [29], CMTL [61], LSF-CNN [62], ASNet [63], P2PNet [64], MANet [65], CrowdCLIP [66] and MPCount [67]

A review of previous work [68–71] revealed that all these methods require image labeling during training. Detection-based approaches demand full identification and outlining of each object, leading to the highest labeling costs. In contrast, regression-based approaches only necessitate labeling the total object count, which results in the lowest annotation costs. Density estimation strikes a balance between the two, requiring only the heads of individuals to be labeled, thus incurring moderate costs. But in extremely high-density crowd scenes, automatic crowd counting annotation techniques struggle to achieve accurate labeling, often showing significant discrepancies from actual data. Manual annotation in such densely populated scenes is also highly time-consuming and labor-intensive, as annotators must meticulously identify individual people within the crowd. This is undeniably a technically challenging and costly task. Additionally, in modern urban development, certain areas require long-term monitoring and counting tasks, necessitating the acquisition of large amounts of continuous data. It is impractical for annotators to label frames over extended periods manually. In such cases, training crowd counting networks with fully labeled data is unrealistic. Therefore, relying on unsupervised or semi-supervised methods for counting becomes particularly important. The focus of this survey is on crowd counting using unsupervision and semi-supervision. The advent of deep learning technology in computer vision has greatly enhanced counting accuracy and accelerated the study of weakly supervised and semi-supervised crowd counting methods. Fig. 2 outlines a crowd counting model based on deep learning methods. First, the crowd image is input into the designed counting network, which extracts the image features and generates the corresponding density map. The core of this process is the design of a counting network that calculates the number of people in the input image by integrating the density values of the pixels. For network training, supervised learning is applied, where labeled images are used to establish the ground truth, indicating the number of people in each image. Finally, the loss function adjusts the network parameters, minimizing the discrepancy between the generated density map and the ground truth.

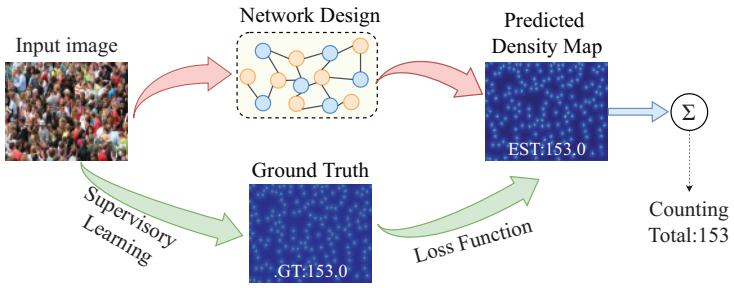

Figure 2: A brief deep learning crowd counting framework

2 Crowd Counting Datasets and Evaluation Protocols

As crowd counting has evolved, numerous datasets have been introduced, inspiring the development of various methods to tackle challenges such as scale change, background interference, and lighting changes. In this section, we provide an overview of several major crowd counting datasets, spanning from the early stages of the field to the present day. The primary evaluation protocols for crowd counting are detailed below. Table 1 summarizes some representative datasets.

Table 1: Some mainstream crowd counting datasets

Dataset	Total count	Images	Resolution
UCSD [22]	–	2000	238 × 158
Mall [46]	62,325	2000	640 × 480
UCF_CC_50 [44]	–	50	Vary
WorldExpo'10 [57]	199,923	3920	576 × 720
ShanghaiTech-A [29]	241,677	482	Vary
ShanghaiTech-B [29]	88,488	716	768 × 1024
UCF-QNRF [32]	1,251,642	1535	Vary
JHU-Crowd [72]	25,791	4372	1450 × 900
NWPU Crowd [73]	2,133,238	5109	Vary

UCSD [22] dataset is the first dataset specifically applied to crowd counting and was developed by the University of California, San Diego (UCSD). The dataset consists of video sequences captured by fixed cameras on a sidewalk at a university campus. The crowd density in the videos varies from sparse to dense, adding a certain level of complexity to the scene. Each frame in the video sequence has a resolution of 238 × 158 pixels, and each frame includes manually annotated locations and counts of pedestrians. This labeling information is widely used to train and evaluate crowd counting models.

Mall [46] dataset is a benchmark dataset used for crowd counting and behavior analysis. It was collected from surveillance footage in a shopping mall and includes a video sequence recorded by a fixed camera. The video sequence comprises 2000 frames, each with a resolution of 640 × 480 pixels and a frame rate of 2 frames per second. The Mall dataset is highly representative due to the specific lighting conditions and complex background disturbances in the mall environment, making it a combination of challenging and practical applications.

UCF_CC_50 [44], published by the University of Central Florida, is a challenging benchmark dataset specifically tailored for crowd counting and density estimation. The dataset includes 50 high-resolution images that span a variety of complex scenes. The crowd density in these images varies widely, from a few dozen people to several thousand, with extremely high density in some extreme scenes. Due to the high density and diversity of the images, UCF_CC_50 is regarded as one of the most challenging baseline datasets in crowd counting. Since the dataset contains only 50 images, researchers often use k-fold cross-validation methods to make the most of the dataset for training and evaluation.

WorldExpo'10 [57] is a widely recognized benchmark for crowd counting and density estimation research. Released by the Chinese University of Hong Kong, this dataset comprises 108 short video sequences captured at the 2010 Shanghai World Expo, featuring a variety of indoor and outdoor scenes. The scenes represent different crowd activity areas, including plazas, entrances, and exhibition halls. The videos have a resolution of 576×720 pixels. The dataset is divided into a training set with 3380 frames and a test set with 600 frames. Each test scenario involves different video sequences, each presenting distinct challenges.

ShanghaiTech [29] dataset, released by Shanghai University of Science and Technology, is divided into two parts: Part A and Part B, tailored for different density scenarios. Part A contains high-density crowd images primarily sourced from the internet. This section comprises 300 training images and 182 test images, featuring very dense crowd scenes, such as protests and parties, with complex backgrounds and significant occlusions. Part B focuses on images of low-density crowds collected mainly from the streets of Shanghai. It includes 400 training images and 316 test images, depicting everyday scenes with lower crowd density and simpler backgrounds.

UCF-QNRF [32] is a hyperscale dataset focused on crowd counting and density estimation. Known for its high-resolution images, extremely high crowd density, and diverse scenes, this dataset is widely utilized in crowd counting research. It stands as one of the most challenging benchmark datasets available. The UCF-QNRF dataset contains 1535 images representing a variety of complex scenes, including streets, festivals, religious gatherings, and sports events. The images are of very high resolution, with some reaching thousands of pixels. Crowd density ranges from a few dozen people in sparse settings to several thousand in highly crowded scenes.

JHU-CROWD [72] dataset, published by Johns Hopkins University, is a large and complex dataset for crowd counting. It contains 4372 high-resolution images that cover a wide range of crowd scenes, from sparse to extremely dense. These scenes include city streets, public gatherings, religious events, sporting events, and more, representing the diverse distribution of people in real-world environments. The dataset features a variety of image resolutions, from low to high, to ensure a broad coverage of visual conditions.

NWPU-Crowd [73] dataset is a dataset focused on crowd counting and density estimation, published by Northwestern Polytechnical University (NWPU) in China. It consists of 5109 high-resolution images that span a wide range of crowd scenes, from sparse to extremely dense. The dataset includes various environments, such as urban streets, business districts, religious gatherings, sports events, and transportation hubs, showcasing the complexity of real-world population distribution. The image resolutions are up to thousands of pixels, ensuring that the head position of each pedestrian can be clearly identified even in dense scenes.

Evaluation Protocols: In assessing the experiment's performance, most methods employ the mean absolute error (MAE) and the root mean square error (RMSE). The specific formula they propose is

as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|, \tag{1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|^2}, \tag{2}$$

where N represents the number of test images. y_i and $\hat{y}_i$ represent the predicted and ground truth values for the i-th image, respectively.

3 Weakly and Semi-Supervised Learning Model

Weak supervision is a technique for training crowd counting models under conditions of insufficient supervision. Compared to traditional fully supervised methods, weak supervision allows for effective model training and better counting performance, even with limited data or incomplete information. In crowd counting, fully labeling the position of each pedestrian's head is often both time-consuming and costly. The weak supervision method reduces the labeling burden by using partial labels or other forms of weak annotation. Such networks have found widespread use in tasks like object detection, segmentation, and classification, where dense annotations are hard to obtain or expensive. Although weakly supervised networks have certain limitations, their application in crowd counting holds significant promise, especially when labeled data is costly or limited. With ongoing advancements in algorithms and computing power, the potential of weak supervision methods in practical applications will be further explored and realized.

Since the advancement of crowd counting, the use of weakly labeled data for crowd counting [74–78] has emerged steadily. These approaches effectively extract image features from noisy annotations [79,80]. Traditional crowd counting networks based on density map estimation demand precise image annotation, which is both time-consuming and expensive. Hierarchical Attention-based Crowd Counting Network (HA-CCN) [81] generates pseudo-density maps from image-level labels, effectively reducing the cost of manual annotation and offering an efficient weakly supervised approach. By fusing feature maps from different convolutional layers, it addresses the scale variation challenge in crowd counting. The integration of multi-scale feature maps with an attention module significantly improves counting accuracy and generalization. However, the accuracy of these pseudo-density maps significantly impacts the effectiveness of weak supervision, with dense and complex scenes potentially introducing substantial errors. Further optimization of pseudo-label generation is necessary.

Traditional semi-supervised crowd counting models struggle with regression tasks due to biases introduced by focusing solely on a single prediction objective. As shown in Fig. 3, the proposal of Dual-Goal Generative Adversarial Networks (DG-GAN) [82] addresses the bias issue in traditional Generative Adversarial Networks (GANs) by introducing two independent objectives: "regression prediction" and "real/fake" classification. Specifically, the discriminator produces two outputs: one predicting the crowd count (regression) and another performing binary classification of the input as "real" or "generated". This approach broadens the applications of GANs and helps reduce annotation costs in dense crowd counting tasks. However, the GAN training process can lead to model instability or mode collapse, which, though not discussed in detail in the paper, may impact model performance in practical applications. Semi-supervised Regression Generative Adversarial Networks (SR-GAN) [83] further optimizes the generator by adjusting the generated images to align their feature distributions

more closely with that of unlabeled data, thereby improving the discriminator's regression prediction accuracy. This approach significantly reduces the reliance on labeled data during training. Additionally, SR-GAN improves model accuracy and robustness by minimizing the differences between real and generated data features. However, feature contrast and feature matching require multiple rounds of adversarial training, which increases computational costs, particularly in high-resolution images and complex feature tasks where computational resources are highly demanded.

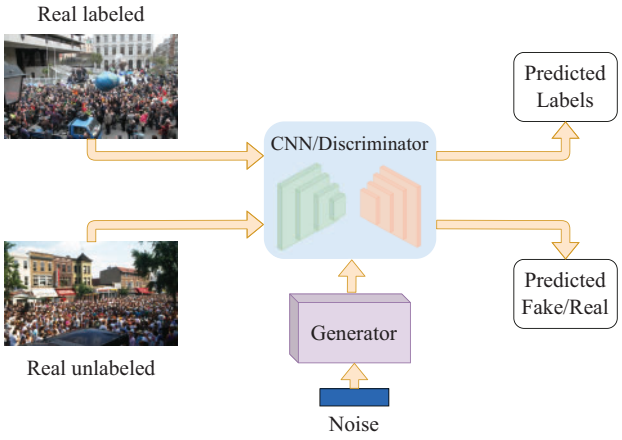

Figure 3: A brief crowd counting semi-supervised network architecture

As semi-supervised crowd counting tasks continue to evolve, several effective methods [84–87] have successfully reduced the need for costly annotations. We have summarized several notable semi-supervised methods developed in recent years, and their counting performance is presented in Table 2. Lei et al. [84] utilized a small amount of fully supervised location-annotated data along with a larger quantity of count-labeled data. They employed the Multiple Auxiliary Tasks Training (MATT) strategy to add multiple auxiliary branches, each generating distinct but equivalent density maps to prevent the generation of unconstrained density maps with count-level annotations alone. This approach significantly reduces annotation costs. However, the instability and complexity of auxiliary tasks may lead to convergence issues during training. Meng et al. [74] utilized a teacher-student model framework, where the teacher model generates spatial uncertainty maps to guide the student model's feature learning, thereby reducing the impact of noise in unlabeled data. This approach employs a well-designed unsupervised consistency mechanism using hard/soft uncertainty maps, enhancing the model's resilience to high-noise unlabeled data. While uncertainty awareness aids in noise suppression, misestimations may still occur in boundary regions of ultra-dense scenes, affecting the final counting accuracy.

Although semi-supervised models reduce the need for expensive annotations in crowd counting, they still encounter challenges in dense crowd scenes, particularly with occlusion, accuracy, and robustness. Khan et al. [23] introduced a spatial uncertainty-aware semi-supervised approach. This method uses a teacher-student framework with spatial uncertainty awareness to address noise in unlabeled data, leveraging an auxiliary binary segmentation task in the teacher model to estimate spatial uncertainty maps. These maps guide density regression and feature learning in the student model. A difference transformation layer enhances spatial consistency between tasks, generating high-quality uncertainty maps and addressing task-level perturbations. Wei et al. [88] handled varying density representations by training multiple models, each focusing on a distinct density distribution but

consistent in total count. Count consistency across models was used to supervise unlabeled data, with kernel mean embedding providing an implicit density representation that avoids strong parametric assumptions. Qian et al. [89] used masking on unlabeled data, guiding the model to predict masked regions based on overall context clues. A fine-grained density classification task enhanced feature learning, strengthening the model's ability to estimate counts in low-density regions and predict high-density regions accurately. Miao et al. [90] extended the traditional smoothness assumption to a many-to-many regional feature smoothness framework to address uneven density distributions. Hypergraph representation was used to capture complex relationships between crowd regions, with a multi-scale dynamic hypergraph convolution module and hyper-edge contrastive loss boosting the model's capacity to handle occlusion. In conclusion, semi-supervised models reduce dependency on labeled data, enhance generalization, and improve counting accuracy and robustness in dense scenes. However, they still face challenges such as increased model complexity, pseudo-label accuracy issues, and dependency on unlabeled data quality.

Semi-supervised crowd counting methods estimate the number of pedestrians in active scenes by combining a limited number of labeled frames with a substantial set of unlabeled frames. This approach is crucial for intelligent transportation systems, as it supports better emergency response planning and reduces congestion risks at transportation hubs such as train stations and airports [91]. Additionally, by implementing passive radio frequency (RF) sensor networks, crowd sizes of thousands can be accurately monitored across various environments, making it an invaluable asset for safety management in large-scale events [92].

4 Unsupervised and Self-Supervised Learning Model

The convergence speed of unsupervised and supervised learning differs significantly, depending on the specific algorithm and application scenario. Supervised learning, utilizing labeled data, generally converges more rapidly to a local optimum. For example, a crowd density estimation model for drones (DroneNet) [93], employs Self-organized Operational Neural Networks (Self-ONN) based on a Multi-column Convolutional Neural Network (MCNN) architecture, replacing convolutional layers with Self-ONN layers. To achieve faster convergence, it does not rely on transfer learning. The Lightweight Crowd Density estimation model (LCDnet) [94], another lightweight model, reduces model complexity through the use of smaller matrix filters and adopts a curriculum learning (CL) approach to enhance convergence. The Curriculum Learning with Iterative data Pruning (CLIP) [95] model improves convergence speed by iteratively pruning the least relevant samples, gradually shrinking the dataset size used during curriculum learning.

Compared to supervised learning, unsupervised learning typically has a slower convergence rate due to the lack of labeled data. For example, the batch Expectation Maximization (EM) algorithm converges slowly, but online EM algorithms (such as incremental EM and stepwise EM) can significantly accelerate convergence [96]. However, the convergence speed of unsupervised learning is also influenced by the learning rate setting, which, if improperly configured, may slow down or hinder convergence [97]. Notably, certain unsupervised learning methods can achieve faster convergence under specific conditions. For instance, deep semi-supervised learning methods theoretically converge faster than multi-parameter supervised learning [98]. Additionally, some self-supervised learning methods improve training efficiency and convergence speed by introducing masking strategies [99]. In summary, supervised learning generally converges more quickly to a local optimum with the support of labeled data, while unsupervised learning takes longer to uncover hidden patterns or structures in the

data. However, under specific conditions, unsupervised methods can also achieve faster convergence by optimizing algorithms and strategies.

The majority of crowd counting methods necessitate a mass of labeled data for training. While crowd counting requires costly labeling, unlabeled data are widely available and inexpensive [100,101]. As shown in Fig. 4, Liu et al. [102] leveraged the property that sub-images of a crowd image contain a number of people less than or equal to the total in the parent image. They segmented unlabeled crowd images into a series of progressively smaller sub-images, training the network with a ranking loss function from learning to rank. This approach aids the network in learning effective feature representations from unlabeled data. However, the ranking task relies on the assumption of a consistent relationship between the crowd counts of parents and sub-images. If significant density variations or occlusions occur in the image, this may lead to inaccurate ranking.

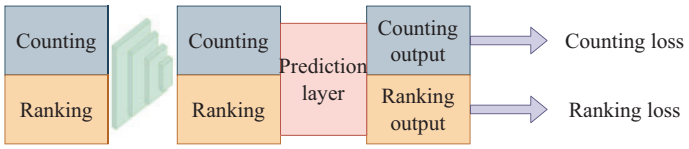

Figure 4: A brief self-supervised learning model

There is a growing body of work aimed at developing more efficient crowd counting networks using unlabeled data, and it often relies on unsupervised learning to obtain the necessary network parameters. As shown in Fig. 5, Sam et al. [103] extended the traditional autoencoder for unsupervised crowd counting tasks. Unlike standard autoencoders, their approach divides each convolutional feature map into predefined grid blocks, allowing only one neuron in each block to implement a "Grid Winner-Take-All" mechanism. By learning features through Grid Winner-Take-All (GWTA), the model can capture local patterns without label supervision, significantly reducing dependency on labeled data. However, the learned features may not always be optimal for task-specific objectives and may underperform in ultra-dense scenes. However, although these self-supervised or unsupervised methods are effective in crowd counting, they require significant data resources and considerable time to train the network. Several advanced methods for unsupervised crowd counting are outlined in Table 2.

With the development of unsupervised crowd counting tasks, numerous effective methods have been proposed. Cross-domain unsupervised crowd counting aims to enhance the generalization capability of crowd counting models across different domains or scenarios. Due to significant variations between images in different scenes as viewing angles, illumination, and density distribution-models trained in one domain (e.g., a specific scene) often perform poorly when applied to another distinct domain. Cross-domain unsupervised crowd counting addresses this problem by enhancing model performance on unlabeled or sparsely labeled target domain data. Liu et al. [104] treated regression and detection models as two distinct knowledge sources suitable for high-density and low-density regions, respectively. They proposed two conversion modules, "Det-to-Reg" and "Reg-to-Det" These modules generate pseudo-labels, which are then used to fine-tune the regression and detection models in the target domain. However, noise in the pseudo-labels can impact the final model performance, particularly in high-density or heavily occluded scenes. The development of domain adaptation techniques has also facilitated the application of unsupervised learning methods. For example, by using style-level transfer learning and scene-aware estimation [105], better performance can be achieved in crowd counting across different styles and scenes. Additionally, some studies have introduced domain-invariant feature extraction modules and the dynamic-β Model-Agnostic Meta-Learning (MAML)

algorithm [106], effectively reducing domain gaps and generating more refined density maps, leading to improved performance in cross-domain crowd counting scenarios. Although these methods leverage domain adaptation techniques to improve the model's generalization ability across different crowd scenarios, a significant drawback remains: the model's accuracy tends to decrease sharply when handling high-density crowds. Compared to unsupervised learning, domain adaptation techniques in supervised learning for crowd counting also focus on transferring models trained on the source domain (labeled datasets) to the target domain (unlabeled or sparsely labeled datasets). The key difference is that supervised learning can leverage ample labeled data in the source domain, enabling the model to achieve higher counting accuracy.

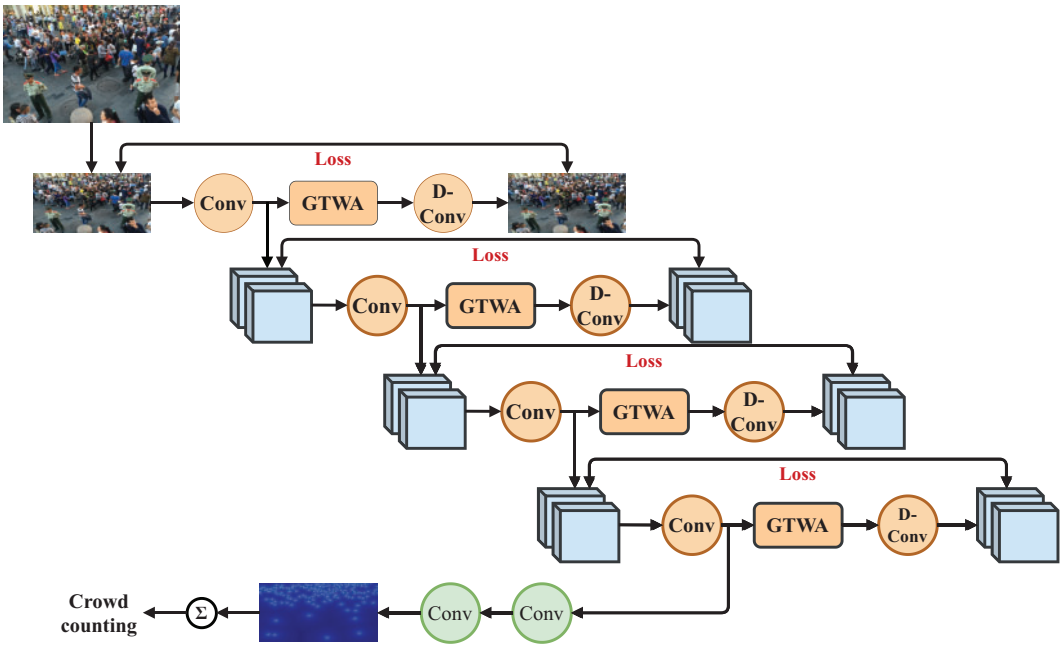

Figure 5: Grid winner-Take-all based crowd counting CNN (GWTA-CCNN): A brief unsupervised learning model

Diffusion Models, a form of generative model, have achieved impressive results in image generation and unsupervised learning tasks in recent years. The key concept behind diffusion models is to progressively add noise to the data and then learn the reverse process to reconstruct the original data from the noise. By training the diffusion model to generate density maps that reflect the actual crowd distribution, the model can capture the overall distribution structure of crowds from unlabeled data. D'Alessandro et al. [107] used a stable diffusion model to generate images containing a specific number of pedestrians. However, there is often a discrepancy between the actual number of people in the generated images and the preset target, creating "noisy" labels. These image pairs form a weak ranking signal, allowing the model to learn crowd counting features during pretraining. However, due to inaccuracies in the generated noisy count data, the model may experience errors in high density scenarios. The application of diffusion models in unsupervised crowd counting tasks shows potential, but it faces several major challenges, including high computational complexity, long training times, reliance on large-scale data, insufficient accuracy in high-density scenarios, and the lack of clear supervision signals. Additionally, evaluating and optimizing the generated results of diffusion models is challenging. Future research can explore how to improve the efficiency and generation accuracy

of diffusion models, as well as how to better integrate them with unsupervised learning methods to address these challenges in practical applications.

In recent years, with the successful application of large models (such as Contrastive Language-Image Pretraining (CLIP) [108] and Bootstrapping Language-Image Pre-training (BLIP) [109]) in visual and verbal tasks, researchers have increasingly explored the application of these models to unsupervised crowd counting tasks. One significant advantage of large models lies in their pre-training on extensive datasets, which enables them to capture diverse feature information and perform effectively across a wide range of tasks. Liang et al. [66] explored visual language integration to address crowd counting challenges, applying the contrastive pre-trained visual language model (CLIP) to unsupervised crowd counting. During training, CrowdCLIP learns from image encoders by aligning crowd patches with corresponding text prompts. This approach effectively addresses the challenge of insufficient image information learning in unsupervised settings, leading to successful unsupervised crowd counting performance. However, the model's use of a multi-stage filtering strategy and ranking-based contrastive training with the CLIP model requires significant computational resources and time, which may make it unsuitable for low-power devices.

Crowd counting has seen broad application in urban planning and public safety. Although fully supervised crowd counting methods have achieved considerable progress, they are highly dependent on a large amount of labeled crowd density maps, which is especially time-consuming and costly in high-density crowds. In contrast, unsupervised crowd counting research holds the potential for more effective real-world public safety applications. Unsupervised anomaly detection algorithms can identify specific abnormal behaviors, such as crowd gathering and dispersal, based on counting data, employing foreground segmentation algorithms and latent energy models for global crowd counting [110]. This facilitates the timely identification of crowd dynamics that may pose safety risks. UrbanCount [111] introduced a fully distributed crowd counting protocol based on communication among mobile devices to conduct crowd estimation. This protocol is designed to achieve accurate local estimates in high-density urban environments, ensuring alignment with global crowd counts while preserving node privacy.

5 Comparison of State-of-the-Art (SOTA) Methods for Crowd Counting

Table 2 presents a selection of current SOTA methods for crowd counting. These methods are categorized into four main groups based on their supervision type: Fully Supervised (FS), Semi-Supervised (SS), Unsupervised (US), and Cross-Domain Supervised (CS). Additionally, we provide the model parameter counts (Params) and inference times (Time) for each method. All models were evaluated for efficiency on an RTX 3090 Ti, with the input image size standardized to 224×224. Most methods lacking efficiency parameters have either unpublished or incomplete code, which hinders our ability to verify their model efficiencies. The data in the table indicates that semi-supervised and unsupervised methods generally have fewer model parameters compared to fully supervised approaches, while their inference times are comparable to most fully supervised counting methods. Overall, although semi-supervised and unsupervised models exhibit some accuracy gaps compared to their fully supervised counterparts, their efficiency is largely equivalent, and in some cases, even superior. This suggests that the proposed semi-supervised and unsupervised methods are viable for practical deployment.

Table 2: Performance and efficiency comparison of various crowd counting models across different datasets and supervisory modes. "**FS**" stands for Fully Supervised methods, "**SS**" for Semi-Supervised methods, "**US**" for Unsupervised methods, and "**CS**" for Cross-Domain Supervised methods

Mode	Method	Part A		Part B		UCF_CC_50		UCF-QNRF		JHU++		NWPU		Time (ms)	Params (M)
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE		
FS	DLPTNet [112]	58.4	95.0	9.3	15.6	–	–	121.0	225.8	77.7	340.1	103.3	421.9	12.53	29.08
	SA^{2}Net [113]	58.6	108.6	7.4	11.7	153.1	275.4	92.2	169.9	66.5	276.5	–	–	87.99	79.24
	GAPNet [114]	67.1	110.4	9.8	15.2	202.8	246.9	118.5	217.2	–	–	174.1	514.7	4.00	2.85
	AHNet [5]	67.5	106.0	7.7	11.9	197.3	268.5	108.2	186.8	–	–	100.2	364.1	7.21	24.98
	DA^{2}Net [4]	74.1	128.4	7.9	13.2	169.5	237.0	111.7	204.3	–	–	102.6	378.5	–	–
	SFCN [85]	64.8	107.5	7.6	13.0	214.2	318.2	102.0	171.4	77.5	297.6	–	–	7.65	38.60
	RAZ [115]	65.1	106.7	8.4	14.1	–	–	116.0	195.0	–	–	–	–	–	–
	AMFNet [116]	66.8	107.6	7.7	12.2	217.3	354.6	106.8	195.9	–	–	115.2	379.3	3.83	22.53
	AMSNet [117]	56.7	93.4	6.7	10.2	208.4	297.3	101.8	163.2	–	–	–	–	–	–
	ChfL [118]	57.5	94.3	6.9	11.0	–	–	80.3	137.6	57.0	235.7	76.8	343.0	2.71	21.51
	HMoDE [119]	54.4	87.4	6.2	9.8	159.6	211.2	–	–	55.7	214.6	–	–	5.13	82.62
	CLTR [120]	56.9	95.2	6.5	10.6	–	–	85.8	141.3	59.5	240.6	74.3	333.8	2.03	43.40
	P2PNet [64]	52.7	85.1	6.2	9.9	172.7	256.1	85.3	154.5	–	–	77.4	362.0	2.56	138.37
	PET [121]	49.3	78.7	6.1	9.6	159.9	223.7	79.5	144.3	58.5	238.0	74.4	328.5	2.54	21.60
	APGCC [122]	48.8	76.7	5.6	8.7	154.8	205.5	80.1	136.6	54.3	225.9	71.7	284.4	2.56	18.68
SS	DACount [123]	74.9	115.5	11.1	19.1	–	–	137.4	230.0	75.9	282.3	–	–	3.64	29.30
	DREAM [124]	86.5	121.2	15.1	23.8	251.5	341.1	109.0	187.2	75.9	282.3	–	–	8.35	16.26
	CSRNet [125]	72.8	111.6	12.0	18.7	294.0	443.1	128.1	218.1	129.7	400.5	178.7	1080.4	–	–
	OT-M [126]	81.6	127.1	10.9	18.1	–	–	107.9	180.6	75.5	287.9	–	–	–	–
	Meng et al. [74]	68.5	121.9	14.1	20.6	–	–	130.3	226.3	80.7	290.8	111.7	443.2	3.67	18.83
	MRC-Crowd [89]	67.3	106.8	10.3	18.2	–	–	93.4	153.2	70.7	261.3	–	–	3.58	34.75
	Liu et al. [127]	79.6	127.5	12.7	20.3	–	–	128.6	226.4	–	–	–	–	–	–
	IRAST [78]	86.9	148.9	14.7	22.9	–	–	135.6	233.4	–	–	–	–	2.57	17.37
	Li et al. [128]	70.8	116.6	9.7	17.7	–	–	104.0	164.3	74.9	281.7	108.8	458.0	–	–
US	D'Alessandro et al. [107]	196.0	295.2	49.0	60.3	–	–	390.0	697.5	194.0	583.9	–	–	–	–
	CrowdCLIP [66]	146.1	236.3	69.3	85.8	438.3	604.7	283.3	488.7	213.7	576.1	–	–	–	–
	GWTA-CCNN [103]	154.7	229.4	–	–	433.7	583.3	–	–	–	–	–	–	–	–
CS	Liu et al. [101]	112.2	218.2	13.4	29.3	368.0	518.9	175.0	294.8	–	–	–	–	–	–
	Ding et al. [129]	116.5	182.2	13.2	23.4	–	–	137.7	253.9	–	–	–	–	4.90	16.34
	CDANet [130]	106.5	162.5	13.5	22.3	–	–	169.2	308.0	–	–	–	–	–	–
	Liu et al. [131]	109.2	168.1	11.4	17.3	336.5	486.1	198.3	332.9	–	–	–	–	–	–
	RDBT [104]	103.6	200.8	11.6	21.0	361.3	504.5	172.8	291.9	–	–	–	–	–	–

6 Crowd Counting Challenges

While the use of CNNs in crowd counting models for density map estimation has significantly enhanced performance, certain challenges remain. These challenges are likely to continue posing considerable difficulties for future crowd counting models.

6.1 Challenges in the Data

Occlusion: High-density crowd images, such as Fig. 6, reveal that people frequently overlap and occlude one another. This poses significant challenges for traditional detection-based crowd counting methods in accurately identifying individuals. To overcome this problem, researchers transitioned from object detection to density map estimation as the preferred approach for crowd counting.

To date, most crowd counting methods rely on density map estimation; however, occlusion remains a significant challenge to counting accuracy. In response to occlusion challenges, numerous effective approaches have been developed. To address the occlusion problem in dense scenes, Chen et al. [112] proposed the Halo Attention module, which enhances the perception of the surrounding environment of objects through a large receptive field, thereby obtaining fine-grained image information.

Figure 6: Occlusion in the image

Scale changes: A person's size in an image changes based on their distance from the camera, as shown in Fig. 7. A person farther from the camera appears smaller than one who is closer. This makes it challenging for computers to identify all individuals accurately. Vertical scale variation is a common issue across nearly all datasets, necessitating crowd counting methods to address this challenge. This problem solutions include using multi-scale networks [5,14,44,116,132], and Single Shot MultiBox Detector (SSD) [43] or You Only Look Once (YOLO) [42], or employing multi-column [2,29,133–135]. Zhai et al. [113] proposed a multi-scale feature aggregator module that integrates multi-scale features to establish correlations across different scales, effectively addressing the issue of scale variation. Guo et al. [12] introduced a scale region recognition network featuring a scale-level awareness module that encodes the representation of counting objects across multiple scales, effectively addressing the issue of scale variation.

Figure 7: Scale variations in the image

Diverse illumination and weather: Illumination differences encompass variations between natural and artificial light sources. Natural light primarily influences the brightness of images at night rather than affecting their color. Furthermore, weather-induced blurring, such as from rain and fog, greatly affects background clarity and the overall sharpness of images. As illustrated in Fig. 8, varying illumination changes feature colors, and numerous interferences can be seen in the image background on snowy days. These challenges impose significant tests on the model's robustness.

Background confusion: Background chaos refers to situations where the background of an image has similar colors or textures to the foreground. In Fig. 9, the presence of a complex background can result in the model misidentifying certain elements. This challenge is typically addressed by applying regions of interest, semantic segmentation, and attention mechanisms [36,132,136]. Zhai et al. [137] introduced a channel-space self-attention mechanism that derived a context-sensing module to suppress background interference, thereby effectively mitigating its impact on counting accuracy.

Figure 8: Diverse illumination and weather

Figure 9: The challenges of background confusion in crowd counting

6.2 Challenges in the Network Algorithm

Model performance is impacted by various aspects, including model input, design, and the training process. The quality of the input plays a crucial role in determining the model's effectiveness, as it directly affects its overall performance. In contrast, the design of the model directly impacts network performance, making it a primary area of focus in research. Furthermore, the training process is critical, largely depending on the feedback mechanisms used. The loss function is directly related to the parametric performance of the model. Equally important are the evaluation criteria used to assess network performance beyond the training loss function.

Network input includes both training images and ground truth images. With the advancement of camera technology, most images in datasets are now high resolution. However, many real-world scenarios still involve low-resolution, outdated surveillance cameras. Therefore, to accommodate real-world data conditions, the model usually needs to preprocess the input image. Such preprocessing generally includes scaling, segmenting, and rotating the image. To reduce background interference and eliminate irrelevant information in images, foreground segmentation and image enhancement are employed. These methods can improve the efficacy of extracting relevant features more effectively.

The design of the network remains a crucial step for crowd counting tasks. Practically, the model's size, training speed, and runtime efficiency are critical aspects that deserve close attention. Currently, network model development primarily targets high-density crowd counting, resulting in models that are increasingly large and complex. However, extremely high-density crowd scenarios represent only a small fraction of real-world applications.

Network learning requires continuous iteration, and the loss function plays a critical role in this process. Currently, the most common method for crowd counting is regression prediction based on density maps, with the L2 loss function as the standard loss function. However, relying solely on Euclidean loss may cause the model to miss important spatial information. Designing innovative loss functions can significantly boost network performance. Examples include Adversarial Loss [138], SmoothL1 Loss [39], Tukey's biweight Loss [139], spatial correlation loss [136], and Maximum Excess over Pixels (MEP) loss [62]. Alternatively, multi-column networks can compute losses for each sub-network output to derive a comprehensive loss function. Moreover, using innovative or compound loss functions can greatly improve the effectiveness of crowd counting methods compared to using a single loss function.

7 Conclusion and Future Directions

This survey report discusses key design considerations and recent developments in crowd counting, emphasizing unsupervised and semi-supervised crowd counting methods. It also offers a comprehensive overview of supervised crowd counting techniques. The field of crowd counting research is diverse and rapidly evolving. The emerging trends and potential future research directions in semi-supervised/unsupervised population counting are summarized as follows:

- **Enhancing Model Generalization:** Although current semi-supervised and unsupervised methods show strong performance on specific datasets, they often fall short in generalizing to new scenes or across different datasets. Future research could focus on developing more effective cross-domain adaptation techniques [140,141] to enhance model robustness and adaptability across varied environments.
- **Enhancing Pseudo-Label Quality:** The quality of pseudo-labels directly impacts the effectiveness of semi-supervised learning. Although the use of soft and hard pseudo-labels in existing methods has improved model performance to some extent, issues of noise and inaccuracy persist. Future research could focus on optimizing pseudo-label generation algorithms, such as by employing improved self-training strategies and reliable pseudo-label learning frameworks to reduce noise in pseudo-labels [142,143].
- **Multimodal Information Fusion:** Crowd counting tasks often involve multiple types of data, such as text [66], images, videos, and sensor information [101]. Future research could focus on effectively integrating these multimodal sources to enhance the accuracy and robustness of counting models.
- **Efficient Data Utilization Strategies:** Effectively leveraging unlabeled data is a key challenge in semi-supervised and unsupervised learning. Future research could explore more efficient strategies for mining and utilizing unlabeled data, such as using graph representations and high-order relationship modeling to better capture underlying data structures [90].
- **Theoretical Foundations and Evaluation Standards:** Current semi-supervised and unsupervised learning methods largely rely on experimental validation and lack a solid theoretical foundation. Future research should focus on establishing a more rigorous theoretical framework and developing new evaluation standards to comprehensively assess model performance and generalization ability [144].

Acknowledgement: The authors would like to express their sincere gratitude and appreciation to the anonymous reviewers for their valuable comments to improve the paper.

Funding Statement: This work was supported by Research Project Support Program for Excellence Institute (2022, ESL) in Incheon National University.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Jianyong Wang, Mingliang Gao; data collection: Qilei Li; analysis and interpretation of results: Gwanggil Jeon; draft manuscript preparation: Hyunbum Kim. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] D. Kang and A. Chan, "Crowd counting by adaptively fusing predictions from an image pyramid," 2018, *arXiv:1805.06115*.

[2] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5744–5752.

[3] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1861–1870.

[4] W. Zhai *et al.*, "Da^{2}Net: A dual attention-aware network for robust crowd counting," *Multimed. Syst.*, vol. 29, no. 5, pp. 3027–3040, 2023. doi: 10.1007/s00530-021-00877-4.

[5] W. Zhai *et al.*, "An attentive hierarchy ConvNet for crowd counting in smart city," *Cluster Comput.*, vol. 26, no. 2, pp. 1099–1111, 2023. doi: 10.1007/s10586-022-03749-2.

[6] Y. Wang and Y. Zou, "Fast visual object counting via example-based density estimation," in *2016 IEEE Int. Conf. Image Process. (ICIP)*, IEEE, 2016, pp. 3653–3657.

[7] E. Walach and L. Wolf, "Learning to count with CNN boosting," in *Comput. Vis.–ECCV 2016: 14th Eur. Conf.*, Amsterdam, The Netherlands, Springer, 2016, pp. 660–676.

[8] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. of the 23th Int.Conf. on Neural Inf. Process. Sys.*, Red Hook, NY, USA, Curran Associates Inc., 2010, vol. 1, pp. 1324–1332.

[9] X. Guo, J. Chen, G. Zhang, G. Zou, Q. Li and M. Gao, "Cell counting via attentive recognition network," *Int. J. Comput. Sci. Eng.*, vol. 27, no. 1, pp. 1–8, 2024. doi: 10.1504/IJCSE.2024.136262.

[10] C. Arteta, V. Lempitsky, and A. Zisserman, "Counting in the wild," in *Comput. Vis.–ECCV 2016: 14th Eur. Conf.*, Amsterdam, The Netherlands, Springer, 2016, pp. 483–498.

[11] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Comput. Vis.–ECCV 2016: 14th Eur. Conf.*, Amsterdam, The Netherlands, Springer, 2016, pp. 615–629.

[12] X. Guo, M. Gao, W. Zhai, Q. Li, and G. Jeon, "Scale region recognition network for object counting in intelligent transportation system," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, pp. 15920–15929, 2023. doi: 10.1109/TITS.2023.3296571.

[13] X. Guo, M. Gao, G. Zou, A. Bruno, A. Chehri and G. Jeon, "Object counting via group and graph attention network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 9, pp. –11 884–11 895, 2024.

[14] W. Zhai, M. Gao, Q. Li, G. Jeon, and M. Anisetti, "FPANet: Feature pyramid attention network for crowd counting," *Appl. Intell.*, vol. 53, no. 16, pp. 19199–19216, 2023. doi: 10.1007/s10489-023-04499-3.

[15] J. Chen, M. Gao, X. Guo, W. Zhai, Q. Li and G. Jeon, "Object counting in remote sensing via selective spatial-frequency pyramid network," *Softw.: Pract. Exp.*, vol. 54, no. 9, pp. 1754–1773, 2024.

[16] S. Aich and I. Stavness, "Leaf counting with deep convolutional and deconvolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 2080–2089.

[17] M. V. Giuffrida, M. Minervini, and S. A. Tsaftaris, "Learning to count leaves in rosette plants," in *Proc. Comput. Vis. Probl. Plant Phenotyp. Workshop 2015*, 2015. doi: 10.5244/C.29.CVPPP.1.

[18] G. French, M. Fisher, M. Mackiewicz, and C. Needle, "Convolutional neural networks for counting fish in fisheries surveillance video," in *Workshop Mach. Vis. Anim. Behav., MVAB'15*, Swansea, UK, 2015. doi: 10.5244/C.29.MVAB.7.
[19] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L. -Q. Xu, "Crowd analysis: A survey," *Mach. Vis. Appl.*, vol. 19, pp. 345–357, 2008. doi: 10.1007/s00138-008-0132-4.
[20] J. Shao, K. Kang, C. C. Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4657–4666.
[21] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *2012 IEEE Conf. Comput. Vis. Pattern Recognit.*, IEEE, 2012, pp. 2871–2878.
[22] A. B. Chan, Z. -S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *2008 IEEE Conf. Comput. Vis. Pattern Recognit.*, IEEE, 2008, pp. 1–7.
[23] M. A. Khan, H. Menouar, and R. Hamila, "Revisiting crowd counting: State-of-the-art, trends, and future perspectives," *Image Vis. Comput.*, vol. 129, 2023, Art. no. 104597. doi: 10.1016/j.imavis.2022.104597.
[24] M. A. Khan, R. Hamila, and H. Menouar, "Visual crowd analysis: Open research problems," *AI Mag.*, vol. 44, no. 3, pp. 296–311, 2023. doi: 10.1002/aaai.12117.
[25] A. Bochkovskiy, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[26] H. Zhang, Y. Du, S. Ning, Y. Zhang, S. Yang and C. Du, "Pedestrian detection method based on faster R-CNN," in *2017 13th Int. Conf. Comput. Intell. Secur. (CIS)*, IEEE, 2017, pp. 427–430.
[27] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, 2011. doi: 10.1109/TPAMI.2011.155.
[28] X. Liu *et al.*, "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 350–359.
[29] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 589–597.
[30] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1091–1100.
[31] X. Ding, Z. Lin, F. He, Y. Wang, and Y. Huang, "A deeply-recursive convolutional network for crowd counting," in *2018 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, IEEE, 2018, pp. 1942–1946.
[32] H. Idrees *et al.*, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 532–546.
[33] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6142–6151.
[34] Y. Hou, S. Zhang, R. Ma, H. Jia, and X. Xie, "Frame-recurrent video crowd counting," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 33, no. 9, pp. 5186–5199, 2023. doi: 10.1109/TCSVT.2023.3250946.
[35] I. S. Topkaya, H. Erdogan, and F. Porikli, "Counting people by clustering person detector outputs," in *2014 11th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, IEEE, 2014, pp. 313–318.
[36] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," in *2008 19th Int. Conf. Pattern Recognit.*, IEEE, 2008, pp. 1–4.
[37] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *2005 IEEE Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR'05)*, IEEE, 2005, vol. 1, pp. 878–885. doi: 10.1109/CVPR.2005.272.
[38] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, 2008. doi: 10.1109/TPAMI.2008.260.
[39] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. of the 28th Int. Conf. on Neural Inf. Process. Sys.*, Cambridge, MA, USA, MIT Press, 2015, vol. 1, pp. 91–99.
[41] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
[42] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
[43] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Comput. Vis.–ECCV 2016: 14th Eur. Conf.*, Amsterdam, The Netherlands, Springer, 2016, pp. 21–37.
[44] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2547–2554.
[45] A. B. Chan and N. Vasconcelos, "Bayesian poisson regression for crowd counting," in *2009 IEEE 12th Int. Conf. Comput. Vis.*, IEEE, 2009, pp. 545–551.
[46] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. Brit. Mach. Vis. Conf.*, 2012. doi: 10.5244/C.26.21.
[47] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Crowd counting using multiple local features," in *2009 Digital Image Comput.: Tech. Appl.*, IEEE, 2009, pp. 81–88.
[48] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Seventh IEEE Int. Conf. Comput. Vis.*, IEEE, 1999, vol. 2, pp. 1150–1157. doi: 10.1109/ICCV.1999.790410.
[49] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Gray scale and rotation invariant texture classification with local binary patterns," in *Comput. Vis.-ECCV 2000: 6th Eur. Conf. Comput. Vis.*, Dublin, Ireland, Springer, 2000, pp. 404–420.
[50] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR'05)*, IEEE, 2005, vol. 1, pp. 886–893. doi: 10.1109/CVPR.2005.177.
[51] N. Paragios and V. Ramesh, "A MRF-based approach for real-time subway monitoring," in *Proc. 2001 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR 2001*, Kauai, HI, USA, IEEE, 2001. doi: 10.1109/CVPR.2001.990644.
[52] Y. Tian, L. Sigal, H. Badino, F. De la Torre, and Y. Liu, "Latent gaussian mixture regression for human pose estimation," in *Asian Conf. Comput. Vis.*, Springer, 2010, pp. 679–690.
[53] V. -Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3253–3261.
[54] C. Ren, G. Zhang, D. Jeong, and L. Hao, "CNN-based multi-object tracking networks with position correction and imm in intelligent transportation system," *Métodos numéricos para cálculo y diseño en ingeniería: Revista internacional*, vol. 39, no. 4, pp. 1–20, 2023.
[55] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *Proc. 23rd ACM Int. Conf. Multimed.*, 2015, pp. 1299–1302.
[56] M. Fu, P. Xu, X. Li, Q. Liu, M. Ye and C. Zhu, "Fast crowd density estimation with convolutional neural networks," *Eng. Appl. Artif. Intell.*, vol. 43, pp. 81–88, 2015. doi: 10.1016/j.engappai.2015.04.006.
[57] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 833–841.
[58] X. Wu, G. Liang, K. K. Lee, and Y. Xu, "Crowd density estimation using texture analysis and learning," in *2006 IEEE Int. Conf. Robot. Biomimetics*, IEEE, 2006, pp. 214–219.
[59] S. An, W. Liu, and S. Venkatesh, "Face recognition using kernel ridge regression," in *2007 IEEE Conf. Comput. Vis. Pattern Recognit.*, IEEE, 2007, pp. 1–7.
[60] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2467–2474.

[61] V. A. Sindagi and V. M. Patel, "CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *2017 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, IEEE, 2017, pp. 1–6.
[62] D. B. Sam, S. V. Peri, M. N. Sundararaman, A. Kamath, and R. V. Babu, "Locate, size, and count: Accurately resolving people in dense crowds via detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2739–2751, 2020.
[63] X. Jiang *et al.*, "Attention scaling for crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4706–4715.
[64] Q. Song *et al.*, "Rethinking counting and localization in crowds: A purely point-based framework," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3365–3374.
[65] P. Li, M. Zhang, J. Wan, and M. Jiang, "Multiscale aggregate networks with dense connections for crowd counting," *Comput. Intell. Neurosci.*, vol. 2021, no. 1, 2021, Art. no. 9996232. doi: 10.1155/2021/9996232.
[66] D. Liang, J. Xie, Z. Zou, X. Ye, W. Xu and X. Bai, "CrowdCLIP: Unsupervised crowd counting via vision-language model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2893–2903.
[67] Z. Peng and S. -H. G. Chan, "Single domain generalization for crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 28025–28034.
[68] S. -F. Lin, J. -Y. Chen, and H. -X. Chao, "Estimation of number of people in crowded scenes using perspective transformation," *IEEE Trans. Syst., Man, Cybern.-Part A: Syst. Humans*, vol. 31, no. 6, pp. 645–654, 2001. doi: 10.1109/3468.983420.
[69] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, pp. 137–154, 2004. doi: 10.1023/B:VISI.0000013087.49260.fb.
[70] Viola and Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proc. Ninth IEEE Int. Conf. Comput. Vis.*, IEEE, 2003, pp. 734–741.
[71] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Tenth IEEE Int. Conf. Comput. Vis. (ICCV'05)*, IEEE, 2005, vol. 1, pp. 90–97.
[72] V. A. Sindagi, R. Yasarla, and V. M. Patel, "JHU-CROWD++: Large-scale crowd counting dataset and a benchmark method," 2020. Accessed: Aug. 10, 2024. [Online]. Available: http://www.crowd-counting.com/
[73] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-Crowd: A large-scale benchmark for crowd counting and localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, pp. 2141–2149, 2021. doi: 10.1109/TPAMI.2020.3013269.
[74] Y. Meng *et al.*, "Spatial uncertainty-aware semi-supervised crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15549–15559.
[75] Y. Wang, J. Hou, X. Hou, and L. -P. Chau, "A self-training approach for point-supervised object detection and counting in crowds," *IEEE Trans. Image Process.*, vol. 30, pp. 2876–2887, 2021. doi: 10.1109/TIP.2021.3055632.
[76] V. A. Sindagi, R. Yasarla, D. S. Babu, R. V. Babu, and V. M. Patel, "Learning to count in the crowd from limited labeled data," in *Comput. Vis.–ECCV 2020: 16th Eur. Conf.*, Glasgow, UK, Springer, 2020, pp. 212–229.
[77] S. Khaki, H. Pham, Y. Han, A. Kuhl, W. Kent and L. Wang, "DeepCorn: A semi-supervised deep learning method for high-throughput image-based corn kernel counting and yield estimation," *Knowl.-Based Syst.*, vol. 218, 2021, Art. no. 106874. doi: 10.1016/j.knosys.2021.106874.
[78] Y. Liu, L. Liu, P. Wang, P. Zhang, and Y. Lei, "Semi-supervised crowd counting via self-training on surrogate tasks," in *Comput. Vis.–ECCV 2020: 16th Eur. Conf.*, Glasgow, UK, Springer, 2020, pp. 242–259. doi: 10.1007/978-3-030-58555-6_15.
[79] W. Li, Z. Cao, Q. Wang, S. Chen, and R. Feng, "Learning error-driven curriculum for crowd counting," in *2020 25th Int. Conf. Pattern Recognit. (ICPR)*, IEEE, 2021, pp. 843–849.
[80] J. Wan and A. Chan, "Modeling noisy annotations for crowd counting," *Adv. Neural Inform. Process. Syst.*, vol. 33, pp. 3386–3396, 2020.

[81] V. A. Sindagi and V. M. Patel, "HA-CCN: Hierarchical attention-based crowd counting network," *IEEE Trans. Image Process.*, vol. 29, pp. 323–335, 2019. doi: 10.1109/TIP.2019.2928634.

[82] G. Olmschenk, J. Chen, H. Tang, and Z. Zhu, "Dense crowd counting convolutional neural networks with minimal data using semi-supervised dual-goal generative adversarial networks," in *IEEE Conf. Comput. Vis. Pattern Recognit.: Learn. Imperfect Data Workshop*, 2019.

[83] G. Olmschenk, Z. Zhu, and H. Tang, "Generalizing semi-supervised generative adversarial networks to regression using feature contrasting," *Comput. Vis. Image Underst.*, vol. 186, pp. 1–12, 2019. doi: 10.1016/j.cviu.2019.06.004.

[84] Y. Lei, Y. Liu, P. Zhang, and L. Liu, "Towards using count-level weak supervision for crowd counting," *Pattern Recognit.*, vol. 109, 2021, Art. no. 107616. doi: 10.1016/j.patcog.2020.107616.

[85] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8198–8207.

[86] 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, 2022, Art. no. 160104. doi: 10.1007/s11432-021-3445-y.

[87] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Exploiting unlabeled data in CNNs by self-supervised learning to rank," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1862–1878, 2019. doi: 10.1109/TPAMI.2019.2899857.

[88] X. Wei, Y. Qiu, Z. Ma, X. Hong, and Y. Gong, "Semi-supervised crowd counting via multiple representation learning," *IEEE Trans. Image Process.*, vol. 32, pp. 5220–5230, 2023. doi: 10.1109/TIP.2023.3313490.

[89] Y. Qian, X. Hong, Z. Guo, O. Arandjelović, and C. R. Donovan, "Semi-supervised crowd counting with contextual modeling: Facilitating holistic understanding of crowd scenes," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 34, pp. 8230–8241, 2024. doi: 10.1109/TCSVT.2024.3392500.

[90] Z. Miao, Y. Zhang, X. Piao, Y. Chu, and B. Yin, "Region feature smoothness assumption for weakly semi-supervised crowd counting," *Comput. Animat. Virtual Worlds*, vol. 34, no. 3–4, 2023, Art. no. e2173. doi: 10.1002/cav.2173.

[91] Q. Zhou, J. Zhang, L. Che, H. Shan, and J. Z. Wang, "Crowd counting with limited labeling through submodular frame selection," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1728–1738, 2018. doi: 10.1109/TITS.2018.2829987.

[92] S. Denis, B. Bellekens, A. Kaya, R. Berkvens, and M. Weyn, "Large-scale crowd analysis through the use of passive radio sensing networks," *Sensors*, vol. 20, no. 9, 2020, Art. no. 2624. doi: 10.3390/s20092624.

[93] M. A. Khan, H. Menouar, and R. Hamila, "DroneNet: Crowd density estimation using self-onns for drones," in *2023 IEEE 20th Consum. Commun. Netw. Conf. (CCNC)*, IEEE, 2023, pp. 455–460.

[94] M. A. Khan, H. Menouar, and R. Hamila, "LCDnet: A lightweight crowd density estimation model for real-time video surveillance," *J. Real Time Image Process.*, vol. 20, no. 2, 2023, Art. no. 29. doi: 10.1007/s11554-023-01286-8.

[95] M. A. Khan, R. Hamila, and H. Menouar, "CLIP: Train faster with less data," in *2023 IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, IEEE, 2023, pp. 34–39.

[96] F. Maire, E. Moulines, and S. Lefebvre, "Online em for functional data," *Comput. Stat. Data Anal.*, vol. 111, pp. 27–47, 2017. doi: 10.1016/j.csda.2017.01.006.

[97] A. A. Patel, *Hands-On Unsupervised Learning Using Python: How to Build Applied Machine Learning Solutions from Unlabeled Data*. Sebastopol, California: O'Reilly Media, 2019.

[98] L. -Z. Guo, Z. -Y. Zhang, Y. Jiang, Y. -F. Li, and Z. -H. Zhou, "Safe deep semi-supervised learning for unseen-class unlabeled data," in *Int. Conf. Mach. Learn.*, PMLR, 2020, pp. 3897–3906.

[99] R. Zhu *et al.*, "SD-DiT: Unleashing the power of self-supervised discrimination in diffusion transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 8435–8445.

[100] D. B. Sam, A. Agarwalla, J. Joseph, V. A. Sindagi, R. V. Babu and V. M. Patel, "Completely self-supervised crowd counting via distribution matching," in *Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 186–204.

[101] Y. Liu, Z. Wang, M. Shi, S. Satoh, Q. Zhao and H. Yang, "Towards unsupervised crowd counting via regression-detection bi-knowledge transfer," in *Proc. 28th ACM Int. Conf. Multimed.*, 2020, pp. 129–137.

[102] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Leveraging unlabeled data for crowd counting by learning to rank," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7661–7669.
[103] D. B. Sam, N. N. Sajjan, H. Maurya, and R. V. Babu, "Almost unsupervised learning for dense crowd counting," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 8868–8875, 2019. doi: 10.1609/aaai.v33i01.33018868.
[104] Y. Liu, Z. Wang, M. Shi, S. Satoh, Q. Zhao and H. Yang, "Discovering regression-detection bi-knowledge transfer for unsupervised cross-domain crowd counting," *Neurocomputing*, vol. 494, pp. 418–431, 2022. doi: 10.1016/j.neucom.2022.04.107.
[105] N. Jiang, X. Wen, and Z. Shi, "DAPC: Domain adaptation people counting via style-level transfer learning and scene-aware estimation," in *2020 25th Int. Conf. Pattern Recognit. (ICPR)*, IEEE, 2021, pp. 1067–1074.
[106] X. Hou, J. Xu, J. Wu, and H. Xu, "Cross domain adaptation of crowd counting with model-agnostic meta-learning," *Appl. Sci.*, vol. 11, no. 24, 2021, Art. no. 12037. doi: 10.3390/app112412037.
[107] A. D'Alessandro, A. Mahdavi-Amiri, and G. Hamarneh, "SYRAC: Synthesize, rank, and count," 2023, *arXiv:2310.01662*.
[108] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 8748–8763.
[109] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Int. Conf. Mach. Learn.*, PMLR, 2022, pp. 12888–12900.
[110] F. Xu, Y. Rao, and Q. Wang, "An unsupervised abnormal crowd behavior detection algorithm," in *2017 Int. Conf. Secur., Pattern Anal., Cybern. (SPAC)*, IEEE, 2017, pp. 219–223.
[111] P. Danielis, S. T. Kouyoumdjieva, and G. Karlsson, "UrbanCount: Mobile crowd counting in urban environments," in *2017 8th IEEE Annu. Inform. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, IEEE, 2017, pp. 640–648.
[112] J. Chen *et al.*, "Privacy-aware crowd counting by decentralized learning with parallel transformers," *Internet of Things*, vol. 26, 2024, Art. no. 101167. doi: 10.1016/j.iot.2024.101167.
[113] W. Zhai, M. Gao, X. Guo, G. Zou, Q. Li and G. Jeon, "Scale attentive aggregation network for crowd counting and localization in smart city," *ACM Trans. Sens. Netw.*, 2024. doi: 10.1145/3653454.
[114] X. Guo, K. Song, M. Gao, W. Zhai, Q. Li and G. Jeon, "Crowd counting in smart city via lightweight ghost attention pyramid network," *Future Gener. Comput. Syst.*, vol. 147, pp. 328–338, 2023. doi: 10.1016/j.future.2023.05.013.
[115] C. Liu, X. Weng, and Y. Mu, "Recurrent attentive zooming for joint crowd counting and precise localization," in *2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 1217–1226. doi: 10.1109/CVPR.2019.00131.
[116] X. Guo *et al.*, "Crowd counting via attention and multi-feature fused network," *Human-Centric Comput. Inform. Sci.*, vol. 13, 2023.
[117] Y. Hu *et al.*, "NAS-Count: Counting-by-density with neural architecture search," in *Comput. Vis.–ECCV 2020: 16th Eur. Conf.*, Glasgow, UK, Springer, 2020, pp. 747–766.
[118] W. Shu, J. Wan, K. C. Tan, S. Kwong, and A. B. Chan, "Crowd counting in the frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19618–19627.
[119] Z. Du, M. Shi, J. Deng, and S. Zafeiriou, "Redesigning multi-scale neural network for crowd counting," *IEEE Trans. Image Process.*, vol. 32, pp. 3664–3678, 2023. doi: 10.1109/TIP.2023.3289290.
[120] D. Liang, W. Xu, and X. Bai, "An end-to-end transformer model for crowd localization," in *Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 38–54.
[121] C. Liu, H. Lu, Z. Cao, and T. Liu, "Point-query quadtree for crowd counting, localization, and more," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 1676–1685.
[122] I. Chen *et al.*, "Improving point-based crowd counting and localization based on auxiliary point guidance," 2024, *arXiv:2405.10589*.
[123] H. Lin, Z. Ma, X. Hong, Y. Wang, and Z. Su, "Semi-supervised crowd counting via density agency," in *Proc. 30th ACM Int. Conf. Multimed.*, 2022, pp. 1416–1426.

[124] J. Gao *et al.*, "Deep rank-consistent pyramid model for enhanced crowd counting," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, 2023. doi: 10.1109/TNNLS.2023.3336774.
[125] Y. Xu *et al.*, "Crowd counting with partial annotations in an image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15570–15579.
[126] W. Lin and A. B. Chan, "Optimal transport minimization: Crowd localization on density maps for semi-supervised counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21663–21673.
[127] Y. Liu *et al.*, "Reducing spatial labeling redundancy for semi-supervised crowd counting," *arXiv preprint arXiv:2108.02970*, 2021.
[128] C. Li, X. Hu, S. Abousamra, and C. Chen, "Calibrating uncertainty for semi-supervised crowd counting," in *2023 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, IEEE, 2023, pp. 16685–16695.
[129] G. Ding, D. Yang, T. Wang, S. Wang, and Y. Zhang, "Crowd counting via unsupervised cross-domain feature adaptation," *IEEE Trans. Multimed.*, vol. 25, pp. 4665–4678, 2022. doi: 10.1109/TMM.2022.3180222.
[130] A. Zhang, J. Xu, X. Luo, X. Cao, and X. Zhen, "Cross-domain attention network for unsupervised domain adaptation crowd counting," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 32, no. 10, pp. 6686–6699, 2022. doi: 10.1109/TCSVT.2022.3179824.
[131] W. Liu, N. Durasov, and P. Fua, "Leveraging self-supervision for cross-domain crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5341–5352.
[132] Y. Chen, C. Gao, Z. Su, X. He, and N. Liu, "Scale-aware rolling fusion network for crowd counting," in *2020 IEEE Int. Conf. Multimed. Expo (ICME)*, IEEE, 2020, pp. 1–6.
[133] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "Decidenet: Counting varying density crowds through attention guided detection and density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5197–5206.
[134] Y. Yang, G. Li, D. Du, Q. Huang, and N. Sebe, "Embedding perspective analysis into multi-column convolutional neural network for crowd counting," *IEEE Trans. Image Process.*, vol. 30, pp. 1395–1407, 2020. doi: 10.1109/TIP.2020.3043122.
[135] 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, 2022. doi: 10.1089/big.2022.0039.
[136] X. Jiang *et al.*, "Crowd counting and density estimation by trellis encoder-decoder networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6133–6142.
[137] W. Zhai, M. Gao, X. Guo, Q. Li, and G. Jeon, "Scale-context perceptive network for crowd counting and localization in smart city system," *IEEE Internet Things J.*, vol. 10, no. 21, pp. 18930–18940, 2023. doi: 10.1109/JIOT.2023.3268226.
[138] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[139] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab, "Robust optimization for deep regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2830–2838.
[140] S. Peng, B. Yin, Y. Xia, Q. Yang, and L. Wang, "Semi-supervised crowd counting based on patch crowds statistics," in *2022 Asia Conf. Algorithms, Comput. Mach. Learn. (CACML)*, IEEE, 2022, pp. 749–755.
[141] C. Li, H. Yin, Y. Xu, and J. Wan, "Semi-supervised dense object counting via mutual consistency learning," in *2022 10th Int. Conf. Inform. Syst. Comput. Technol. (ISCTech)*, IEEE, 2022, pp. 386–391.
[142] P. Zhu, J. Li, B. Cao, and Q. Hu, "Multi-task credible pseudo-label learning for semi-supervised crowd counting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 8, pp. 10394–10406, 2023.
[143] H. Li, Y. Song, and T. Geng, "Semi-supervised crowd counting based on hard pseudo-labels," in *2024 Int. Joint Conf. on Neural Networks (IJCNN)*, IEEE, 2024, pp. 1–8.
[144] E. Tu and J. Yang, "A review of semi supervised learning theories and recent advances," 2019, *arXiv:1905.11590*.