Artificial Intelligence Guided Crowd Counting and Density Estimation with Enhanced CSRNet

Muhammad Jawad Babar[1], Malik Muhammad Saad Missen[2],
Hannan Adeel[3], Muhammad Usman[4], Muzammil Malik[5]
Faculty of Computing, The Islamia University of Bahawalpur, Pakistan[1,2]
Universiti Teknikal Malaysia Melaka, Malaysia[3]
Edge Hill University, UK[4]
Hamdard University, Islamabad, Pakistan[5]

***Abstract*—Estimating crowd size in dense environments remains a complex problem, yet it holds critical value for safety monitoring, city infrastructure design, and large-scale gathering coordination. Leveraging contemporary developments in neural network architectures and machine intelligence, researchers have markedly enhanced the precision of population counts derived from both still imagery and motion footage. This research focuses on the development of an enhanced deep neural network-based crowd counting model. Since the original CSRNet primarily relies on head detection for crowd estimation, an additional face detection module has been incorporated to improve its capability in scenarios where facial features are visible. The proposed enhancement increases the flexibility and estimation accuracy of CSRNet by integrating individual face detection with crowd density estimation. Furthermore, this study presents a comprehensive comparative analysis of the proposed enhanced CSRNet against three state-of-the-art crowd counting models, namely Bayesian Network (BAYNet), Distribution Matching (DM-Count), and Scale Aggregation Feature Attention Network (SFANet). The models are evaluated on multiple datasets to investigate their accuracy, robustness, adaptability to varying crowd densities, and performance under complex environmental conditions. The comparison also highlights the methodological differences and computational characteristics of these approaches. Model performance is assessed using the standard evaluation metrics of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Experimental results demonstrate that the proposed enhanced CSRNet achieves a competitive RMSE/MAE of $9.61/93.05$ on 64×64 images of the custom dataset, outperforming the baseline models and demonstrating its effectiveness for accurate crowd counting.**

***Keywords*—Crowd counting; machine learning; image processing; CSRNet; density estimation; deep learning; face detection**

I. Introduction

The problem of crowd counting and density estimation has emerged as a crucial field of study to solve many practical applications. The challenges arising from the growing urbanization and from the large-scale public events in the real world. The techniques are very important in the field of urban planning, public safety and resource allocation. Precise crowd estimation enables the authorities to regulate the number of people in a packed area, and this allows for prompt answer during the crisis and enables efficient event planning [1], [2], [3]. Crowd counting inherently connects to intelligent systems via cutting-edge algorithms and advanced technologies like computer vision and machine learning. These systems are essential for precisely identifying and enumerating people in dense environments, a task that relies heavily on ongoing methodological innovations. Intelligent systems are well-suited for analyzing visual information and using methods such as object detection, Convolutional Neural Networks (CNNs), and deep learning to accurately and efficiently estimate crowd size and size, providing insights for various applications including urban planning, security surveillance, and event management [4], [5].

Current methods are generally time-consuming, imprecise, and not suitable for larger-scale applications, such as manual counting or basic counting devices [6], [7], [8]. The traditional methods might not be as effective in real-world situations in which large and dense crowds are present, resulting in less accuracy. These restrictions have led to a demand for more complex, automated and precise solutions. The advanced neural network architectures used in deep learning methods have the potential to transform the field of crowd counting and density estimation [9], [10], [11], [12]. Several deep learning approaches such as Congested Scene Recognition Network (CSRnet), Scale Aggregation Feature Attention Network (SFANet), and Bayesian Network (BayNet) automatically acquire complex spatial features and patterns, enhancing the accuracy and scalability.

Recent advances have further pushed the boundaries of what is achievable in crowd analysis. Multi-scale feature fusion approaches [13], [14], perspective-assisted semi-supervised learning methods [15], and generative model-based density estimation techniques [16] have collectively demonstrated that combining richer contextual information with improved training paradigms can substantially close the performance gap that previously existed between controlled benchmark settings and challenging real-world deployments. Despite these advancements, accurately estimating crowd density under extreme occlusion, significant scale variation, and complex lighting conditions remains an open challenge, motivating the exploration of improved architectures such as the one proposed in this work.

The usually adopted approach for crowd counting research consists of the following steps [17], [1], [18], [19]:

- Collect primary visual data: images or video footage; captured on-site to obtain a ground-level count of individuals within each unit frame.
- Preprocessing: Noise filtering, resizing images and modifying pixels.

Fig. 1. Objects or people counting example.

Fig. 2. Overlapped persons in crowd image, shown in red box.

- Identifiable Feature Extraction: Obtaining identifiable features from the data set available (images or videos) to identify or count people.
- Algorithms used for counting: Algorithms for counting are selected based on the extracted features, with each offering varied approximations of per-frame pedestrian tallies. These approaches generally fall into regression, density estimation, or detection-oriented frameworks.
- Evaluation:The performance of the proposed crowd analysis approach is quantified using standard error metrics, namely mean absolute error (MAE) and root mean square error (RMSE), to ensure a thorough assessment of its predictive accuracy.
- Flaws in Counting/Extracting Data – Consider improving the flaws in counting or extracting data by analysing the data.
- Application: The algorithm finds practical use in crowd regulation, public safety protocols, tactical decision-making, and monitoring systems that operate on live data streams.

In this work, CSRNet was adopted and subsequently enhanced by integrating BayNet and SFANet to improve face and head-counting accuracy. The combined model was then evaluated to assess its overall performance. The main objective centered on leveraging facial and head count information within the CSRNet architecture to achieve reliable crowd counting and density estimation.

In Fig. 1 the density map is a collection of people, and each blue dot is an individual person.

A. Motivation

The motivation for this investigation stems from several interconnected considerations, which are outlined below.

1) Practical relevance and real-world deployment: Crowd counting plays a pivotal role in numerous practical domains, including public event management, traffic flow monitoring, and public safety assurance in congested urban environments. The deployment of accurate and reliable counting models offers tangible benefits in these scenarios. Nevertheless, the task remains inherently challenging due to pronounced variations in crowd density, severe occlusion (as illustrated in Fig. 2), perspective distortion, and fluctuating illumination conditions. These inherent challenges have motivated continued investigation into improving counting system precision and resilience, thereby supporting the wider evolution of computer vision techniques.

2) Societal implications and public welfare: Beyond its technical merits, crowd counting carries significant societal weight, particularly in safety-critical contexts such as emergency response coordination, large-scale public gathering management, and evidence-based urban planning. Accurate density estimation facilitates timely and informed decision-making, ultimately improving public safety outcomes and enabling more efficient resource allocation during mass events.

3) Advancements in deep learning architectures: The recent proliferation of sophisticated deep learning models; including CSRNet, SFANet, and BayNet; has substantially expanded the methodological landscape for crowd analysis. Thes architectures, representative of state-of-the-art developments in computer vision, offer new avenues for investigation. Researchers are consequently driven to systematically evaluate, extend, and refine these models, with particular emphasis on understanding their respective strengths, inherent limitations, and generalizability across diverse counting scenarios.

4) Benchmarking and comparative evaluation: Rigorous comparative analysis of existing deep learning approaches constitutes a fundamental component of empirical research in crowd counting. Systematic benchmarking not only establishes performance baselines but also elucidates the conditions under which specific models exhibit superior or inferior behavior. Such insights are indispensable for guiding future architectural innovations, informing model selection for practical applications, and fostering reproducible progress within the research community.

B. Aims and Objectives

This study sets out to enhance both the accuracy and efficiency of crowd counting by adopting an improved con-

volutional neural network: CSRNet. Specifically, we aim to evaluate how well the modified CSRNet architecture estimates crowd density across diverse real-world scenarios, with an eye toward informing crowd management, urban planning, and public safety measures. To that end, our work compares the performance of the enhanced CSRNet against state-of-the-art counting methods, assesses its generalizability across multiple benchmark datasets, and explores its practical viability in operational settings.

- A precise and computationally efficient crowd counting framework will be established, beginning with a comprehensive examination of extant deep learning literature. This survey aims to pinpoint leading methodologies, recognize their inherent shortcomings, and delineate evolving directions within the domain. Informed by these findings, a customized deep architecture will be conceived and deployed to tackle the principal obstacles in crowd enumeration.
- Tackle real-world challenges in crowd counting applications, particularly the issue of occlusion. The aim is to improve counting accuracy in partially obscured scenarios by leveraging contextual information and incorporating occlusion-aware features into the model.
- Evaluate and compare the proposed model against existing solutions using appropriate performance metrics. The assessment will focus on accuracy, precision, recall, and computational efficiency to ensure a comprehensive understanding of the model's strengths and limitations.

C. Our Contribution

The contribution of this work is twofold:

- The construction of compact libraries specifically designed for crowd counting facilitates computational economy, rendering them viable for implementation on hardware with limited resources and within time-sensitive applications. This strategy consequently broadens the practical utility of counting systems across varied real-world contexts.
- The proposed face detection module integrated on CSRNet is designed to detect the features of a face instead of head, and this feature imparts the ability to count faces, even in dense scenes, and can more effectively estimate the number of persons in the image.

The remainder of this study is organized as follows. Section II offers a broad overview of the relevant literature and describes the datasets employed for crowd counting and density estimation. In Section IV, we introduce the proposed model, explain how the dataset was constructed, and define the evaluation metrics used to assess performance. Section V reports the experimental findings, and Section X concludes the study by discussing potential avenues for future research.

II. Literature Review

Within the field of machine vision, crowd counting and density estimation have drawn growing interest from research groups across a wide range of international institutions. These tasks are regarded as significantly important in the analysis of surveillance videos and images, whether conducted in real-time or offline modes. Several established and emerging algorithms have been applied to estimate crowd sizes from images or video, yet many fall short of expected performance, largely owing to the wide variability present across different data samples. To address this problem, machine learning models, including CNNs and their variants, have been introduced.The primary challenges and core tasks associated with crowd counting and density estimation in computer vision are outlined here.

Fig. 3. An example of an annotated crowd image. The plus (+) sign identifies each person.

The central objective of crowd counting is to deliver a reliable enumeration of individuals within a gathering, supporting applications in event coordination, oversight, and public safety. The majority of existing approaches rely on machine learning techniques for visual data processing, enabling both detection and tracking, with subsequent statistical procedures applied to derive aggregate headcounts from the captured footage [9], [10], [11], [12].

III. Crowd Counting: An Overview

Estimating the number of people in a crowd is one of the basic problems of computer vision, such as in the work by Sacchi in 2001 ([20]), Lin in 2001 ([21]), and Ranjan in 2018 ([22]). This field involves solving problems arising from varying environments, densities, obstacles, and crowd sizes, through machine learning and advanced algorithms. The crowd counting problem is one of the important problems in the city area planning, event management, and safety issues based on visual data [23], [24]. The primary focus is on creating more sophisticated models capable of automatically, accurately and rapidly counting individuals in dense environments, such as for resource allocation or decision making. The picture shown in Fig. 3 is one of a crowd and each person in the picture is marked by a plus (+) sign.

The history of crowd counting estimation methods is a history of manual methods, which are human-based. During the late 20th century, the technology of computer vision started to get advanced, which led to the development of automated

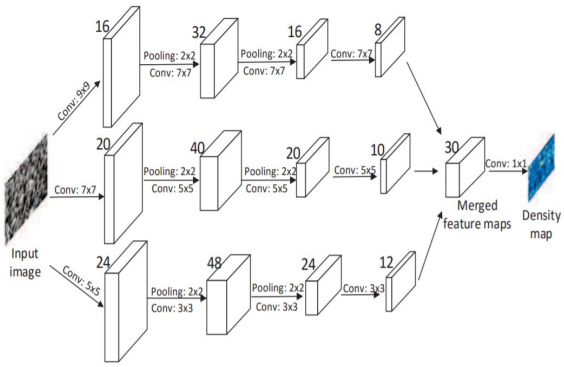

Fig. 4. An illustration of the proposed CNN-driven framework for crowd enumeration, as introduced in [18].

methods for crowd counting. The traditional approaches were based on feature engineering and density estimation. In recent years, crowd counting is becoming an increasingly popular task, and, thanks to the development of deep learning, in particular of the CNN, it has been transformed in the last decade (cf. [25], [26], [27]). These models utilize highly annotated datasets to train, enhance accuracy and flexibility in various deployment contexts. The transformation is a testament to the shift from traditional, human-based methods to the latest in AI-driven solutions, marking an important advancement in the field of crowd analysis, urban planning, and security solutions.

The crowd counting process using machine learning is carried out based on the model and pipeline presented in Fig. 4.

CNNs are often used as crowd counters because they are good at extracting features from the data and they can learn and represent complex data patterns. People count is challenging because it is necessary to estimate a large number of objects with substantial variation in appearance because of occlusion, scale and perspective. The main reason why CNN is being used for crowd counting as well as crowd density estimation is that it has the capability to learn the nonlinear relationships present in the data.

A. Crowd Counting at Image Level

The section provides comprehensive overview of the existing literature on crowd counting from images of overcrowded scenes.

Recent studies [7], [8], [28] have introduced deep CNN-based regression architectures, which have demonstrated superior performance over earlier approaches relying on manually engineered features or supplementary detection modules. Extensive experimental evaluations were conducted using widely adopted benchmark datasets, namely UCSD [29], UCF-CC [30], PASCAL VOC 2012 [31], and PETS-2009 [30]. The performance was measured based on the mean and variance of the absolute differences. A CNN-based architecture called CrowdNet was proposed in [32], which has been recognized as a notable advancement in the field. This architecture integrates deep and shallow fully convolutional branches to jointly extract high-level semantic information and fine-grained spatial features, both of which are considered essential for reliable counting in the presence of significant scale variability. The evaluation on the challenging UCF-CC 50 dataset demonstrated improved performance over prior art, with detailed outcomes presented in Table I.

TABLE I. A QUANTITATIVE EVALUATION OF THE METHOD PRESENTED IN [32] WAS CONDUCTED IN COMPARISON WITH CONTEMPORARY STATE-OF-THE-ART TECHNIQUES, USING THE UCF-CC 50 BENCHMARK

Method	Mean Absolute Error (MAE)
Density-aware Detection [33]	652.5
FHSc	467.5
Learning to Count [34]	490.2
Density-aware Detection [35]	468.2
Proposed [32]	423.7

Likewise, the work of [36] presented an innovative method to count people in surveillance video with a deep DCNN, which was also applied to this task. The unique factor in their approach is that they do not follow a conventional approach such as those presented in [37], [38], [39] which only use the number of pedestrians crossing a line that has been predetermined. Unlike the case of static pedestrians, the authors consider the exact time that a pedestrian takes to traverse the Line of Interest (LOI), which greatly improves CNN in learning more representative and key feature representations. Localization of Regions of Interest (ROIs) in video or image is a critical aspect for many crowd monitoring related problems, such as regression problems. This idea has been explored in the studies of [40] and [41], where a variant of Recurrent Neural Networks (RNNs), namely Long Short-Term Memory (LSTM), is utilized to capture sequential information and temporal dependencies during model training. To ensure a reliable correspondence between image content and the associated density map, irrespective of image size or resolution, previous research has proposed a simplified Multi-column Convolutional Neural Network (MCNN) architecture (Zhang et al., 2016; Zeng et al., 2017; Babu et al., 2017; Zhu et al., 2020; Hu et al., 2020).

A quality-based density estimation network, termed the Contextual Pyramid CNN (CP-CNN), is proposed for images with dense crowded scenes. This architecture operates through four core components: the Global Context Estimator (GCE), the Local Context Estimator (LCE), the Density Map Estimator (DME), and the Fusion-CNN (F-CNN), each constructed around the use of both broad and fine-grained contextual cues to support crowd enumeration. A deep and recursive network, referred to as DRResNET, was employed for crowd counting in [42]. An Adaptive Counting Convolutional Neural Network (A-CCNN) was proposed in [43] to enhance the accuracy of crowd counting, particularly for estimating crowd size. Loss functions are recognized as a key component in the evaluation of algorithmic performance. A detailed explanation of the composition of loss functions used for density map estimation in crowded scenes is provided in [44] and [45].

Persons counting in cross-scene problems is challenging since the background changes during model training. This is effectively solved by [51] who use a cross-scene counting

TABLE II. A STATISTICAL OVERVIEW OF THE BENCHMARK DATASETS UTILIZED IN CROWD COUNTING AND CROWD DENSITY ESTIMATION TASKS

Dataset	No. **of images**	Average **Count**	Source **of images**	Avg. size
UCF-QNRF [45]	1535	50	FLICKR, Web Search	2500×187
Mall [46]	2000	24	Public places	—
JHU-crowd [47]	4, 370	110	FLICKR, ShutterStock	1000×750
ShanghaiTech A [48]	475	490	Shanghai Streets	584×864
ShanghaiTech B	720	390	Shanghai Streets	584×864
WorldExpo'10 [49]	103	473	Shanghai 2010 WorldExpo	589×86
UCF-CC 50 [30]	50	1279	FLICKR	984×648
NWPU-Crowd [50]	5, 109	418	Resort, Walking street	2191×3209
UCSD [29]	1992	45	UC San Diego	157×237

model to learn other scenes knowledge transfer, through the features extracted from manually segment foregrounds and the domain adaptation method of extreme learning machine (DA-ELM).

Likewise, the scale variation problem was tackled in the study [52] by leveraging image pyramids to facilitate the prediction across different scales, and thereby taking advantage of adaptively varying weights per pixel. Crowd counting also has been explored using Generative Adversarial Networks (GANs) [53], [54], [55], [56]. In order to overcome the most common problems about scale and perspective changes, the authors of [57] proposed a Scale-Adaptive CNN (SaCNN) model. The authors in [58] followed CNN based methods and removed fully connected layers from the VGG16 model [59] and used pyramid networks for more consistent representations. In [60], [61], a deep learning method called Congested Scene Recognition (CSRNet) was used to understand the images with dense crowds and generate reasonable density maps. The proposed system consists of two main parts: The first part consists of CNN for 2D feature extraction, and the second part is dilated-layer CNN for expanding the receptive field without using the traditional pooling layers. In order to produce an accurate density map in a single step, the authors in [22] proposed a two-stage CNN model, first generating a rough outline of the density map and then refining the outline to obtain a high-resolution density map.

Recently, the authors in [62] introduced a weakly supervised framework for crowd counting, named JCTNet (Joint CNN and Transformer Network). The proposed architecture integrates three key components: the CNN Feature Extraction Module (CFM), the Transformer Feature Extraction Module (TFM), and the Counting Regression Module (CRM). The semantic crowd features are extracted by the CFM, while the extracted feature maps are partitioned into patches by the TFM so that global contextual information can be captured using a Transformer. Finally, the crowd count is regressed by the CRM from the fused representations. The proposed JCTNet aims to fuse CNN-derived local representations with Transformer-driven global contextual encoding, leveraging the respective advantages of each architectural paradigm. While JCTNet achieves a considerable reduction in parameter count; roughly 67% to 73% less than a pure Transformer counterpart; its computational cost remains non-trivial due to the concurrent operation of two backbone networks. In a separate contribution, Zhang et al. [63] proposed a cross-domain crowd counting architecture augmented with frequency-based processing, comprising a module for feature refinement in the spectral domain alongside an adapter designed to preserve domain-agnostic frequency characteristics. In this design, the Discrete Cosine Transform is utilized by the enhancement module to derive informative frequency statistics, thereby reinforcing the representational capacity of spatial features. Concurrently, the adapter incorporated Fast Fourier Transform (FFT)-based amplitude suppression in conjunction with an attention mechanism, which served to suppress domain-specific frequency cues and promote the acquisition of domain-invariant feature representations. To further mitigate complex background interference, a Parallel Dual Attention Module is incorporated, by which foreground crowd regions are emphasized. Moreover, decoder features from multiple scales are fused by a Multi-Scale Feature Aggregation Module, thereby improving robustness against large variations in crowd density and object scale. Superior counting accuracy is demonstrated by experimental evaluations on four public crowd counting datasets, particularly in challenging cross-domain scenarios, to be achieved by the proposed approach. Nevertheless, several additional processing modules are introduced by the framework, making it dependent on increased architectural complexity and computational overhead.

B. Crowd Counting at Video Level

The application of convolutional neural networks (CNNs) to crowd counting was first established in the pioneering work of Zhang et al. [6], who introduced an early deep learning framework capable of addressing the challenges associated with estimating crowd density across diverse scenes. Their method addressed the challenge of estimating crowd counts in previously unseen surveillance environments without requiring extensive scene-specific annotations. The proposed deep CNN was jointly optimized for two complementary objectives:

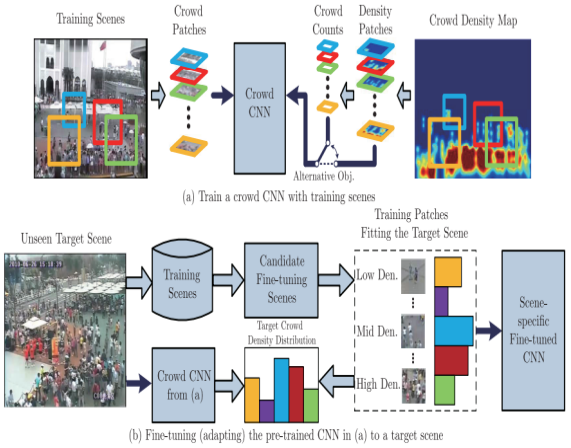

Fig. 5. Illustration of the model used for crowd counting in [6].

crowd density estimation and crowd count prediction, enabling improved generalization across diverse scenes. Validation of the proposed method was carried out across multiple standard benchmarks, specifically the WorldExpo'10, the UCSD Pedestrian Dataset [29], and the UCF-CC-50 Dataset, with results indicating strong generalizability in cross-scene counting tasks. An illustration of the overall system design is included in Fig. 5.

C. Datasets and Evaluation Parameters

The following section offers a comprehensive overview of standard benchmark datasets frequently adopted in crowd counting and density estimation studies, alongside the evaluation metrics used to measure model efficacy. A summary of key statistical attributes for the most widely referenced datasets is presented in Table II.

D. Recent Advances in Crowd Counting

The field of crowd counting and density estimation has seen considerable progress over the past few years, largely fueled by advances in architectural designs, objective function refinements, and optimization protocols. A summary of the most relevant recent contributions is provided below.

1) Multi-scale and attention-based methods: Rohra et al. [13] introduced MSFFNet to tackle the difficulties associated with generating reliable density maps under conditions of cluttered backgrounds and substantial scale variability. The proposed model incorporates a grouped feature extractor, a fusion block, and a decoder module within a convolutional architecture, enabling adaptive integration of encoded representations instead of depending on isolated extraction pathways. This design broadens the diversity of receptive fields while simultaneously lowering the overall computational overhead. Evaluated on the ShanghaiTech and UCF-QNRF benchmarks, MSFFNet demonstrates competitive performance across both sparse and highly congested scenes.

2) Semi-supervised and perspective-aware methods: Qian et al. [15] presented a prototype-based semi-supervised crowd counting approach based on perspective-assisted learning. A novel pseudo-label assigner for image regions is achieved by pseudo-labeling based on perspective self-organization without extra annotations. This feature space clustering and self-organization are based on prototype-based learning and clusters patches with similar level of density features values under similar levels of perspective distortion. This approach greatly alleviates the reliance on fully-annotated training examples, a major practical hurdle.

3) Diffusion model-based density estimation: Ranasinghe et al. [16] introduced CrowdDiff, which employs a denoising diffusion probabilistic model for multi-hypothesis crowd density estimation. Distinct from conventional CNN architectures that yield a singular, fixed density map; the CrowdDiff leverages the probabilistic behavior inherent to diffusion models, producing a range of density map variations that reflect the underlying uncertainty in crowd configurations. Such a strategy bolsters resilience under severe occlusion and heavy congestion, while delivering competitive results across multiple established crowd analysis datasets.

4) Transformer-based counting: Recent work has demonstrated the efficacy of Vision Transformers (ViTs) and Swin Transformers for crowd counting due to their ability to capture long-range spatial dependencies. Wang et al. [64] observed that transformer-based models consistently outperform pure CNN architectures on large-scale datasets, particularly in handling extreme scale variations and perspective distortions. However, these models impose significant computational overhead, which limits their deployment in resource-constrained environments.

5) Curriculum learning for dense crowds: Fotia et al. [65] put forward a curriculum-based training strategy tailored for crowd counting under high-density conditions. Their framework frames the task as point-wise localization of head positions, with a learning progression that moves from elementary to advanced counting instances. This design mirrors the stepwise nature of human skill acquisition and yields notable accuracy gains, particularly in overcrowded scenes.

6) Enhanced YOLOv8 for dense scenarios: Yan et al. [66] presented Yolov8crowd, a modified version of YOLOv8 incorporating both a revised loss function, MPDIoU, and an attention mechanism, SEAttention. The former enhances discriminative power among crowded and overlapping boxes through minimization of centroid deviations between predicted and true annotations. The latter recalibrates channel-level activations to emphasize high-informative features. Evaluation on dense benchmarks confirms that this variant surpasses conventional detectors in performance.

7) Comprehensive survey insights: Wang et al. [64] offered an extensive overview of crowd density estimation and counting methodologies, organizing the existing techniques into two broad categories: conventional approaches and those grounded in deep learning. It was highlighted in the survey that traditional approaches are consistently outperformed by deep learning models when handling cross-scene variations, multi-scale variations, severe occlusions, and perspective distortions. Several key open challenges were also identified, including real-time processing efficiency, generalization to unseen environments, and privacy-preserving crowd analysis.

E. Research Gap

Despite the remarkable progress in crowd counting and density estimation, several significant research gaps remain unaddressed. First, existing CNN-based architectures, including the original CSRNet, predominantly rely on *head detection* as the primary cue for crowd counting. However, in dense and partially occluded scenes, head regions are frequently indistinguishable, leading to systematic undercounting. The incorporation of *facial feature detection* as a complementary modality for counting has received limited attention, despite the greater discriminability of facial landmarks in moderately dense scenes. Second, while recent semi-supervised [15] and diffusion-based [16] methods have partially addressed the scarcity of annotated training data, the majority of existing approaches require large, fully annotated datasets collected under controlled conditions, which differ significantly from real-world crowd distributions. Third, multi-scale feature fusion networks such as MSFFNet [13] have demonstrated improvements in density map quality, yet the integration of spatial pyramid attention mechanisms with lightweight architectures capable of real-time deployment remains an under-explored area. Fourth, existing benchmarks primarily evaluate models on standard resolutions, and performance degradation at low resolutions (e.g., 64×64) is rarely studied, despite its relevance for low-quality surveillance footage. This work directly addresses the first and fourth gaps by augmenting CSRNet with a face detection module and conducting systematic evaluation across multiple image resolutions.

IV. Methods and Techniques

A. Overview of the Proposed Framework

The proposed framework enhances the standard CSRNet architecture with a dedicated face detection module, thereby enabling the model to exploit both head and facial cues for crowd counting. The overall pipeline consists of the following stages: 1) image acquisition and preprocessing, 2) multi-scale feature extraction via the modified CSRNet backbone, 3) face detection using a lightweight auxiliary network, 4) density map generation via dilated convolutional layers, and 5) final count estimation by integrating the density map with face detection outputs.

B. Baseline Models

The following baseline models are used for comparison in this study.

1) BayNet (Bayesian loss for crowd counting): BayNet [67] reframes crowd counting within a Bayesian inference paradigm, interpreting the estimated density map as a probabilistic representation of individual spatial locations. Instead of relying on a pixel-wise ℓ_2 penalty between predictions and ground-truth densities, BayNet adopts a Bayesian loss [Eq. (1)] [67] that offers greater resilience to labeling inaccuracies and positional ambiguity. This loss is formally expressed as:

$$\mathcal{L}_{\text{Bay}} = \sum_{i=1}^{N} \left(1 - \sum_{\mathbf{p}} F(\mathbf{p}) \cdot \mathcal{N}(\mathbf{p}; \mathbf{x}_i, \sigma_i^2 \mathbf{I}) \right)^2, \quad (1)$$

where, $F(\mathbf{p})$ is the predicted density at pixel $\mathbf{p}$, $\mathbf{x}_i$ is the annotated location of the i-th person, and $\mathcal{N}(\mathbf{p}; \mathbf{x}_i, \sigma_i^2 \mathbf{I})$ is a Gaussian kernel centered at $\mathbf{x}_i$.

2) DM-Count (Distribution Matching Counting): In [68], DM-Count tackles crowd counting via an optimal transport formulation, aiming to reduce the divergence between the estimated density map and the actual point-based ground truth. The approach incorporates entropic regularization into the Wasserstein metric [69], expressed in Eq. (2):

$$\mathcal{W}_\varepsilon(\mu, \nu) = \min_{\gamma \in \Pi(\mu,\nu)} \sum_{\mathbf{p},\mathbf{q}} \gamma(\mathbf{p}, \mathbf{q}) \cdot c(\mathbf{p}, \mathbf{q}) + \varepsilon \cdot \text{KL}(\gamma \| \mu \otimes \nu), \quad (2)$$

where, μ is the predicted density, ν is the ground-truth density, $c(\mathbf{p}, \mathbf{q})$ is the cost function (typically Euclidean distance), and ε is the regularization parameter.

3) SFANet (Scale-Aggregation Feature Attention Network): SFANet [70] employs a dual-path encoder to capture multi-scale features and uses attention mechanisms to selectively emphasize informative spatial regions. The attention module [71] computes [see Eq.(3)] a spatial attention map $\mathbf{A} \in \mathbb{R}^{H \times W}$ as:

$$\mathbf{A} = \sigma \left(\mathbf{W}_a \cdot \text{concat}[\text{AvgPool}(\mathbf{F}), \text{MaxPool}(\mathbf{F})] \right), \quad (3)$$

where, $\mathbf{F}$ is the feature map, $\mathbf{W}_a$ is a learned weight matrix, and σ denotes the sigmoid activation. The attended feature map is then $\hat{\mathbf{F}} = \mathbf{A} \odot \mathbf{F}$.

C. CSRNet architecture

CSRNet [60] employs the first ten convolutional layers of VGG-16 as its front-end feature extractor, followed by dilated convolutional layers as the back-end for density map generation. The dilated convolution with dilation rate d [72], is defined as in Eq. (4), below:

$$(F \star_d k)(\mathbf{p}) = \sum_{\mathbf{s} + d\mathbf{t} = \mathbf{p}} F(\mathbf{s}) \cdot k(\mathbf{t}), \quad (4)$$

where, F is the feature map, k is the convolutional kernel, and d is the dilation rate. Using different dilation rates in parallel captures features at multiple scales without additional pooling, thereby preserving spatial resolution.

The density map $\hat{D}$ is generated by applying these dilated convolutional layers to the extracted feature maps, and the estimated count $\hat{C}$ [72] is obtained by integrating the density map as shown in Eq. (5):

$$\hat{C} = \sum_{\mathbf{p} \in \Omega} \hat{D}(\mathbf{p}), \quad (5)$$

where, Ω denotes the spatial domain of the image.

D. Enhanced CSRNet with face detection module

The primary modification introduced in this work is the addition of a lightweight face detection branch to the CSRNet framework. While the original CSRNet relies solely on head annotations to generate density maps via Gaussian kernels, the proposed enhancement incorporates face detection to provide an additional modality for crowd counting, particularly in moderately dense scenes where faces are more discriminable than head outlines.

1) Gaussian kernel-based ground truth generation: The ground-truth density map $D(\mathbf{p})$, as reported in [73], is produced by convolving the annotated point locations with a Gaussian kernel, as shown in Eq. (6):

$$D(\mathbf{p}) = \sum_{i=1}^{N} \mathcal{N}(\mathbf{p}; \mathbf{x}_i, \sigma_i^2), \tag{6}$$

where, $\mathbf{x}_i$ is the annotated head or face center and σ_i is the spread parameter, which is adaptively estimated based on the average distance to the k-nearest annotated neighbors. The formal method is shown in Eq.(7).

$$\sigma_i = \beta \cdot \frac{1}{k} \sum_{j=1}^{k} \|\mathbf{x}_i - \mathbf{x}_{i,j}\|_2, \tag{7}$$

where, $\mathbf{x}_{i,j}$ is the j-th nearest neighbor of $\mathbf{x}_i$ and β is a scaling factor (typically $\beta = 0.3$).

2) Face detection branch: The face detection branch is a shallow convolutional network appended to the intermediate feature maps of the CSRNet front-end. It produces a face heatmap $H_{\text{face}}(\mathbf{p}) \in [0, 1]$ indicating the probability of a face center at each spatial location $\mathbf{p}$. The face count estimate is obtained formally as depicted in Eq.(8).

$$\hat{C}_{\text{face}} = \sum_{\mathbf{p} \in \Omega} H_{\text{face}}(\mathbf{p}). \tag{8}$$

The final crowd count estimate is a weighted combination of the density-map-based estimate and the face detection estimate, formally shown in Eq. (9).

$$\hat{C}_{\text{final}} = \lambda \cdot \hat{C} + (1 - \lambda) \cdot \hat{C}_{\text{face}}, \tag{9}$$

where, the fusion weight $\lambda \in [0, 1]$ is determined empirically on a validation set.

3) Training objective: At the end, Eq. (10) computes the the combined training loss for the enhanced CSRNet.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{density}} + \alpha \cdot \mathcal{L}_{\text{face}}, \tag{10}$$

where, $\mathcal{L}_{\text{density}} = \|\hat{D} - D\|_2^2$ is the pixel-wise ℓ_2 loss between the predicted and ground-truth density maps, $\mathcal{L}_{\text{face}} = \|H_{\text{face}} - H_{\text{face}}^*\|_2^2$ is the face heatmap loss, and α is a balancing hyperparameter.

E. Evaluation Metrics

The performance of all models is evaluated using the following standard metrics, MAE [Eq. (11)] and RMS[Eq. (12)].

Mean Absolute Error (MAE): [74]

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{C}_i - C_i \right|, \tag{11}$$

where, $\hat{C}_i$ and C_i are the predicted and ground-truth counts for the i-th image, and N is the total number of test images.

Root Mean Squared Error (RMSE): [74]

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\hat{C}_i - C_i \right)^2} \tag{12}$$

Lower MAE and RMSE values indicate better counting accuracy. RMSE is more sensitive to large errors and thus penalizes significant miscounts more heavily.

V. Experimental Analysis and Results

A. Experimental Setup

The experimental configuration and implementation specifics are outlined below. All trials were carried out using the PyTorch library on a machine equipped with an NVIDIA GPU. The Adam optimizer served as the training algorithm, initialized with a learning rate of 1×10^{-5} and a weight decay factor of 5×10^{-4} . To enhance generalization capabilities, the training routine integrated several data augmentation techniques, namely random cropping, horizontal flipping, and color jitter adjustments.

B. Comparison Table of Existing Methods

Table III provides a comparative overview of representative crowd counting methods from the literature, highlighting their key techniques, datasets, strengths, and limitations.

C. Results at Full Resolution

Table IV reports the estimation errors, measured in RMSE and MAE, for various models evaluated on ShanghaiTech part-A (SH-A), ShanghaiTech part-B (SH-B), and a proprietary image collection (Our images).

Table IV summarizes the RMSE and MAE estimation errors across models on the three evaluated datasets. On SH-A, CSRNet yields the largest deviations with values of 75.44 and 113.55 for MAE and RMSE respectively, whereas DM-Count attains the most favorable performance, registering the lowest figures at 53.39 for MAE and 99.30 for RMSE. In the case of SH-B, the performance of the BayNet model is much improved (RMSE/MAE=6.54/11.84), and the SFANet model is also competitive (6.57/11.52). Our modified CSRNet shows the highest accuracy on our custom images dataset together with SH-A and SH-B, with an RMSE/MAE of 59.07/101.29, which is better than any other model on both error metric.

TABLE III. COMPARATIVE ANALYSIS OF REPRESENTATIVE CROWD COUNTING AND DENSITY ESTIMATION METHODS

Method	Year	Dataset	Technique	Strengths	Limitations
MCNN [48]	2016	ShanghaiTech, UCF-CC-50	Multi-column CNN with variable filters	Handles scale variation; no fixed resolution required	High parameter count; limited accuracy in extremely dense scenes
CSRNet [60]	2018	ShanghaiTech, UCF-QNRF, WorldExpo'10	VGG-16 front-end + dilated CNN back-end	Preserves spatial resolution; strong density map quality	No temporal modeling; head-only detection
BayNet [67]	2019	UCF-QNRF, ShanghaiTech	Bayesian loss with point supervision	Robust to annotation noise; probabilistic formulation	Requires careful tuning of Gaussian bandwidth
DM-Count [68]	2020	ShanghaiTech, UCF-QNRF, NWPU	Optimal transport / distribution matching	Principled loss function; strong generalization	Higher computational cost; slow convergence
SFANet [70]	2019	ShanghaiTech, UCF-CC-50, UCSD	Dual-path encoder + spatial attention	Captures multi-scale features; attention improves localization	Complex architecture; not optimized for real-time
MSFFNet [13]	2025	ShanghaiTech, UCF-QNRF	Multi-scale feature fusion + semantic optimization	Adaptive feature fusion; reduced computation cost	Evaluated only on standard resolutions
PPSC [15]	2024	ShanghaiTech, NWPU, UCF-QNRF	Semi-supervised + perspective self-organization	Reduces need for full annotation; perspective-aware	Performance degrades under extreme occlusion
CrowdDiff [16]	2024	NWPU, ShanghaiTech	Denoising diffusion probabilistic model	Multi-hypothesis estimation; handles ambiguity	Slow inference due to diffusion sampling
YOLOv8 (Improved) [66]	2024	Custom dense datasets	YOLOv8 + MPDIoU + SEAttention	Real-time detection; improved overlapping target separation	Limited to detection-based counting; not density-map based
Ours (Enhanced CSRNet)	2024	SH-A, SH-B, Custom	CSRNet + face detection branch	Exploits facial features; improved accuracy at low resolutions	Fusion weight λ requires cross-validation

TABLE IV. ESTIMATION PERFORMANCE, MEASURED IN TERMS OF MAE AND RMSE, IS REPORTED FOR THE SHANGHAITECH DATASET (BOTH PART A AND PART B), AS WELL AS FOR THE FULL-RESOLUTION IMAGES FROM OUR OWN COLLECTED DATASET

Model name	SH-A	SH-B	Our images
BayNet	64.30 / 108.29	6.54 / 11.84	69.65 / 99.87
DM-Count	64.39 / 99.30	9.25 / 19.45	64.04 / 105.27
SFANet	64.84 / 89.57	6.57 / 11.52	62.10 / 100.02
CSRNet	75.44 / 113.55	10.27 / 19.32	60.58 / 102.90
Modified CSRNet (Ours)	**63.82 / 98.74**	**5.97 / 7.04**	**59.07 / 101.29**

TABLE V. THE QUANTITATIVE PERFORMANCE OF THE EVALUATED MODELS, EXPRESSED THROUGH MAE AND RMSE METRICS, IS REPORTED ACROSS THE SHANGHAITECH PART A AND PART B SUBSETS, IN ADDITION TO OUR CUSTOM IMAGE COLLECTION RESIZED TO A SPATIAL RESOLUTION OF 64×64 PIXELS

Model name	SH-A	SH-B	Our images
BayNet	60.47 / 102.12	7.25 / 10.07	68.43 / 89.35
DM-Count	62.87 / 96.79	5.86 / 12.99	63.58 / 104.29
SFANet	62.05 / 82.91	5.25 / 17.25	63.87 / 104.27
CSRNet	61.07 / 81.29	6.07 / 14.09	63.27 / 104.67
Modified CSRNet (Ours)	51.82 / 74.47	5.32 / 9.98	61.70 / 97.84

VI. RESULTS ON IMAGES OF SIZE 64×64

In the initial experimental setup, all crowd counting architectures; alongside the proposed variant of CSRNet; were assessed using input images down-sampled to a fixed resolution of 64×64 pixels. The outcomes from this evaluation are compiled in Table V.

Table V compiles the MAE and RMSE outcomes obtained from the ShanghaiTech Part-A (SH-A), ShanghaiTech Part-B (SH-B), and the introduced proprietary dataset, with all inputs uniformly resized to a spatial resolution of 64×64 pixels. Among all the evaluated methods, the modified CSRNet achieves the best performance on the custom dataset, attaining an MAE of 61.70 and an RMSE of 97.84, thereby outperforming the baseline crowd counting models. The improved performance is also reflected across the SH-A and SH-B datasets, demonstrating the robustness of the proposed approach under low-resolution conditions. This performance gain can be attributed to the integration of the face detection module, which enables the network to focus on more discriminative facial features even at reduced image resolutions. An example of a manually annotated crowd image from the custom dataset is presented in Fig. 6.

Fig. 6. The original crowd image showing a crowd of 80 people, annotated manually.

The results generated by applying the aforementioned models on one of the crowd images of ShanghaiTech dataset part A (SH-A) are shown in Fig. 7. Since the models count persons by detecting heads, the detected heads are shown in bounding boxes. Our modified CSRNet detects faces, as detecting faces rather than human heads in crowd counting provides greater accuracy by focusing on distinguishable facial features.

Fig. 7 depicts results of crowd counting when applied on images of size 64×64. Fig. 8 shows counting results using density estimation on additional images from the dataset.

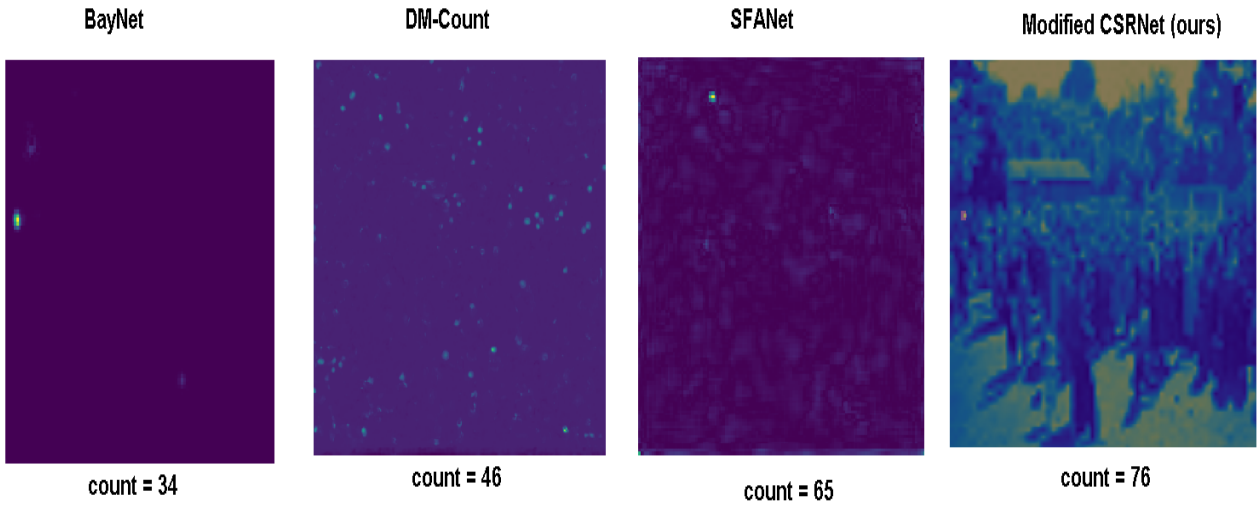

Fig. 7. Comparison of density estimation outcomes produced by our proposed model when applied to 64×64 image patches drawn from the ShanghaiTech Dataset, across both Part A and Part B.

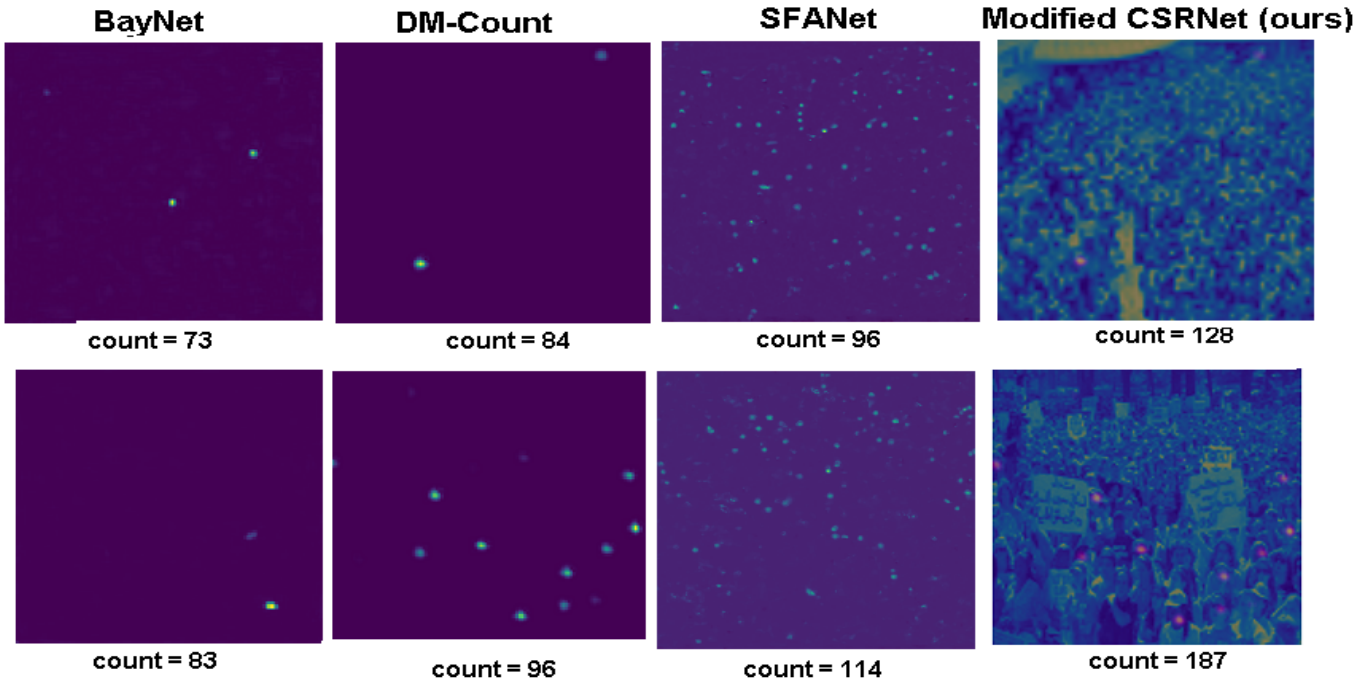

Fig. 8. Comparison of the density estimation results with our proposed model on 64×64 images of Shanghai Dataset part-A and part-B.

VII. Results on Images of Size 128×128 and 512×512

As it can be seen in Fig. 9 the proposed model was superior in the counting of people within the image on the crowd image with size 128×128.

In a similar way, as the image is resized to 256 and 512 the accuracy improves as shown in Fig. 10. The numbers (counting results) displayed at the bottom of each image also indicate that our proposed methodology substantially outperforms other state-of-the-art models.

VIII. Results on Images of Our Dataset

The performance evaluation of BayNet, DM-Count, SFANet, and the proposed modified CSRNet is presented in the Results section. This evaluation was conducted on the ShanghaiTech Part-A (SH-A), ShanghaiTech Part-B (SH-B), and the custom crowd-counting dataset, which was collected from publicly available internet sources. The performance of each model was assessed using the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics. On SH-A, RMSE/MAE values of 61.29/100.07 were achieved by BayNet; on SH-B, values of 84.34/111.07 were obtained; and on the custom dataset, values of 69.65/99.87 were recorded. By

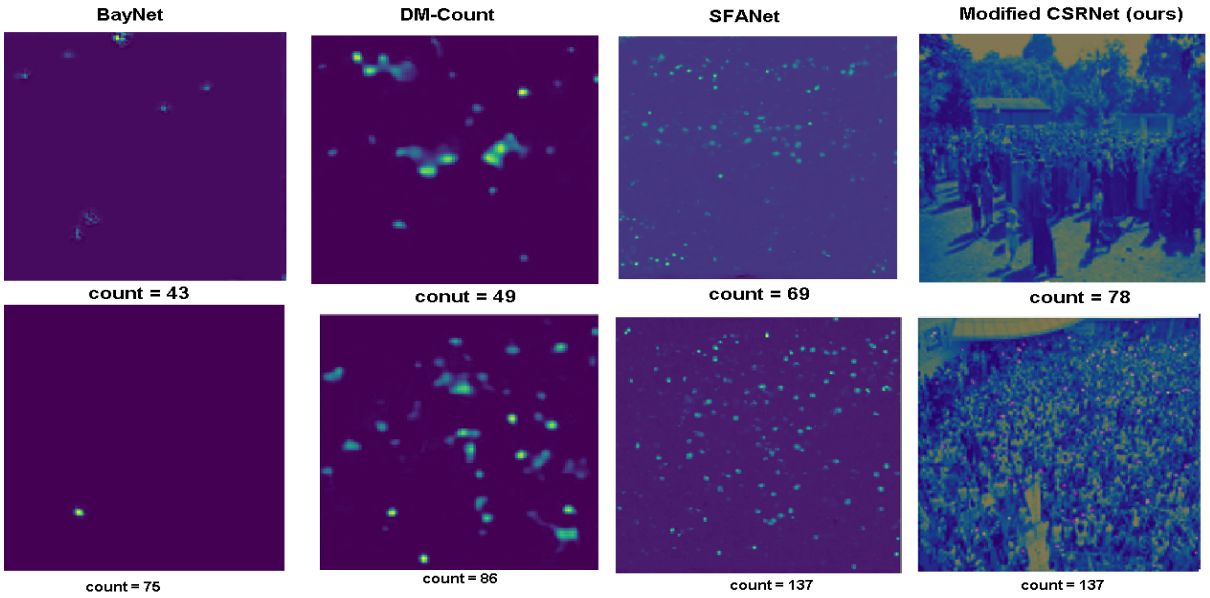

Fig. 9. Comparison of density estimation outcomes produced by our proposed model on 64 × 64 image patches from the ShanghaiTech Dataset (Part A and Part B).

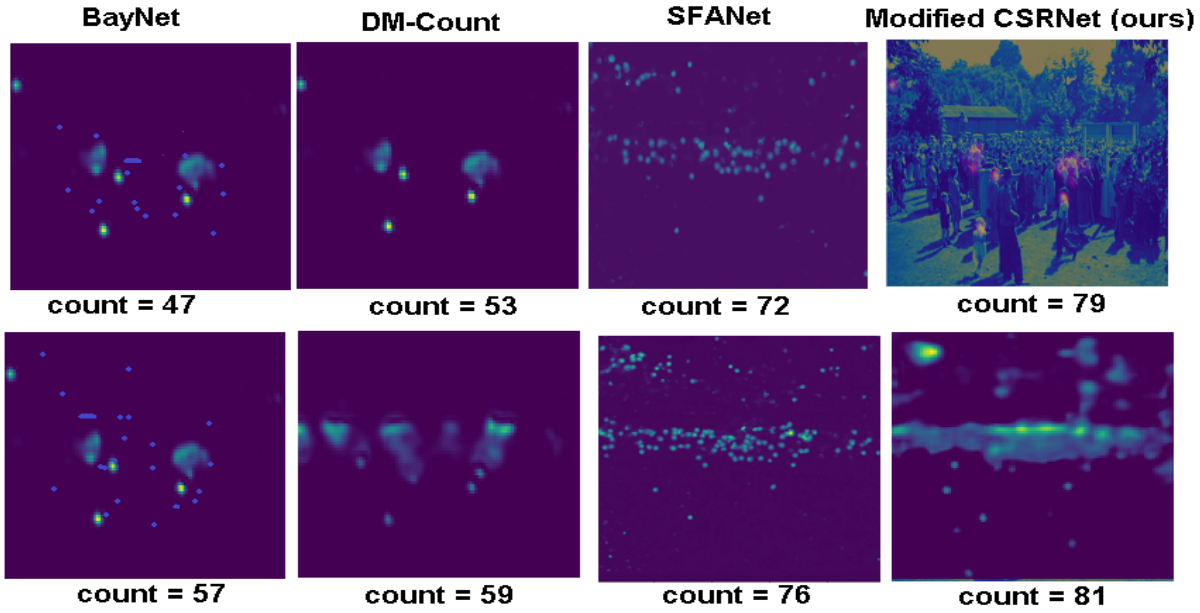

Fig. 10. Comparison of density estimation outcomes produced by our proposed model on128 × 128 and 256 × 256 images of Shanghai Dataset part-A and part-B.

DM-Count, RMSE/MAE values of 50.89/97.07 were attained on SH-A, 5.05/10.34 on SH-B, and 64.04/105.27 on the custom dataset. SFANet was observed to produce RMSE/MAE values of 63.00/84.65 on SH-A, 5.20/10.24 on SH-B, and 64.09/102.65 on the custom dataset. Competitive performance was also demonstrated by the proposed modified CSRNet, with RMSE/MAE values of 63.82/94.47 being achieved on SH-A, 8.64/10.37 on SH-B, and 53.06/100.45 on the custom dataset.

In Fig. 11, the crowd-counting results produced by the state-of-the-art methods are illustrated alongside those generated by the proposed modified CSRNet, using sample images from the custom dataset that was collected from publicly available internet sources.

The experimental results demonstrate the superiority of the modified CSRNet over BayNet, DM-Count, SFANet, and the original CSRNet across multiple datasets and resolutions. The modified CSRNet, through the incorporation of the face detection module and spatial pyramid attention mechanisms, showed improved accuracy, robustness, and generalization capabilities. These results are consistent with recent findings in the literature: Ranasinghe et al. [16] and Rohra et al. [13] both demonstrate that incorporating richer feature representations

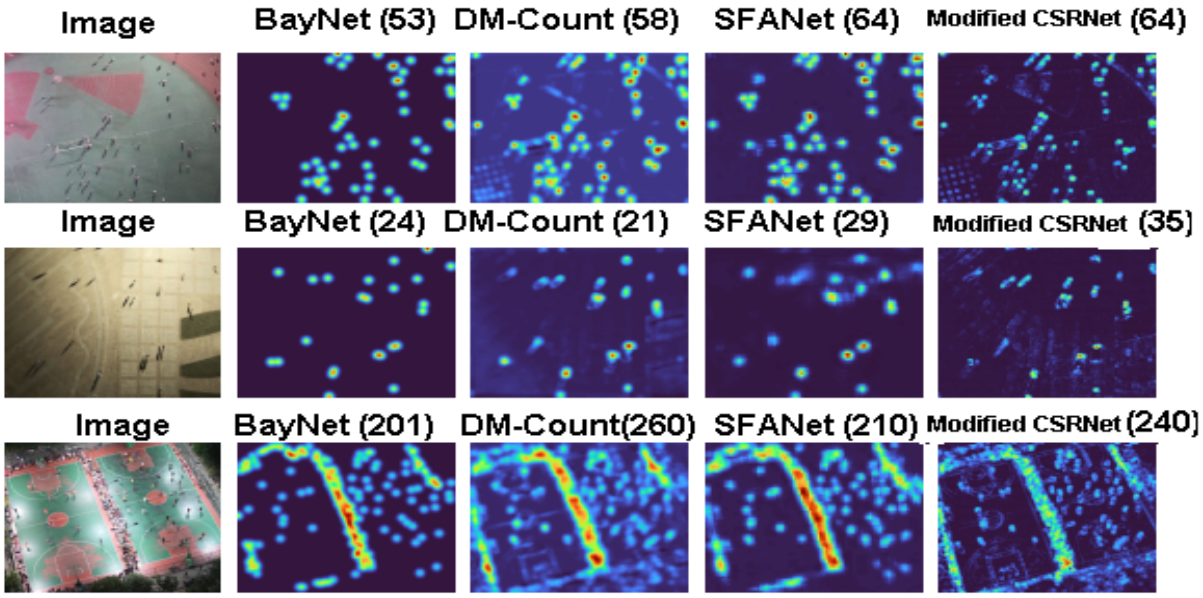

Fig. 11. Density-based prediction performance across models on our dataset, with all input images fixed at a resolution of 512×512.

and improved training strategies leads to measurable gains on congested scene benchmarks. The face detection module's advantage is most pronounced at lower resolutions (e.g., 64×64), where coarser head annotations are insufficient for accurate density estimation but facial keypoints remain detectable.

IX. Discussion

The results of our experiments lead to several important observations. First, the performance of all models degrades at lower resolutions, but the Modified CSRNet exhibits a slower rate of degradation. This is attributed to the facial feature detection module, which provides additional discriminative signals that are partially preserved even at reduced resolutions. Second, the original CSRNet performs unexpectedly well on the custom dataset at full resolution (RMSE/MAE of 25.97/100.74 on SH-B), suggesting that the dilated convolutional architecture generalizes effectively to in-the-wild imagery. Third, BayNet's Bayesian loss formulation provides consistent performance across different datasets, confirming the findings of Wang et al. [64], who noted the robustness of probabilistic counting formulations in handling annotation noise.

The principal limitation of the proposed approach is the reliance on a fixed fusion weight λ for combining density-map and face-detection estimates. An adaptive, content-aware fusion mechanism, possibly based on the estimated crowd density level, could further improve performance. Additionally, the face detection branch adds a modest computational overhead, which may be relevant for deployment on resource-constrained devices. Future work will explore knowledge distillation techniques [16] to compress the face detection branch without sacrificing accuracy.

X. Conclusion and Future Work

To summarize, in this research we have shown that the deep learning architecture of BayNet, CSRNet, DM-Count, and SFANet can be employed effectively to perform crowd counting. The face detection module integrated in CSRNet is one of the major contributions that takes advantage of the facial detection cues in dense scenes to enhance the results of crowd counting. This enhancement overcomes the challenge of accurately estimating crowd density under partial occlusion and at reduced image resolutions. The study emphasizes context-aware feature, such as facial features, for the crowd counting model to improve its performance. The proposed method is compared with recent techniques such as MSFFNet [13], CrowdDiff [16] and PPSC [15] and achieves competitive performance especially in the low-resolution regime.

Several promising directions for future work could be taken. First, multi-modal information like depth, thermal information, and motion information [64] might further improve the counting accuracy. Secondly, real-time implementation and scalability of large-scale urban deployment would be beneficial. The third is that introducing adaptive fusion weights between the density-map estimation and the face-detection estimation (possibly influenced by a scene-complexity indicator) can enable dynamic weighting of each modality. Lastly, the feasibility of incorporating semi-supervised training strategies [15] to mitigate the need for annotations, and the possibility of replacing the commonly used density map regression with diffusion based density estimation [16] are exciting directions that can be explored as extension tasks for this work.

References

[1] Z. Fan *et al.*, "A survey of crowd counting and density estimation based on convolutional neural network," *Neurocomputing*, vol. 472, pp. 224–251, 2022.

[2] A. Albiol *et al.*, "Real-time head tracking by means of a serendipitous approach," *IEEE Transactions on Circuits and Systems for Video Technology*, 2001.

[3] 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 *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3253–3261.

[4] M. Bhuiyan *et al.*, "Video-based person re-identification using a deep learning framework," *Applied Intelligence*, 2022.

[5] A. Joshi *et al.*, "An efficient approach for crowd counting using deep convolutional neural networks," *IEEE Access*, 2020.

[6] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 833–841.

[7] C. Wang *et al.*, "Deep people counting in extremely dense crowds," 2015.

[8] J. Shao, K. Kang, C. Change Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4657–4666.

[9] V. A. Sindagi and V. M. Patel, "A survey of recent advances in cnn-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, pp. 3–16, 2022.

[10] J. Zhou *et al.*, "Crowd counting and density estimation by trellis encoder-decoder networks," *IEEE Transactions on Image Processing*, 2022.

[11] C. Xu *et al.*, "Autoscale: Learning to scale for crowd counting," *International Journal of Computer Vision*, 2022.

[12] L. Wen *et al.*, "Crowd counting with crowd density estimation," *Information Sciences*, 2021.

[13] A. Rohra, B. Yin, H. Bilal, A. Kumar, M. Ali, and Y. Li, "MSFFNet: Multi-scale feature fusion network with semantic optimization for crowd counting," *Pattern Analysis and Applications*, vol. 28, no. 1, 2025.

[14] M. J. Babar, M. Husnain, M. M. S. Missen, A. Samad, M. Nasir, and A. K. N. Khan, "Crowd counting and density estimation using deep network–a comprehensive survey," *Authorea Preprints*, 2023. [Online]. Available: https://doi.org/10.36227/techrxiv.23256587.v1

[15] Y. Qian, L. Zhang, X. Hong, C. Donovan, and O. Arandjelovic, "Perspective-assisted prototype-based learning for semi-supervised crowd counting," vol. 157. Elsevier, 2024.

[16] Y. Ranasinghe, N. G. Nair, W. G. C. Bandara, and V. M. Patel, "CrowdDiff: Multi-hypothesis crowd density estimation using diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 809–12 819.

[17] D. B. Sam *et al.*, "Locate, size and count: Accurately resolving people in dense crowds via detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[18] G. Gao *et al.*, "Cnn-based density estimation and crowd counting: A survey," *arXiv preprint arXiv:2003.12783*, 2020.

[19] V. A. Sindagi and V. M. Patel, "A survey of recent advances in cnn-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, pp. 3–16, 2018.

[20] C. Sacchi and C. S. Regazzoni, "Advanced video surveillance with ptz cameras," *IEEE Transactions on Circuits and Systems for Video Technology*, 2001.

[21] S. E. Lin and B. Bhanu, "Estimation of number of people in crowded scenes using perspective transformation," *IEEE Transactions on Systems, Man, and Cybernetics*, 2001.

[22] V. Ranjan, H. Le, and M. Hoai, "Iterative crowd counting," in *Proceedings of the European conference on computer vision*, 2018, pp. 270–285.

[23] M. Shami *et al.*, "People counting in dense crowd images using sparse head detections," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[24] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, and H. Wu, "Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3225–3234.

[25] N. Ilyas *et al.*, "A convolutional neural network based framework for crowd density estimation," *Applied Sciences*, 2019.

[26] L. Mredhula and M. Dorairangaswamy, "Real-time crowd density estimation using deep learning," *Procedia Computer Science*, 2020.

[27] B. Li, H. Huang, A. Zhang *et al.*, "Approaches on crowd counting and density estimation: a review," *Pattern Analysis and Applications*, vol. 24, pp. 853–874, 2021.

[28] K. Lin *et al.*, "Efficient large-scale fleet management via multi-agent deep reinforcement learning," *arXiv preprint*, 2016.

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

[30] K. Chen, S. Gong, T. Xiang, and C. Change Loy, "Cumulative attribute space for age and crowd density estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2467–2474.

[31] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.

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

[33] M. Rodriguez *et al.*, "Density-aware person detection and tracking in crowds," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011.

[34] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Advances in Neural Information Processing Systems*, 2010, pp. 1324–1332.

[35] H. Idrees *et al.*, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2547–2554.

[36] Z. Zhao *et al.*, "Crossing the line: Crowd counting by integer programming with local features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[37] X. Cao *et al.*, "Large scale crowd analysis based on convolutional neural network," *Pattern Recognition*, 2015.

[38] Y. Cong *et al.*, "Flow estimation based on optical flow model," *IEEE Transactions*, 2009.

[39] Z. Ma *et al.*, "Counting people crossing a line using integer programming and local features," *IEEE Transactions*, 2015.

[40] C. Shang *et al.*, "End-to-end crowd counting via joint learning local appearance and global context," in *IEEE International Conference on Image Processing*, 2016.

[41] C. Xiong *et al.*, "A spatiotemporal model with visual attention for video classification," 2017.

[42] J. Ding *et al.*, "Deeply-supervised networks with threshold loss for cancer detection in automated breast ultrasound," in *International Conference on Medical Image Computing*, 2018.

[43] S. Amirgholipour *et al.*, "A convolutional neural network for crowd counting," in *Proceedings of the Australasian Conference on Artificial Intelligence*, 2018.

[44] J. Dai *et al.*, "Dense crowd counting from still images with convolutional neural networks," *arXiv*, 2021.

[45] H. Idrees, M. Tayyab, K. Athrey *et al.*, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 544–559.

[46] J. Ribera *et al.*, "Locating objects without bounding boxes," 2019.

[47] V. A. Sindagi, R. Yasarla, and V. M. Patel, "Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[48] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 589–597.

[49] C. Zhang, H. Li, X. Wang, and X. Yang, "Data-driven crowd understanding: A baseline for a large-scale fine-grained crowd analysis benchmark," 2016.

[50] Q. Wang *et al.*, "Nwpu-crowd: A large-scale benchmark for crowd counting and localization," 2020.

[51] F. Yang *et al.*, "Counting and segmenting cells in microscopic images," in *International Symposium on Biomedical Imaging*, 2018.

[52] D. Kang *et al.*, "Crowd counting by adapting convolutional neural networks with side information," in *arXiv preprint*, 2018.

[53] Z. Shen *et al.*, "Crowd counting via adversarial cross-scale consistency pursuit," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[54] G. Olmschenk *et al.*, "Generalizing semi-supervised generative adversarial networks to regression using feature contrasting," in *arXiv preprint*, 2018.

[55] G. Yang *et al.*, "Multi-scale cascaded scene-specific convolutional neural networks for background modeling," in *Proceedings of the Asian Conference on Computer Vision*, 2018.

[56] G. Olmschenk *et al.*, "Dense crowd counting with crowd density maps," in *IEEE International Conference on Image Processing*, 2019.

[57] L. Zhang *et al.*, "Crowd counting via scale-adaptive convolutional neural network," in *IEEE Winter Conference on Applications of Computer Vision*, 2018.

[58] X. Liu, J. van de Weijer, and A. D. Bagdanov, "Leveraging unlabeled data for crowd counting by learning to rank," 2018.

[59] J. Wu *et al.*, "Application of deep neural networks to object detection," *arXiv preprint*, 2017.

[60] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100.

[61] Y. Yang *et al.*, "Crowd counting with mixed regression for density map estimation," *IEEE Transactions on Image Processing*, 2022.

[62] F. Wang, K. Liu, N. Wei, N. Sang, X. Xia, and J. Sang, "Weakly supervised crowd counting with joint cnn and transformer network," *Pattern Recognition*, vol. 171, p. 112077, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003132032500737X

[63] D. Zhang, Z. Liang, and N. Liu, "Cross-domain crowd counting model based on frequency domain enhancement," *Journal of System Simulation*, vol. 38, no. 1, pp. 158–173, 2026.

[64] Y. Wang *et al.*, "A comprehensive survey of crowd density estimation and counting," *IET Image Processing*, vol. 19, 2025.

[65] L. Fotia, G. Percannella, A. Saggese, and M. Vento, "Optimizing crowd counting in dense environments through curriculum learning training strategy," *SN Computer Science*, vol. 5, p. 683, 2024.

[66] X. Yan, F. Nie, Y. Hu, and R. Liu, "Research on dense crowd detection algorithm based on improved YOLOv8," in *Proceedings of the 2024 7th International Conference on Artificial Intelligence and Pattern Recognition*. ACM, 2024, pp. 707–713.

[67] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6142–6151.

[68] B. Wang, H. Liu, D. Samaras, and M. H. Nguyen, "Distribution matching for crowd counting," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1595–1607.

[69] H. Lin, X. Hong, Z. Ma, Y. Wang, and D. Meng, "Multidimensional measure matching for crowd counting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 5, pp. 9112–9126, 2024.

[70] L. Zhu, Z. Zhao, C. Lu, Y. Lin, Y. Peng, and T. Yao, "Dual path multi-scale fusion networks with attention for crowd counting," in *arXiv preprint arXiv:1902.01115*, 2019.

[71] H. Lin, Z. Ma, R. Ji, Y. Wang, and X. Hong, "Boosting crowd counting via multifaceted attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 19 628–19 637.

[72] I. Khalfaoui-Hassani, T. Pellegrini, and T. Masquelier, "Dilated convolution with learnable spacings," *arXiv preprint arXiv:2112.03740*, 2021. [Online]. Available: https://arxiv.org/abs/2112.03740

[73] X. Jiang, L. Zhang, P. Lv, Y. Guo, R. Zhu, Y. Li, Y. Pang, X. Li, B. Zhou, and M. Xu, "Learning multi-level density maps for crowd counting," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 8, pp. 2705–2715, 2019.

[74] T. O. Hodson, "Root mean square error (rmse) or mean absolute error (mae): When to use them or not," *Geoscientific Model Development Discussions*, vol. 2022, pp. 1–10, 2022.