SJER Scientific Journal of Engineering Research

futuristech PT. Teknologi Futuristik Indonesia

Received: April 2, 2026; Revised: May 12, 2026; Accepted: May 15, 2026; Published: May 16, 2026; Current version: May 16, 2026. e-ISSN: 3109-1725
Digital Object Identifier 10.64539/sjer.v2i3.2026.494

Article

NRCC-LC: Noise-Robust Crowd Counting with Dynamic Label Correction under Noisy Supervision

Abubakar Abdinur Hersi[1,*], Miaogen Ling[1], Muhammad Raza[2], Abdirahman Mohamed Hassan[3], Idris Aweis Hussien[3]

[1] School of Computer Science, Nanjing University of Information Science & Technology, Nanjing, 210044, China; e-mail: arabkau89@gmail.com (A. A. Hersi).
[2] School of Computer and Software, Nanjing University of Information Science & Technology, Nanjing, 210044, China.
[3] School of Electronics and Communication Engineering, Nanjing University of Information Science & Technology, Nanjing, 210044, China.
* Correspondence Author

The authors received no financial support for the research, authorship, and/or publication of this article.

Abstract: Crowd counting remains a challenge within computer vision due to many factors that affect the performance of available methods such as occlusion, scale variability, and perspective distortion. Additionally, many labels associated with crowd counting systems have high levels of noise caused by various real-world conditions. Although crowd counting methodologies have improved accuracy over recent years, the majority of crowd counting models still rely on clean real-time supervision and lack systems that can correct for dynamically corrupted labels, resulting in low robustness for crowd counting models when deployed in real-world applications. In this work we present a Noise-Robust Crowd Counting with Label Correction (NRCC-LC) framework to obtain reliable density estimates from noisy supervision. To accomplish this, our approach uses a combined CNN-Transformer architecture to capture both locally- and globally-relevant visual information (i.e., image content and context), along with a Noise-Robust Module (NRM) and a Dynamic Label Correction (DLC) mechanism. Our principle experimental results evaluated across four benchmark datasets: ShanghaiTech Part A, ShanghaiTech Part B, NWPU-Crowd, and JHU-Crowd++, indicate that the NRCC-LC exhibits competitive performance with respect to existing state-of-the-art crowd-counting methods; most notably, producing per-image MAEs of 97.8 and 392.3 on NWPU-Crowd. These experimental results additionally have real-world implications for improving public safety and urban planning; thus, through our novel method of noise-aware feature learning combined with iterative label correction, we can establish the potential of automated monitoring systems in complex, real-world environments to be significantly more reliable.

Keywords: Crowd counting, Density estimation, Learning with noise, Transformer/CNN, Correcting labels, Teacher-student learning.

1. Introduction

The area of crowd counting is an important area of research within computer vision because of its many applications across a range of settings, including public safety, intelligent surveillance, transportation management, urban planning, and the monitoring of events [1]. Accurate crowd counting provides vital information to support decision-making in environments (e.g., railway stations, religious gatherings, stadiums, shopping areas, political demonstrations) where it is necessary to quantify the size of a crowd for planning purposes and to enhance public safety by minimizing the potential for incidents related to crowds [2]. In particular, accurate crowd counting plays an important role in crowded public spaces, where over-crowding can result in health and safety threats, including congestion, panic, stampedes, and business or emergency response failures. The study of crowd counting began with simple manual observation and traditional visual estimates of people through images or video frames, but these methods were typically too subjective (and

therefore unreliable) for use in crowds with high density because individuals may be partially occluded, heavily overlapped, or too small to be easily distinguished from one another [3].

Consequently, the first crowd counting research efforts used automated, model-based methods to identify and count individuals in real-time. Although the performance of these automated methods was acceptable in low-density locations, their performance decreased substantially at higher levels of crowd density (due to the increased difficulty of determining the visibility of individuals). The transition from traditional methods of crowd detection to regression- and density-based methods was a significant milestone in the evolution of crowd counting [3], [4]. Regression- and density-based methods can estimate total crowd size and predict a density map instead of detecting each individual and using convolution neural networks for feature extraction, and density estimation. Therefore, researchers can better capture highly-congested scenes. More recently, transformers and hybrid CNN-transformer approaches were introduced to further develop the idea of crowd counting by introducing global contextual reasoning while maintaining local spatial detail [5], [6]. Even with the proliferation of advanced techniques for crowd counting, real-world scenes are still challenging due to occlusion, scale change, perspective distortion, background noise, and annotation noise. As a result, the development of robust and reliable crowd counting algorithms remains a motivation for researchers in the field [7].

There have been considerable improvements in both the design of the architecture and the scale of the datasets, yet crowd counting continues to be a highly complex and difficult problem. The problem is not limited to the actual mathematical density estimation but also has to do with the multiple environmental and structural constraints that interact with each other. For instance, the majority of challenges has to do with occlusion. In addition, with respect to occlusion, many of the head regions are partially visible, while others may be completely occluded and there is also the issue of feature blurring and spatial mapping ambiguity. Even models based on a density approach may lose accuracy in heavily congested areas [7], [8]. Adding to the problem is that the scale of the scene can also vary greatly. Due to this type of foreshortening, there are instances in which two individuals could be standing next to each other in the same picture yet have heads of significantly different relative sizes. Further complicating the building of models that are robust to variations in scale is that there are more and more examples of perspective distortion. Specifically, theoretically, individuals who are closest to the camera will occupy the largest pixel area, while individuals who are located farther away will appear disproportionately smaller. Researchers will continue to pursue a solution to maintain scale invariance in designing their models [9].

One other major obstacle to successful crowd counting is having noisy annotations that are used to create the density maps. Most crowd counting datasets create density maps using manual annotations of head points, but these manual annotations can have missing labels, incorrect point locations, and inconsistencies due to human error. This noise in the annotations will affect the way the model optimizes and generalizes to new people being counted. Additionally, there are environmental issues such as changing light, motion blur on the camera, background clutter, and weather that make it even more difficult to learn how to make a good representation of a crowd. Therefore, the problem of creating crowd counting algorithms that can successfully learn when given noisy supervision is still a major area of research. via the manual annotation. While creating density maps relies on manually annotating the head point, these points could also have missing labels along with positional inaccuracies leading to an additional challenge when converting the noise into the generalization effect of density based upon statistics. In today's computational architectures, they are entirely transformer based, which also results in large compute requirements in terms of both speed and volume [6], [10].

This paper presents the NRCC-LC framework, which combines the features of hybrid CNN–Transformer architecture and enhances them through the addition of a noise-aware supervision refinement process, enabling improved training and learning with corrupted density maps. A Noise-Robust Module (NRM) is integrated into the framework to reduce the negative influence of unreliable feature responses and a Dynamic Label Correction (DLC) process is used to iteratively refine corrupted density labels during the training phase. Additionally, a teacher/student consistency approach will be applied to increase optimization stability and generalization capacity under conditions of noise.

To advance beyond the drawbacks related to density estimation models with dirty labels are presented, providing an NRCC-LC Framework: a model based on CNVCC for crowd counting capable of denoising density maps resulting from a supervised model; Feature Extraction Combiner: a novel CNN-Transformer combination model that both extracts local detail while modelling global context; Noise Robust Module (NRM): a network designed to eliminate unreliable features due to poor quality/dirty images prior to regression; Dynamic Label Correction (DLC): an iterative method to create improved labels through high confidence predictions made by the model; and Consistency Driven Optimisation: a student-teacher education process using ST-AR consistency to stabilise results of the model under ambiguous situations.

In this paper we will begin with: Section 2 discusses the methods of counting crowds by detection and density. Section 3 describes how the NRCC-LC was developed

along with its mathematical representations and base elements. Section 4 includes the procedures, quantitative analysis, and ablation characteristics of the four benchmark datasets. Lastly, Section 5 will summarize the paper and describe what research could be done in future studies.

2. Related Work

Originally, crowd counting focused on two main techniques: Detection and Density Estimation; which have both evolved into what we call Convolutional Neural Networks. Of late, our methods have also expanded to incorporate Transformer networks and noise-aware learning.

The idea behind estimating how many people are in an image has remained the same since early on, but our assumptions on how we arrive at that information have changed greatly. Initially, we viewed crowd counting as a simple detection task; however, it was quickly apparent that when people are present in high numbers, we cannot rely on just detecting each individual person to get an estimate of the total number of people. Instead, researchers have learned to use density estimation in order to estimate the total number of people in an image. In order for a crowd counting system to work well, it also needs to account for varying scale, scene context, supervision reliability, and annotation noise. We will look at the development of crowd counting from three perspectives: early foundational work, CNN-based advancements, and then finally to the most recent research.

2.1. Traditional Crowd Counting Approaches

In the early days of crowd counting, the objective was basically to be able to detect individuals in an image. The simple assumption was that if we were able to detect each individual, the total count could be generated by summing each of those detection responses.

The method worked well in the case where there were few individuals in an image, and they were usually very clear; however, when there are many individuals in an image, and there is significant overlap between individuals, it creates a situation for which the initial approach cannot be used [11], [12].

Chan, Liang, and Vasconcelos made significant early contributions to privacy-preserving crowd monitoring by proposing methods that do not require an explicit model of each person or any type of individual tracking [11]. They raised the foundational premise of whether counting could be done without necessarily detecting every individual. In other related studies, Chan and Vasconcelos proposed methods for counting by using low-level features of vision combined with Bayesian regression [12]. Collectively, these works demonstrated that when dealing with very dense populations, statistical estimates might be more feasible than directly detecting populations. Another major advance came from Lempitsky and Zisserman; their proposal of using density maps to represent the crowd and its corresponding population changed the direction of the area [3].

Rather than detecting individuals to obtain a count for them or predicting a scalar output only, they described the crowd as a continuous function of density generated from annotations of points within the image plane. The population numbers were calculated by integrating the function over the area of the image. This was very influential to the field, as it maintained the integrity of the spatial relationships without requiring each individual to be separated from others in the scene. Thus, their work demonstrated that density estimation would provide a more meaningful framework for estimating the number of people in the crowd than did direct detection or simple regression.

Seibert and Shah studied multi-source and multi-scale counting of crowded images [13]. Their research demonstrated that there is no way to model highly crowded scenes with just one feature type and one fixed scale. They have shown that the scale, as well as the complexity of the scene, is important for crowd-counting by using datasets from several different sources of data. They also developed methods for counting crowds through regression-based methods, which reduced the extreme dependence on exact location in highly dense scenes; but, as they only provided a scalar output, they lost all spatial information. By the end of this foundational period, the research community had all converged on three points regarding the estimation of crowds through direct detection in dense scenes: direct detection has significant reliability problems, scalar regression produces extremely limited results, and estimating crowds through density maps produces the highest level of accuracy relative to the level of spatial representation [3], [11]-[13].

2.2. CNN-Based Crowd Counting Methods

The introduction of deep learning and, in particular, convolutional neural networks (CNNs), has been a major step forward for the crowd-counting process. CNNs enabled crowd-counting to be done using hierarchical features learned from data rather than handcrafted descriptors. This is important to crowd-counting since repeated head-like structures, various textures, and vaguely defined local patterns exist in crowd scenes, making it challenging to manually create a representation of those crowds. Through end-to-end learning, CNNs have created very large increases in density estimation accuracy and robustness for the crowd-counting process [2], [14], [15].

C. Zhang, H. Li, X. Wang, and X. Yang were among the earliest authors to publish an influential paper based on CNNs, introducing cross-scene crowd counting with deep convolutional neural networks [14]. These authors focused on how CNN algorithms would provide generalisation between different scenes, rather than achieving

peak performance within a single fixed scene. Another major breakthrough came from Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, who proposed the Multi-column Convolutional Neural Network (MCNN) [15]. The MCNN architecture directly solved one of the foundational problems of this field, scale variation through the effects of perspective distortion. Through the use of several parallel convolutional branches with varying receptive fields, their model was able to respond to crowd structures with different apparent sizes within the same image. While the MCNN design is very important, the real significance lies in that they introduced a broader principle into crowd counting; that is, that multi-scale learning should be a fundamental design consideration when creating CNNs.

The CNN crowd counting literature has seen rapid growth in multiple directions. For example, Boominathan, Kruthiventi, and Babu presented CrowdNet [2]. In conjunction with this, Shang, Ai, and Bai also provided an approach that combined local and global counting information about the people in the crowd [16]. The importance of combining fine-grained visual detail to help understand the overall crowd scene adds to the value of these two papers. In a similar vein, Onoro-Rubio and López-Sastre demonstrated an approach to perform perspective-free object detection via deep learning [17].

Additionally, Walach and Wolf demonstrated that training strategies also affect the quality of counting when using CNN boosting methods [18]. However, one of the more significant contributions to Crowd Counting is through the work completed by Li, Zhang, and Chen with their work on CSRNet [8]. CSRNet employed a dilated convolution on a VGG front end, to obtain increased receptive field size while avoiding the heavy parallel architecture of MCNN. Therefore, it provided a more efficient method to contextualize reasoning about the density of individuals in the crowd, while also maintaining the details of the density map (through its use of dilated convolutions).

Finally, a multi-scale crowd counting architecture was presented in SANet [9]; the authors indicated that similar to other multi-scale methods, features should not only be extracted from multiple scales but should also be considered as a mean aggregation. As the number of papers on CNN-based crowd counting continues to increase, a greater emphasis will be placed on considering context and/or multi-level of supervision to improve overall counting performance. An example of this work is Liu, Salzmann, and Fua, who presented "Context-Aware Crowd Count [19].

Bayesian loss was originally introduced by Ma, Wei, Hong and Gong [20] to demonstrate the effect of uncertainty in point data, while Jiang *et al*. improved upon that with their work on density estimation, which relied on trellis encoder-decoder networks [21]. SFANet was subsequently proposed by Zhu *et al* [22]. Other works that contribute to the development of the CNN paradigm include those on compositional loss by Idrees *et al* [23], DSRM by Yao *et al* [24] and improvements for perspective awareness by Shi *et al* [25].

Data sets are increasingly gaining importance as well as they were used in the creation of the NWPU-Crowd benchmark for crowd counting and localization by Wang *et al* [7]. The introduction of the contextual pyramid CNN by Sindagi and Patel helps produce high quality density maps [26] and improved methods for locating, sizing and counting people in dense scenes were provided by Sam *et al* [27]. Thus, the establishment of the CNN era illustrated that deep learning has a significant impact on crowd counting when models used support multi-scale reasoning, have broader context and provide better supervision and that the methods that are CNN based remained fundamentally local in nature, which has impeded their abilities to model long range dependencies in complicated scenes [1], [2], [7]-[9], [14]-[28].

2.3. Transformer and Robust Hybrid Crowd Counting Methods

Crowd count research has entered an advanced stage that broadly involves three essential areas: the ability to reason specifically about global context, accurate evaluation of large crowds, and learning to do so despite imperfectly supervised data. The focus of research has changed significantly; it is no longer sufficient to simply improve accuracy of counts. Also of concern is how reliable a model will continue to be when exposed to extremely realistic complexity in the scenes being scanned; variation in the environments being scanned; etc. [4]-[6], [29]-[42].

One avenue of research aims at improving the way data is supervised and lost during counting. Liu *et al*. proposed an approach that utilized unlabeled data with rank-based learning to increase the density of counts [29]. Wang *et al*. built upon this work by developing a methodology called DM-Count, which used optimal transportation loss criteria to refine density measures [30]. Ma *et al.* continued this research utilizing a different, less frequent application of the optimal transport method of denoting density data [31]. These researchers demonstrate that advances in data collection techniques are not alone sufficient to create an accurate count. Equally the manner in which density (as determined by points) is supervised and optimized. Another key area of current research (with respect to both crowdedness and learning) relates to annotation noise/non-specificity and semi-supervised learning. Wan and Chan focused on explicitly modeling the existence of point noise in the context of counting large crowds [33]. They undoubtedly supported the assertion that point labels are not always representative of dense scenes being scanned. Meng *et al*. developed methods utilizing a spatial

configuration-aware methodology to estimate semi-supervised counts (as related to recognition of uncertainty) in crowds [34].

Finally, Li *et al*. focused on developing metrics to calibrate uncertainty levels of annotations in semi-supervised systems. [35] The point-to-region loss for semi-supervised point-based crowd counting has been introduced by Lin and colleagues in 2023. These works represent a significant change in how researchers define supervision: instead of assuming that all supervision is correct, recent research now treats supervision as uncertain and unreliable spatially. hybrid approaches, based on CNNs and transformers. The transformer model proposed in the work of Vaswani *et al*. in 2017's paper on self-attention [43] served as a basis for many of the recent vison-specific developments, including the following: the vision transformer (ViT) by Dosovitskiy *et al*. [44]; the swin transformer by Liu *et al*. [45]; and the training of data-efficient image transformers by Touvron *et al*. [10]. Several transformer-based crowd counting methods that have been developed include Gramformer by Lin *et al*. in 2023 [37]; CountFormer by Mo *et al*. in 2022 [38]; and an end-to-end transformer for crowd localization by Liang Xu and Bai in 2023 [39]. Performing scale-aware crowd counting using transformers is also on the rise with recent work by Jia *et al*.'s scale-aware transformer for crowd counting (STCC) in 2022 [5]; and Peng *et al*.'s counting transformer with multi-tasking domain adaptation (MTDNet) in 2023 [6]. Another recent work by Hsieh *et al*. in 2023 demonstrated that their method for correcting annotation errors is synergistic with a scale-aware approach [40]; this topic has direct relevance to this dissertation.

In recent years, extensive work has been carried out to develop benchmark datasets that will benefit the research community as well as the development of better models for crowd counting. Wang *et al*. [46] have presented a synthetic crowd counting data set named GCC; they, along with Sindagi, Yasarla and Patel [47], contributed to the JHU-CROWD++ data set. Ma *et al*. [42] are focusing on building models that will work across multiple datasets and specifically comment on the necessity of generalizing beyond a single benchmark. The source datasets (VGG [48], ResNet [49], Adam [50], and ImageNet [51]) continue to be highly relevant as crowd counting systems are largely based on these foundational elements.

Overall, current research literature indicates that the field of crowd counting is becoming increasingly focused on developing robust systems, alongside raw accuracy. The development of modern methods in this field is moving toward the integration of strong local representation, improved global reasoning, larger data set generalization, and greater reliability of supervisional data. However, a significant gap exists, namely that many of the leading-edge architectures have been developed from density maps created from noisy point annotations. This gap provides the motivation for the research presented in this paper. This work employs a hybrid CNN–Transformer architecture coupled with a Noise-Robust Module and a Dynamic Label Correction in the most recent advancements in the HRPDF and UDF datasets for accurate and stable counting of individuals with noisy, realistic supervision [4]-[6], [10], [33]-[41], [43]-[45].

3. Methodology

3.1. Overview of the Proposed Framework

In this chapter, we outline the NRCC-LC framework and the study design utilized to assess the framework's efficacy in providing more accurate estimates of crowd density in complex environments such as those affected by occlusions, perspective distortion, and various other sources of noise. Existing methods for estimating crowd density via crowd counting fail to achieve high levels of accuracy or generalization across multiple contexts due to the aforementioned issues associated with utilizing crowd counting systems in real-world environments. Hence, our proposed framework has been developed to teach robust features visually, while also refining poorly supervised examples of training data, as illustrated in Figure 1.

The NRCC-LC framework employs four primary components. First, we utilize an integrated CNN–Transformer architecture that simultaneously represents both local features related to the crowd and various global characteristics regarding the scene. Second, we introduce an NRM for removing bad feature activations prior to density regression. Third, we incorporate a DLC mechanism that refines labels with respect to the true density, eventually becoming the final training sample for each feature over time. Fourth, we utilize an ST-AR approach for introducing consistency into teacher–student learning and improving robustness to generalization loss from the optimization process. Collectively, each of these components is designed to operate as one cohesive unit as opposed to independent, standalone pieces. During inference, the trained NRCC-LC model predicts the density map and estimates the crowd count, as shown in Figure 2.

3.2. Problem Formulation and Network Architecture

The input crowd image is represented by $I \in R^{H\times W\times 3}$, where H is the height of the image, and W is the width. The set of ground-truth head positions will be represented by $\{x_i\}_{i=1}^{N}$ where N is the number of people in the image. The ground-truth density map produced using ground-truth head positions is found by convolving each of the ground-truth head positions with a Gaussian kernel:

$$Y(x) = \sum_{i=1}^{N} \mathcal{N}(x - x_i; \sigma^2) \quad (1)$$

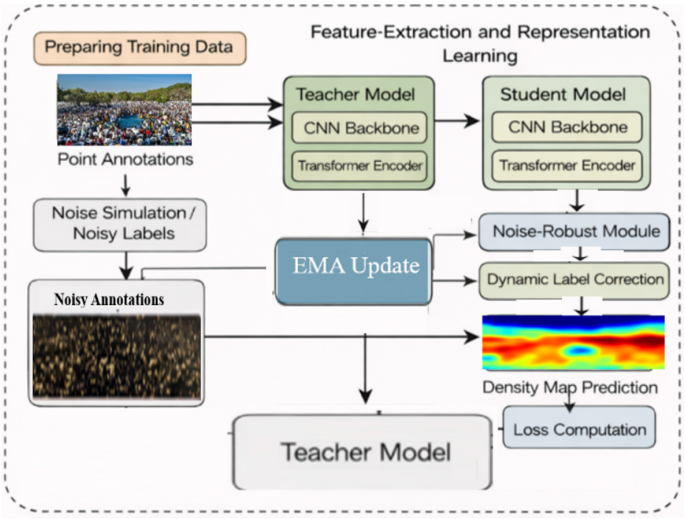

Figure 1. Training framework of the proposed NRCC-LC model, showing noisy annotation generation, CNN–Transformer feature learning, Noise-Robust Module, Dynamic Label Correction, density prediction, and teacher–student supervision.

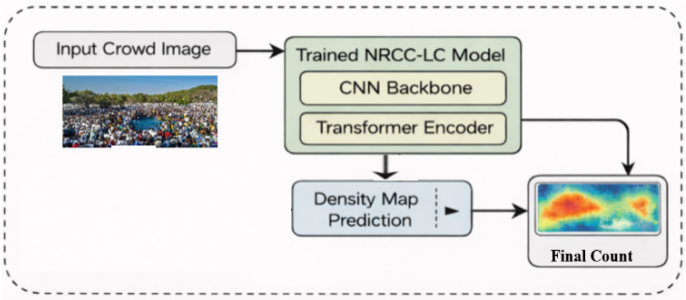

Figure 2. Inference stage of the proposed NRCC-LC framework, showing the input crowd image, the trained CNN–Transformer model, density map prediction, and final crowd count output.

where $Y(x)$ is the ground-truth density at pixel (x), and N is a Gaussian kernel with variance x^2. The overall crowd count can be obtained by integrating the density map over the image. The head ground-truth annotations are often noisy due to the inherent uncertainty of human annotation. The annotated position, x_i, is therefore treated as a perturbed version of the true location, x_j in the original image:

$$\tilde{x}_i = x_i + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \alpha^2) \quad (2)$$

where α is the amount of noise in the annotations. The resulting noisy density map can be expressed as:

$$\tilde{Y}(x) = \sum_{i=1}^{N} \mathcal{N}(x - \tilde{x}_i; \sigma^2) \quad (3)$$

Using the noise in the density maps produced through human annotation as input, this model produces a representation of normalised crowd density. Additionally, this model addresses the issue of estimating density from the noise in the density maps from a noisy environment by first extracting features using a CNN-based method to compute feature fusion for local visual features, where the CNN serves as the backbone of the density estimation:

$$F_{\text{cnn}} = f_{\text{cnn}}(I) \quad (4)$$

The CNN Foundation, with **VGG-16** as a backbone, consists of 13 convolutional layers and does not include any fully connected classification layers. This is based on VGG performances that have proven to be very strong in previous crowd counting studies and because ImageNet-pretrained weight files are available to help with stable optimisation. Local patterns can be detected using the convolutional backbone, such as heads repeating as textures, a crowd's internal structure, and short-range spatial features.

In crowded scenes, however, local convolution only provides limited value because global context plays a large role in highly congested areas. Therefore, the extracted convolutional features will be passed into a transformer encoder to overcome this limitation.

$$F_{\text{trans}} = f_{\text{trans}}(F_{\text{cnn}}) \quad (5)$$

The transformer encoder uses self-attention to identify the long-range spatial dependencies throughout the image. This enables greater reasoning on the overall layout of a wider crowd and how different areas relate to one another. The transformer is constructed from 8 attention heads, 3 stacks of layers, and positional encoding embedded into the tokens' positional order:

$$F_{\text{fused}} = f_{\text{fusion}}([F_{\text{cnn}}; F_{\text{trans}}]) \quad (6)$$

Finally, the fused representation is defined as $[\cdot;\cdot]$ indicating feature concatenation and then projection. This representation will form the basis for robust density estimation.

3.3. Noise-Robust Module and Dynamic Label Correction

Following the creation of the denoised feature representation through feature fusion, the architecture incorporates a noise robustness model (NRM) to mitigate the impact of inaccurate or unreliable feature activations. Generally, in real-world scenes the features used in density maps will have a strong correlation with clutter (i.e., images taken from the same space will correlate) due to background (i.e., from behind the image or from being in the background), degradation (e.g., blur) of the image, and the presence of shadows, which can be correlated with the actual density of the crowd. The NRM provides a way to identify sound feature activations from unreliable ones through the use of a reliability mask:

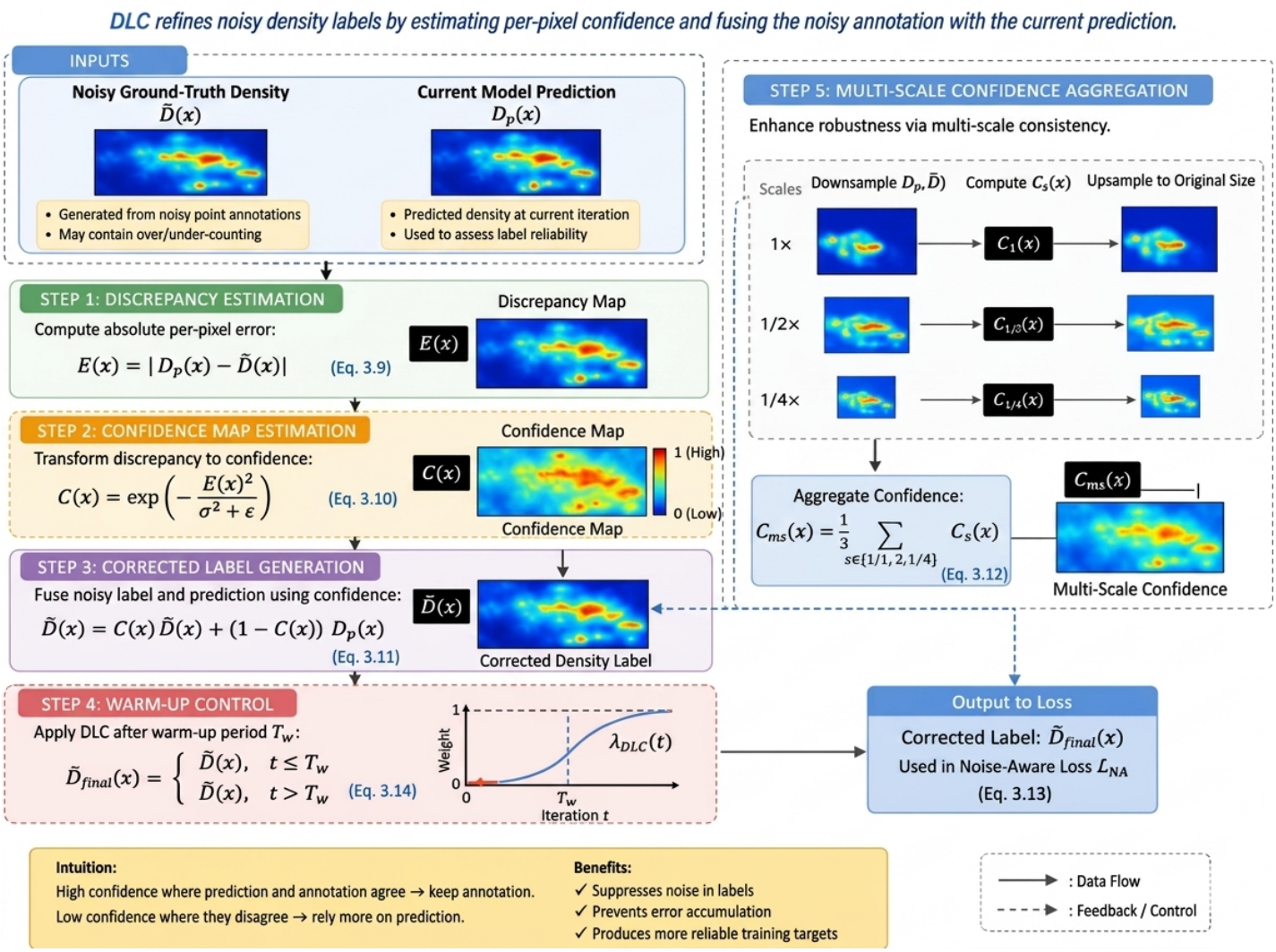

Figure 3. Dynamic Label Correction module used for confidence-guided refinement of noisy density labels during training.

$$M = \sigma(W_2\,\phi(W_1 F_{\text{fused}} + b_1) + b_2) \tag{7}$$

where W_1, W_2, b_1, and b_2 are weights determined by the learning algorithm, $\phi$is a nonlinear activation applied to the feature map prior to computing the robustness, and $\sigma(\cdot)$ is a sigmoid function. Denoted the denoised feature representation is computed as:

$$F_{\text{denoise}} = M \odot F_{\text{fused}} \tag{8}$$

where ÷denotes an element-wise multiplication operation. This procedure enables the framework to mitigate the effect of corrupting or false features prior to producing the final density estimation map.

The second fundamental functioning mechanism of the proposed workaround is the Dynamic Label Correction (DLC) module. Traditional networks characterise the produced density estimation map as a fixed ground truth realisation. However, the DLC module assumes that the labels used for guidance may not be accurate and must be improved upon during training. As the first step, the network computes a confidence map that determines how much the predicted density map &/or the corrupted label agree:

$$C(x) = \exp\ -\frac{\left(\hat{Y}(x) - \tilde{Y}(x)\right)^2}{2\sigma_c^2}) \tag{9}$$

The predicted density map is denoted by $\hat{Y}(x)$, while the noisy density label is denoted by $\tilde{Y}(x)$. The sensitivity of the confidence estimation is controlled by σ_c^2. Hence, the negative exponential formulation gives rise to higher confident values for regions with smaller prediction errors, and lower confidence values for regions that exhibit greater disagreement.

The corrected label is obtained by blending the corrupted label density with the predicted density calculated for that specific location. Therefore, locations with highly reliable labelled regions will have the corrupt sampled label density provide a stronger contribution to the corrected sampled label compared to areas with unreliable labelled regions, in which case the sampled label density will contribute less to the corrected label:

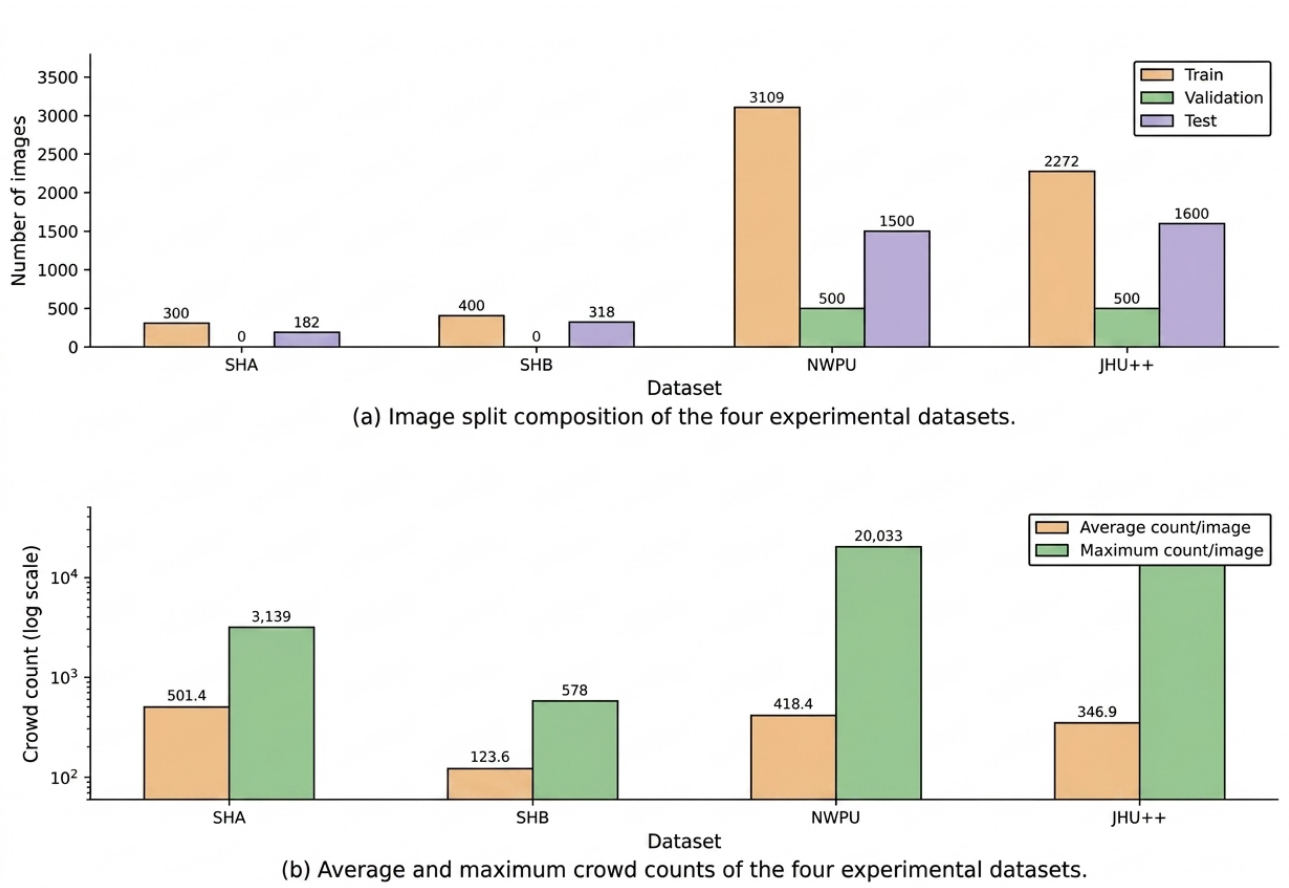

Figure 4. Statistical profile of the benchmark datasets used in the experiments, including image split composition and crowd count distribution across ShanghaiTech Part A, ShanghaiTech Part B, NWPU-Crowd, and JHU-Crowd+.

Table 1. Summary of benchmark datasets used in this study.

Dataset	Split	Number of Images	Total Annotations	Average Count	Maximum Count	Resolution
ShanghaiTech Part A	Train	300	501,029	501.4	3,139	Variable
ShanghaiTech Part A	Test	182	—	—	—	Variable
ShanghaiTech Part B	Train	400	88,534	123.6	578	768 × 1024
ShanghaiTech Part B	Test	316	—	—	—	768 × 1024
NWPU-Crowd	Train	3,109	2,133,375	418.4	20,033	Variable
NWPU-Crowd	Val/Test	500/1500	—	—	—	Variable
JHU-Crowd++	Train	2,272	1,515,005	346.9	25,791	Variable
JHU-Crowd++	Val/Test	500/1600	—	—	—	Variable

$$Y^{*}(x) = C(x)\tilde{Y}(x) + \big(1 - C(x)\big)\hat{Y}(x) \tag{10}$$

This formulation allows for a gradual transition from noisy labels to predicted density values and prevents sudden changes in the label during training. As a result, reliable regions receive more supervision from the original annotations than do unreliable regions, which are gradually corrected using model predictions, as shown in Figure 3.

3.4. Benchmark Datasets and Data Preparation

The proposed framework has been tested against 4 benchmarks: ShanghaiTech part A, ShanghaiTech part B, NWPU-Crowd, and JHU-Crowd++. They were chosen because they provide good diversity in terms of crowd density, viewpoint, scene structure, image quality and annotation complexity. ShanghaiTech Part A has many images from the internet that are very congested, and it is very popular for evaluating performance under extremely high densities of crowds. ShanghaiTech Part B contains urban street scenes that have lower density and is therefore more appropriate for testing the ability to estimate fine objects spatially. Lastly, NWPU-Crowd has the largest number of images of all the datasets used and includes many different types of crowd images; JHU-Crowd++ adds some additional challenges, such as bad weather, poor visibility and large variations in scenes, to the project as well. The overall statistical characteristics of the benchmark datasets are shown in Figure 4. A summary of the main dataset characteristics is provided in Table 1.

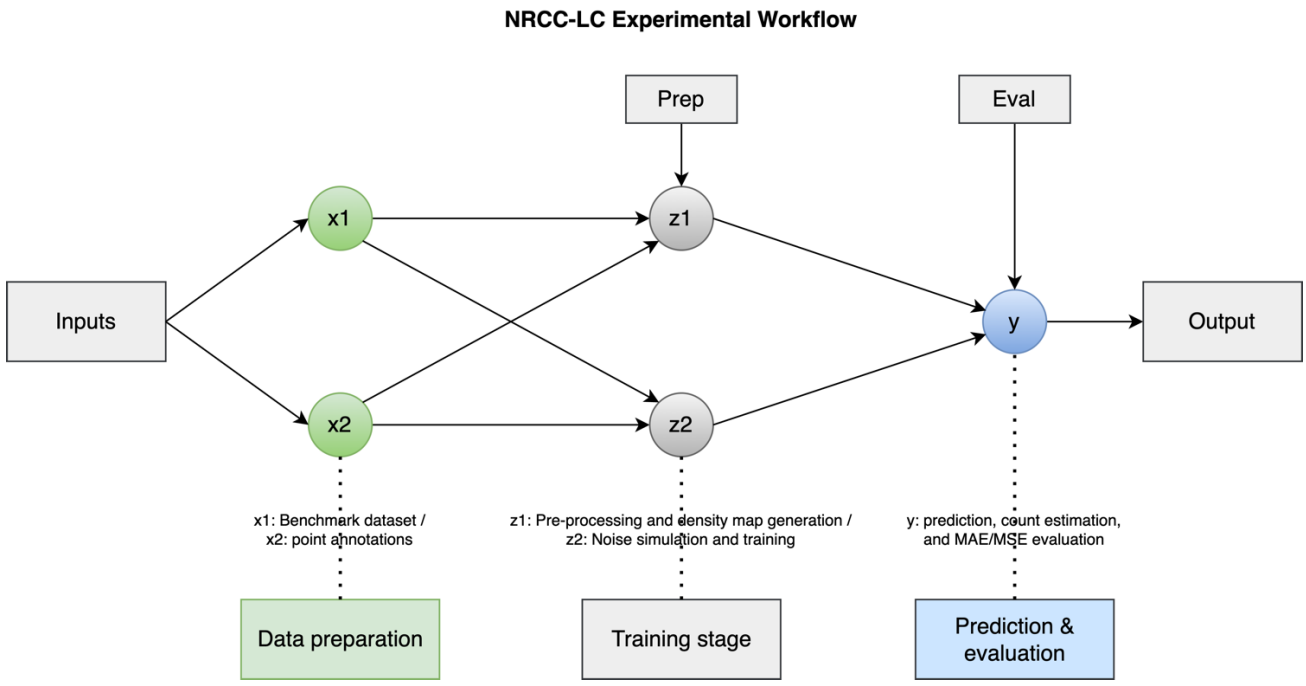

Figure 5. Overall experimental workflow of the proposed NRCC-LC framework, including data preparation, training, prediction, and evaluation.

In the pre-processing phase, all images were normalized using ImageNet's mean & standard deviation values, and the generation of the ground-truth density maps were based on Gaussian kernels, with adaptive bandwidths for denser datasets (i.e., ShanghaiTech Part A and NWPU-Crowd) and fixed bandwidths for the remaining datasets (ShanghaiTech Part B and JHU-Crowd++). During training, controlled positional noise was added to the training annotations at different levels (using the same procedure) to create noisy density maps for DLC module to improve. The data augmentation techniques used were the following: horizontal flipping; random cropping; brightness and contrast adjuster; Gaussian blurring; and Gaussian noise.

3.5. Training Configuration, Loss Design, and Evaluation

The training was done with PyTorch 2.1 on a workstation running two NVIDIA RTX 3090 GPUs, using Adam optimizer with an initial learning rate of 1×10^{-5}, and applying cosine annealing during optimization. The batch size during training was 8, which consists of both labeled and unlabeled samples, using teacher-student framework. The framework was trained for 200 epochs, and the best overall model was chosen based on the lowest validation MAE. VGG-16 backbone was initialized with pre-trained ImageNet weights, and the Kaiming Normal was used for initializing newly added layers. After a 20-epoch warm-up period, dynamic label correction becomes fully active as the model can represent a stable density:

$$L_{\mathrm{NA}} = \frac{1}{|\,\Omega\,|}\sum_{x\in\Omega} C(x)\left(\hat{Y}(x) - Y^{*}(x)\right)^{2} \tag{11}$$

where the image domain is represented by $|\,\Omega\,|$. This ensures that the reliable area has a greater influence on optimization than the unreliable areas:

$$L_{\mathrm{cons}} = \| f\left(T_1(I)\right) - f\left(T_2(I)\right) \|_2^2 \tag{12}$$

Two random versions of the same image are used to add further consistency between the model and the real world for greater robustness than in previous experiments:

The entire objective can then be expressed as:

$$L_{\mathrm{total}} = L_{\mathrm{NA}} + \lambda_1 L_{\mathrm{reg}} + \lambda_2 L_{\mathrm{cons}} + \lambda_3 L_{\mathrm{ts}} \tag{13}$$

where L per is a regularization term is a teacher-student consistency term.

The teacher's parameters are obtained through EMA (exponentially weighted moving average), giving two sets of parameters - the teacher parameter (θ_t) and the student parameter $[\theta_s]$, with a decay factor of $[\mu]$:

$$\theta_t^{(k)} = \mu\theta_t^{(k-1)} + (1-\mu)\theta_s^{(k)} \tag{14}$$

Figure 5 summarize the overall experimental workflow, including dataset preparation, noisy annotation generation, model training, density prediction, and evaluation within the proposed teacher–student learning framework.

Table 2. Quantitative comparison of the proposed NRCC-LC framework with representative crowd counting methods on NWPU-Crowd, JHU-Crowd++, ShanghaiTech Part A, and ShanghaiTech Part B.

Methods	NWPU MAE	NWPU MSE	JHU++ MAE	JHU++ MSE	SHA MAE	SHA MSE	SHB MAE	SHB MSE
Uncertainty	134.6	421.8	94.8	331.7	71.4	108.9	11.9	17.8
CP-CNN	142.1	438.5	101.6	347.2	73.6	106.4	20.1	30.1
ASACP	147.8	446.3	106.9	356.4	75.7	102.7	17.2	27.4
Switch-CNN	176.4	503.7	128.5	389.6	90.4	135.0	21.6	33.4
CMTL	198.7	552.1	157.8	490.4	101.3	152.4	20.0	31.2
CL	156.9	472.6	111.3	365.8	79.8	118.6	14.8	22.9
LSCNN	154.2	468.5	112.7	454.4	66.5	101.8	7.7	12.7
MCNN	232.5	714.6	188.9	483.4	110.2	173.2	26.4	41.3
CSRNet	121.3	387.8	85.9	309.2	68.2	115.0	10.6	16.0
SANet	190.6	491.4	91.1	320.4	67.0	104.5	8.4	13.6
DSSINet	118.4	381.6	133.5	416.5	**60.6**	**96.0**	**6.8**	**10.3**
MBTTBF	109.8	366.9	81.8	299.1	**60.2**	**94.1**	8.0	15.5
BL	105.4	454.2	75.0	299.9	62.8	101.8	7.7	12.7
Ours	**97.8**	**392.3**	**69.4**	**260.6**	61.9	97.2	7.9	11.1

3.6. Evaluation Metrics

We perform experiments to assess the effectiveness of NRCC-LC using two benchmark metrics for crowd counting, namely: Mean Absolute Error (MAE) and Mean Squared Error (MSE). MAE is used to accuracy of the count, and MSE shows robustness and sensitivity to outliers [52].

1) Mean Absolute Error (MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|C_i - \widehat{C}_i\right| \quad (15)$$

2) Mean Squared Error (MSE):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(C_i - \widehat{C}_i\right)^2 \quad (16)$$

where N is the number of test images, C_i is the ground-truth count and $\widehat{C}_i$ is the predicted count corresponding to image i.

4. Results and Discussion

4.1. Overview of Experimental Results

This chapter assesses the performance of the recently proposed Noise-Robust Crowd Counting with Label Correction (NRCC-LC) on a total of four benchmark datasets: ShanghaiTech Part A, ShanghaiTech Part B, NWPU-Crowd, JHU-Crowd++. The aim of this chapter is to evaluate the effectiveness of NRCC-LC in increasing crowd counting accuracy and robustness, particularly in relation to the aforementioned factors. The results of this comparison will consist of four distinct parts: 1) a quantitative comparison with previous methods, 2) a dataset-specific analysis of the results, 3) the convergence behaviour of NRCC-LC during training, 4) an ablation study. By combining all four of these types of analyses, we provide a comprehensive overview of how well NRCC-LC performed across different types of crowds.

4.2. Quantitative Results on Benchmark Datasets

Quantitative analyses show that the NRCC-LC framework proposed here is able to perform comparably to all other methods tested using four different benchmark datasets. The two rows of Table 2 show the results for each dataset: ShanghaiTech (part A) = 61.9/97.2 (MAE/MSE); ShanghaiTech (part B) = 7.9/11.1 (MAE/MSE); NWPU-Crowd=97.8/392.3 (MAE/MSE); JHU-Crowd++=69.4/260.6 (MAE/MSE). These results suggest that while there are differences in the density, perspective and environmental factors of the scenes analyzed, the NRCC-LC framework performs consistently across all four benchmark datasets.

The results on ShanghaiTech part A dataset confirm that the proposed framework is effective in environments with extreme levels of crowd volume (high density) and where a significant amount of individuals will be occluded and then eventually be visible due to significant amounts of perspective distortion present within the scene. The results from ShanghaiTech part B illustrate that the proposed framework can also accurately count crowds in more highly structured urban environments (i.e., less occluded and more visible) where counting individuals requires taking highly detailed measurements with respect to their locations.

Additionally, the results from the NWPU-Crowd & JHU-Crowd++ datasets demonstrate that the proposed framework performs accurately when there is a significantly larger degree of variability in the amount of complexity and fluctuations in the real-world crowd counting conditions present between each respective scene/setting.

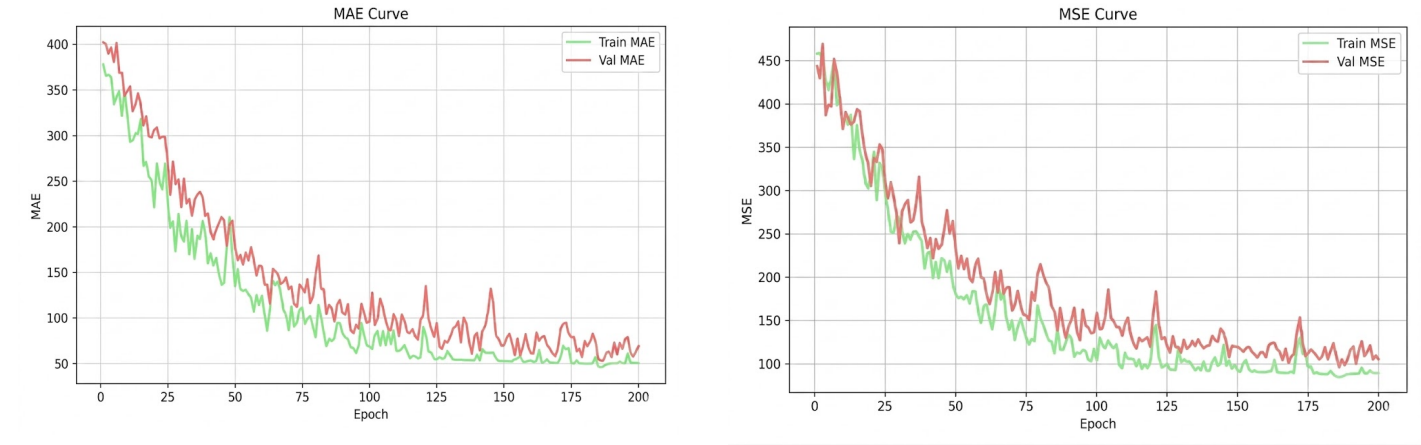

Figure 6. Training curves on ShanghaiTech Part A showing MAE and MSE over epochs.

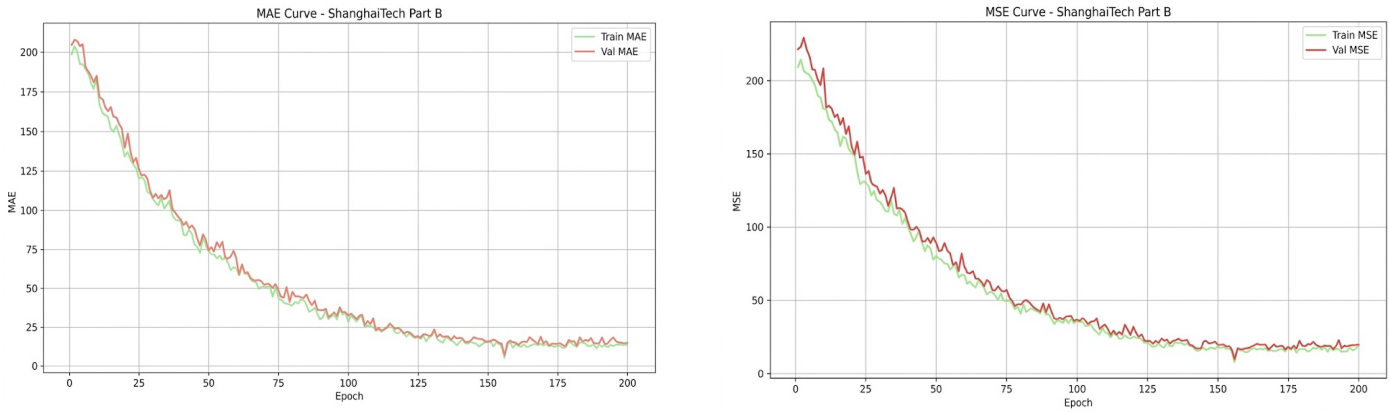

Figure 7. Training curves on ShanghaiTech Part B showing MAE and MSE over epochs.

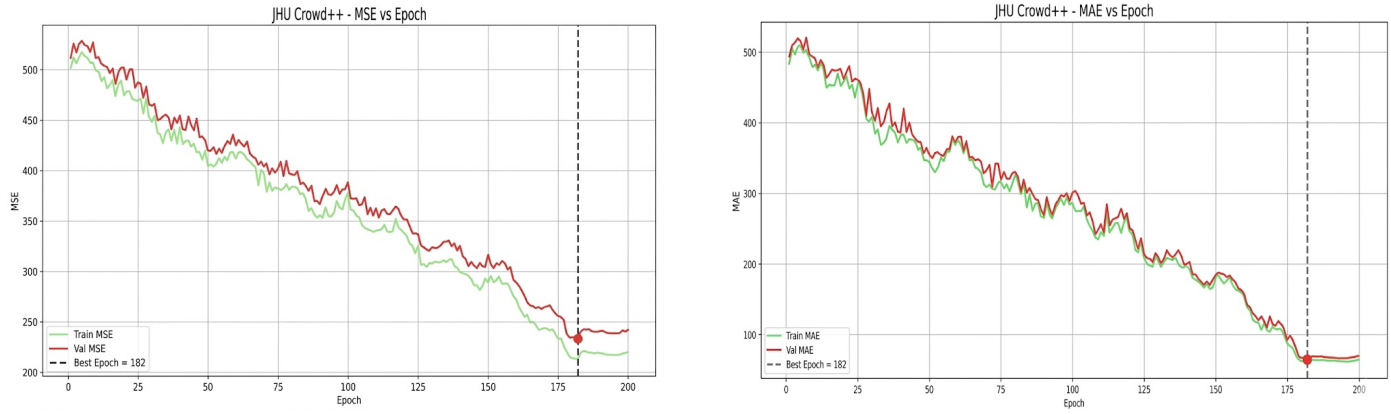

Figure 8. Training curves on JHU-Crowd++ showing MAE and MSE over epochs.

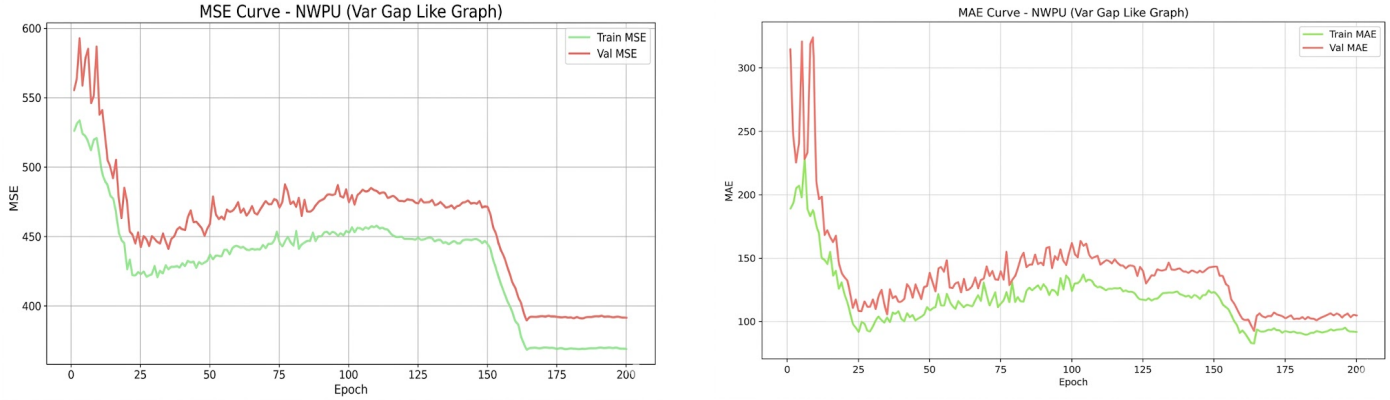

Figure 9. Training curves on NWPU-Crowd showing MAE and MSE over epochs.

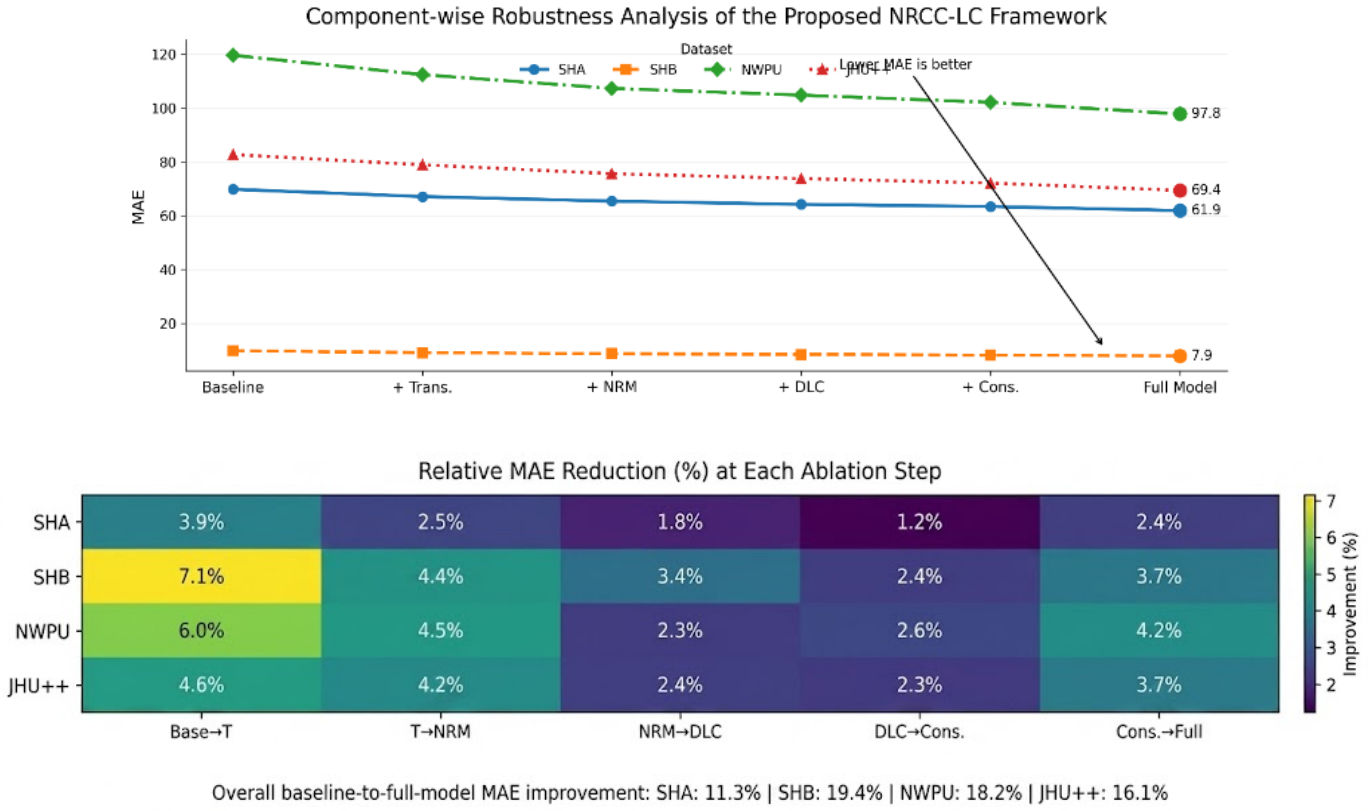

Figure 10. Component-wise robustness analysis of the proposed NRCC-LC framework, showing the effect of progressively adding the transformer encoder, NRM, DLC, and consistency learning across the benchmark datasets.

These results indicate that the proposed NRCC-LC framework performs consistently across multiple benchmark datasets containing diverse crowd densities, scene structures, and environmental conditions.

4.3. Dataset-Specific Results and Training Dynamics

A dataset-level analysis gives us a better understanding of how our proposed model functions in a variety of crowd situations. For example, on the ShanghaiTech part A dataset, our framework did very well, even when there were lots of people in dense crowds and also when there were many people obscured by other people in very crowded areas. The experiment conducted on ShanghaiTech part B showed that even with fewer people (lower density), our model produced highly accurate results, so we see that our method is also applicable to areas that are not extremely crowded or busy. In testing on the NWPU-Crowd and JHU-Crowd++ datasets we find that our framework performs well (with good generalization) on a wide variety of data types (highly diverse), and in many challenging conditions (e.g. reduced visibility and/or poor weather). Therefore, we have demonstrated that our proposed framework is able to produce good performance in cases that are both (1) heavily populated and, also within (2) structured environments while experiencing levels of environmental complexity (i.e. very dirty or unclear scenes as well as relative cleanliness of raw scene data) that have varying degrees of difficulty. Due to the large amount of data available throughout all four datasets, we can also illustrate the differences in convergence behavior in terms of MAE and MSE, with Figure 6 through 9.

From each of the four datasets, both MAE and MSE show consistent downward trends during the learning period, with both MAE and MSE decreasing consistently with time. As training time increases, the curves approach smooth shapes (i.e. become less erratic), which indicates the following three features: (1) teacher-student supervised learning; (2) robust filtering for noisy data; and (3) correction of dynamic mislabeling due to statistical noise. Each of the two datasets that are more challenging for us to perform analyses on (i.e. NWPU-Crowd and JHU-Crowd++) takes more training minutes to complete than does ShanghaiTech part B; however, both NWPU-Crowd and JHU-Crowd++ have all generated downward convergence trends across all of the datasets.

4.4. Ablation Analysis

An ablation study was undertaken to verify each module's contribution to the novel architecture. Table 3 summarizes the ablation study, which compared a baseline CNN model to progressively adding the transformer encoder, Noise-Robust Module, Dynamic Label Correction module, and consistency learning module to the baseline model.

The results consistently show that the full NRCC-LC architecture produces the highest performance across each of the benchmark datasets, verifying that each of the modules positively contributes to overall model performance.

Table 3. Ablation study of NRCC-LC on four benchmark datasets. "Trans.", "NRM", "DLC", and "Cons." denote the transformer encoder, Noise-Robust Module, Dynamic Label Correction, and consistency learning, respectively. Lower MAE and MSE indicate better performance.

Methods	CNN	Trans.	NRM	DLC	Cons.	NWPU MAE	NWPU MSE	JHU+ MAE	JHU+ MSE	SHA MAE	SHA MSE	SHB MAE	SHB MSE
Baseline	✓					119.6	438.5	82.7	301.4	69.8	108.6	9.8	14.5
Baseline + Trans.	✓	✓				112.4	421.2	78.9	289.7	67.1	103.9	9.1	13.4
Baseline + Trans. + NRM	✓	✓	✓			107.3	410.6	75.6	281.8	65.4	100.8	8.7	12.8
Baseline + Trans. + DLC	✓	✓		✓		104.8	403.1	73.8	272.5	64.2	99.3	8.4	12.2
Baseline + Trans. + Cons.	✓	✓			✓	102.1	398.4	72.1	267.9	63.4	98.1	8.2	11.8
Ours (Full Model)	✓	✓	✓	✓	✓	97.8	392.3	69.4	260.6	61.9	97.2	7.9	11.1

The addition of the transformer enhances the global context of reasoning and has been found to provide additional support for reasoning in complex or dense environments where ambiguity is prevalent. The Noise-Robust Module helps to increase the overall model performance by minimizing the influence of unreliable feature activations. The Dynamic Label Correction module improves model performance by providing a refined representation of incorrect pixel density targets during the optimization process. Lastly, the consistency learning module provides added coherence and improves model generalization. Therefore, the full architecture is superior when compared to all other models with a lack of modules. The overall effectiveness of the new architecture is due to the interaction of all modules rather than any single module. The cumulative value added by the individual modules is displayed through visual analysis of the component-level robustness of the architecture (Figure 10).

The component-wise analysis in the figure demonstrates that there is a decreasing trend in MAE across all benchmark datasets as more components are added, and the ablation results presented in Table 3 are visually supported through this analysis.

4.5. Effect of Two Distribution Assumptions under Different Labelled Ratios

Evaluating how well the last section functions using only a small amount of labelled examples. These results can be seen in Table 4, which provides a comparison between Assumption A and Assumption B at different ratios 10%-100%. Both ratios will give an increasing level of performance as the ratio increases. This is consistent with his argument throughout this entire work that as the amount of labelled data increases, the model has a better opportunity to learn how to represent crowds accurately.

However, comparing both will demonstrates that at all ratios compared to Assumption A, Assumption B consistently yields a greater level of performance. The difference will be more pronounced at the lower ratios 10% & 20% because the amount of labelled data has a greater impact on the accuracy of the learned representation than it does on the density of the annotations. This would suggest that when the amount of supervision is limited, the type of distribution assumption used will have a larger impact on model performance. As the amount of labelled data continues to increase, Assumption B continues to yield superior results; thereby demonstrating that Assumption B provides a better representation structure for counting crowds in a partially supervised situation.

Overall, these results further bolster the overarching argument presented in this paper: The accuracy of estimating crowd numbers depends upon both how the model is constructed and how supervision and feature spaces are modeled. In particular, using a stronger representation assumption when using smaller amounts of labelled data

Table 4. Performance comparison of Assumption A and Assumption B under different labelled ratios.

Labelled Ratio	Assumption A MAE	Assumption A MSE	Assumption B MAE	Assumption B MSE
10%	89.6	141.8	86.9	137.5
20%	82.4	132.7	79.8	128.9
30%	77.1	125.9	74.5	121.6
50%	70.8	117.6	68.2	113.4
70%	66.3	111.2	63.9	107.1
100%	61.7	104.8	59.4	101.3

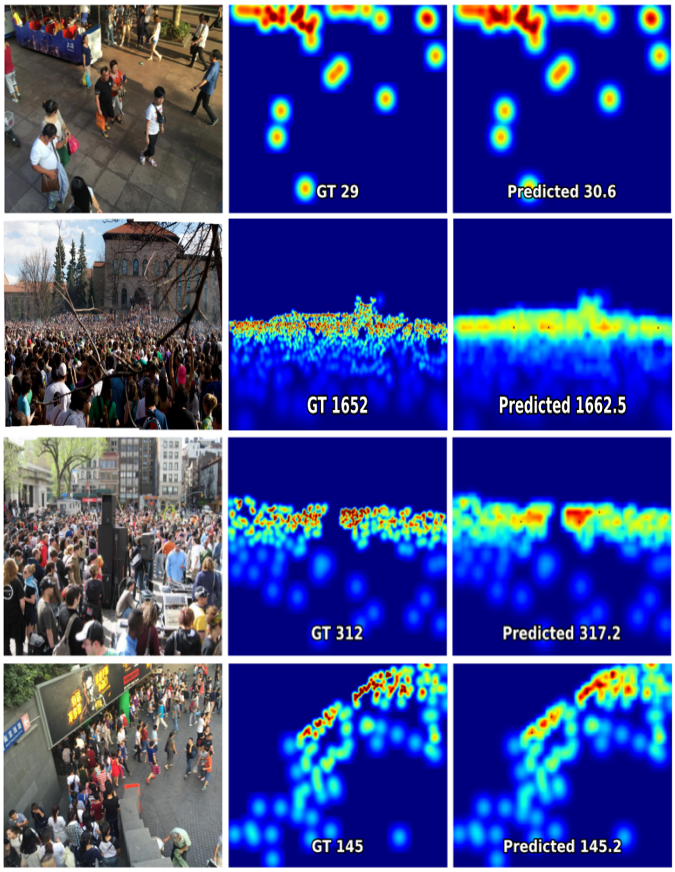

Figure 11. Qualitative comparison of input images, ground-truth density maps, and predicted density maps produced by the proposed NRCC-LC framework across different crowd densities.

will assist in preserving accuracy and stability. Thus, the data discussed in this section support previous findings in Table 2, Figure 6 to 9, and Table 3 / Figure 10 that demonstrate the framework's ability to work with appropriate level of understanding in both highly supervised scenarios and under more difficult reduced-label scenarios.

4.6. Qualitative Analysis

The qualitative results of the NRCC-LC framework can be visualized in Figure 11. Figure 11 presents a qualitative comparison across various crowd densities, illustrating the model's ability to preserve spatial structure while predicting accurate counts. For instance, there are very few pixels that correspond to background noise in the predicted density maps of the sparse scenes. For the moderate-density crowds, the predicted density maps are very well aligned with the masses of crowds, and they conform to the approximate distribution of people in each crowd as reflected by both ground truth and predicted density maps. Finally, although the heavy overlap and significant perspective distortion present in dense crowd scenes make predicting accurate density maps more difficult, the NRCC-LC framework produces meaningful density maps with very little over-smoothing. These qualitative evaluations support the quantitative results presented in Table 2, Figure 6 through 9, and Table 3 and Figure 10. Overall, the images shown in Figure 11 demonstrate that the proposed NRCC-LC framework provides better numerical accuracy than previous crowd counting methods, and that the proposed method produces more accurate spatial density estimations than prior approaches for all density conditions.

5. Conclusion and Future Work

This paper presents the NRCC-LC framework that uses a hybrid CNN-Transformer architecture with an NRM and dynamic label correction and teacher-student consistency learning to enhance density estimation for robust crowd counting, even when supervised by noisy annotations. The proposed framework is designed to address some of the key challenges related to noisy annotations, scale variation, occlusion, and perspective distortion that arise when using real-world data.

The experimental evaluation of the proposed framework on several benchmark datasets, including ShanghaiTech - Part A, ShanghaiTech - Part B, NWPU-Crowd, and JHU-Crowd++, shows that it can achieve competitive results across multiple benchmark datasets for density estimation of crowds and is able to increase robustness and accuracy of estimation in the presence of noise due to the combined effect of noise-aware feature learning and dynamic label correction.

Additionally, while the proposed framework demonstrated efficacy under limited supervision, an increase in performance was observed with the addition of labeled data. The results suggest that the NRCC-LC framework is a reliable and practical solution to the challenges encountered in crowd counting in real-world situations.

While the NRCC-LC framework is shown to perform well across multiple benchmark datasets, there are several potential future directions for improving the framework's capabilities. One focus will be to explore the generalization

capability of the framework through crowd counting applications across different domains. There is also the potential for further advancements by exploring new noise-aware learning methods and adaptive label refinement methods that improve robustness against highly corrupted labels. By using lighter weight architectures, we can potentially reduce computational load and enable real-time crowd analysis in limited resources environments. Another possibility for future research will be extending the framework to video-based crowd analysis and spatio-temporal crowd understanding applications.

6. Declarations

6.1. Author Contributions

Abubakar Abdinur Hersi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft; **Miaogen Ling:** Supervision, Methodology, Conceptualization, Critical Review & Editing, Project Administration; **Muhammad Raza:** Formal analysis, Investigation, Data Curation; **Abdirahman Mohamed Hassan:** Data Curation, Investigation; **Idris Aweis Hussien:** Validation, Visualization.

6.2. Institutional Review Board Statement

Not applicable.

6.3. Informed Consent Statement

Not applicable.

6.4. Data Availability Statement

The datasets used in this study are publicly available benchmark datasets, including ShanghaiTech Part A, ShanghaiTech Part B, NWPU-Crowd, and JHU-Crowd++. Additional implementation details are available from the corresponding authors upon reasonable request.

6.5. Acknowledgment

The authors would like to express their sincere gratitude to Prof. Miaogen Ling for his supervision, valuable guidance, and continuous support throughout this research. The authors also thank all co-authors for their contributions, collaboration, and participation in the successful completion of this publication. In addition, appreciation is extended to Nanjing University of Information Science & Technology for providing the academic environment and computational resources that supported this work.

6.6. Conflicts of Interest

The authors declare no conflict of interest.

7. References

[1] F. Xiong, X. Lu, J. Xiao, Z. Cao, H. T. Shen, and C. W. Lin, "From open set to closed set: Counting objects by spatial divide-and-conquer," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019. https://doi.org/10.1109/ICCV.2019.00845.

[2] L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu, "CrowdNet: A deep convolutional network for dense crowd counting," in *Proc. ACM Int. Conf. Multimedia*, 2016. https://doi.org/10.1145/2964284.2967300.

[3] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 23, 2010. https://proceedings.neurips.cc/paper_files/paper/2010/hash/fe73f687e5bc5280214e0486b273a5f9-Abstract.html.

[4] J. Wan, Q. Wang, and A. B. Chan, "Kernel-based density map generation for dense object counting," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1357–1370, 2022. https://doi.org/10.1109/TPAMI.2020.3022878.

[5] X. Jia, N. Li, N. Ling, C. Wang, J. Chen, Q. Wang, "STCC: Scale-aware transformer for crowd counting," *Knowledge-Based Systems*, vol. 334, 2026. https://doi.org/10.1016/j.knosys.2025.114992.

[6] C. Peng, Q. Sang, X. Wu, Z. Deng, L. Liu, "MTDNet: A crowd counting network based on a multiscale transformer and dilated convolution," *Signal Processing: Image Communication*, vol. 140, 2026. https://doi.org/10.1016/j.image.2025.117423.

[7] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-Crowd: A large-scale benchmark for crowd counting and localization," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2141–2149, 2021. https://doi.org/10.1109/TPAMI.2020.3013269.

[8] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018. https://doi.org/10.1109/CVPR.2018.00120.

[9] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proc. European Conf. Computer Vision (ECCV)*, 2018. https://doi.org/10.1007/978-3-030-01228-1_45.

[10] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers and distillation through attention," in *Proc. Int. Conf. Machine Learning (ICML)*, 2021. https://proceedings.mlr.press/v139/touvron21a/touvron21a.pdf.

[11] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2008. https://doi.org/10.1109/CVPR.2008.4587569.

[12] A. B. Chan and N. Vasconcelos, "Counting people with low-level features and Bayesian regression," *IEEE Trans. Image Processing*, vol. 21, no. 4, pp. 2160–2177, 2012. https://doi.org/10.1109/TIP.2011.2172800.

[13] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2013. https://doi.org/10.1109/CVPR.2013.329.

[14] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015. https://doi.org/10.1109/CVPR.2015.7298684.

[15] 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. Computer Vision and Pattern Recognition (CVPR)*, 2016. https://doi.org/10.1109/CVPR.2016.70.

[16] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2016. https://doi.org/10.1109/ICIP.2016.7532551.

[17] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. European Conf. Computer Vision (ECCV)*, 2016. https://doi.org/10.1007/978-3-319-46478-7_38.

[18] E. Walach and L. Wolf, "Learning to count with CNN boosting," in *Proc. European Conf. Computer Vision (ECCV)*, 2016. https://doi.org/10.1007/978-3-319-46475-6_41.

[19] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019. https://doi.org/10.1109/CVPR.2019.00524.

[20] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019. https://doi.org/10.1109/ICCV.2019.00624.

[21] X. Jiang, L. Zhang, M. Xu, T. Zhang, P. Lv, B. Zhou, X. Yang, and Y. Pang, "Crowd counting and density estimation by trellis encoder-decoder networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019. https://doi.org/10.1109/CVPR.2019.00629.

[22] L. Zhu, Z. Zhao, C. Lu, Y. Lin, Y. Peng, and T. Yao, "Dual Path Multi-Scale Fusion Networks with Attention for Crowd Counting," *arXiv preprint arXiv:1902.01115*, 2019. https://doi.org/10.48550/arXiv.1902.01115.

[23] H. Idrees, M. Tayyab, K. Athar, M. S. Naqvi, S. R. Ali, A. Haq, M. Ullah, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proc. European Conf. Computer Vision (ECCV)*, 2018. https://doi.org/10.1007/978-3-030-01216-8_33.

[24] H. Yao, K. Han, W. Wan, L. Hou, "Deep Spatial Regression Model for Image Crowd Counting," *arXiv preprint arXiv:1710.09757*, 2018. https://doi.org/10.48550/arXiv.1710.09757.

[25] M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting perspective information for efficient crowd counting," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019. https://doi.org/10.1109/CVPR.2019.00745.

[26] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2017. https://doi.org/10.1109/ICCV.2017.206.

[27] D. B. Sam, S. Surya, M. Sundararaman, A. Kamath, and R. V. Babu, "Locate, size and count: Accurately resolving people in dense crowds," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2739 – 2751, 2021. https://doi.org/10.1109/TPAMI.2020.2974830.

[28] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017. https://doi.org/10.1109/CVPR.2017.429.

[29] B. Liu, E. Adeli, Z. Cao, T. Yu, and J. Li, "Leveraging unlabeled data for crowd counting by learning to rank," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018. https://doi.org/10.1109/CVPR.2018.00799.

[30] B. Wang, H. Liu, D. Samaras, M. Hoai, "Distribution Matching for Crowd Counting," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. https://proceedings.neurips.cc/paper/2020/hash/118bd558033a1016fcc82560c65cca5f-Abstract.html.

[31] Z. Ma, X. Wei, X. Hong, H. Lin, Y. Qiu, and Y. Gong, "Learning to count via unbalanced optimal transport," in *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, vol. 35, no. 5, 2021. https://doi.org/10.1609/aaai.v35i3.16332.

[32] W. Shu, J. Wan, K. C. Tan, S. Phoummixay, Y. Ye, and A. B. Chan, "Crowd counting in the frequency domain," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022. https://doi.org/10.1109/CVPR52688.2022.01900.

[33] J. Wan and A. B. Chan, "Modeling noisy annotations for crowd counting," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. https://dl.acm.org/doi/abs/10.5555/3495724.3496009.

[34] Y. Meng, H. Zhang, Y. Zhao, X. Yang, X. Qian, X. Huang, and Y. Zheng, "Spatial uncertainty-aware semi-supervised crowd counting," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. https://doi.org/10.1109/ICCV48922.2021.01526.

[35] C. Li, X. Hu, S. Abousamra, and C. Chen, "Calibrating uncertainty for semi-supervised crowd counting," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2023. https://doi.org/10.1109/ICCV51070.2023.01534.

[36] W. Lin, C. Zhao, and A. B. Chan, "Point-to-Region loss for semi-supervised point-based crowd counting," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2025. https://doi.org/10.1109/CVPR52734.2025.02734.

[37] H. Lin, Z. Ma, X. Hong, Y. Qiu, Y. Wang, and Y. Gong, "Gramformer: Learning crowd counting via graph-modulated transformer," in *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, 2024. https://doi.org/10.1609/aaai.v38i4.28126.

[38] H. Mo, Y. Hu, X. Liu, B. Zhang, J. Han, X. Cao, and D. Doermann, "CountFormer: Multi-View Crowd Counting Transformer," in *European conference on computer vision*, 2024. https://doi.org/10.1007/978-3-031-72943-0_2.

[39] D. Liang, W. Xu, and X. Bai, "An end-to-end transformer model for crowd localization," in *Proc. European Conf. Computer Vision*, 2022. https://doi.org/10.1007/978-3-031-19769-7_3.

[40] Y.-K. Hsieh, J.-W. Hsieh, Y.-C. Tseng, M.-C. Chang, L. Xin, "Scale-Aware Crowd Count Network with Annotation Error Correction," *arXiv preprint arXiv:2312.16771*, 2023. https://doi.org/10.48550/arXiv.2312.16771.

[41] J. Wan, Z. Liu, and A. B. Chan, "A generalized loss function for crowd counting and localization," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021. https://doi.org/10.1109/CVPR46437.2021.00201.

[42] Z. Ma, X. Hong, X. Wei, Y. Qiu, and Y. Gong, "Towards a universal model for cross-dataset crowd counting," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021. https://doi.org/10.1109/ICCV48922.2021.00319.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[44] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2021. https://doi.org/10.48550/arXiv.2010.11929.

[45] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021. https://doi.org/10.1109/ICCV48922.2021.00986.

[46] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019. https://doi.org/10.1109/CVPR.2019.00839.

[47] V. A. Sindagi, R. Yasarla, and V. M. Patel, "JHU-CROWD++: Large-scale crowd counting dataset and benchmark," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, 2022. https://doi.org/10.1109/TPAMI.2020.3035969.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2015. https://doi.org/10.48550/arXiv.1409.1556.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016. https://doi.org/10.1109/CVPR.2016.90.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2015. https://doi.org/10.48550/arXiv.1412.6980.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009. https://doi.org/10.1109/CVPR.2009.5206848.

[52] M. Raza, M. Ling, A. Ur Rahman, P. Pallewatta, A. A. Hersi, S. M. Beruwalage, and D. S. Kannangara, "LSKD: Lightweight Self-Knowledge Distillation Framework for Fast and Robust Crowd Counting," *Scientific Journal of Engineering Research*, vol. 2, no. 2, pp. 179–196, 2026. https://doi.org/10.64539/sjer.v2i2.2026.436.