ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

UDC 004.932.72:621.397.42 **doi: 10.32620/reks.2025.2.08**

Ruslan DOBRYSHEV[1], Sergiy PURISH[2], Mykhaylo LOBACHEV[1], Mykola HODOVYCHENKO[1]

[1]***Odessa Polytechnic National University, Odesa, Ukraine***
[2] ***Enestech Software, Wilmington, USA***

CROWD COUNTING IN INTELLIGENT VIDEO SURVEILLANCE SYSTEMS

This study focuses *on enhancing the accuracy and robustness of crowd counting in intelligent video surveillance systems by incorporating perspective-awareness into deep learning models. Traditional convolutional neural networks often struggle with scale variations caused by perspective distortions and object occlusions in dense crowd scenes.* ***The goal of this study is*** *to develop a method that leverages geometric depth information to improve the spatial consistency of density estimations and provide more reliable predictions across varying scene configurations, including highly congested or irregular environments.* ***The tasks to be*** *accomplished include designing a depth inclusion module, integrating it into an encoder-decoder architecture, generating depth maps from monocular RGB images, and recalibrating feature representations using attention-weighted scale-aware mechanisms.* ***The methods used*** *involve the extraction of depth features from images via a pre-trained depth estimation model, followed by spatial attention-based recalibration of feature maps to highlight foreground objects and suppress irrelevant background signals. A fully differentiable pipeline is implemented to ensure seamless integration into standard CNN frameworks. The network training procedure also incorporates Euclidean loss functions on pixel-level density maps to optimize scale-sensitive prediction. The proposed method is evaluated on benchmark datasets including ShanghaiTech-B, UCF_CC_50, and Mall, where it consistently outperforms state-of-the-art models in terms of MAE and MSE. The experimental results confirm that the explicit incorporation of depth-aware representations significantly enhances counting performance, especially in scenarios with severe perspective-induced scale disparities. The integration of geometric priors into data-driven models offers a promising direction for real-time surveillance and large-scale crowd monitoring applications, providing not only quantitative improvements but also greater spatial fidelity in density map generation and better adaptability to complex visual conditions.*

Keywords: *crowd counting; video surveillance; deep learning; convolutional neural networks; depth estimation; density map; perspective distortion; attention mechanism.*

1. Introduction

1.1. Motivation

Fast urbanization, the growing frequency of major events, and the increasing need to improve public safety have attracted considerable attention for crowd analysis in video surveillance systems. Autonomous, intelligent video surveillance systems assessing crowd behavior help reduce hazards associated with important events, such as criminal activities, stampedes, or panic.

Furthermore, developments in processing power, computer vision, and artificial intelligence techniques help in the creation of more exact and scalable crowd analysis techniques. Two primary fields of research in crowd analysis distinguish themselves: the first stresses crowd size estimation, whereas the second studies and analyzes individual behavior within the crowd [1].

Crowd analysis depends on crowd size estimation, as it allows attendance evaluation without depending on actual observations. Two main approaches might be used for this operation: counting people in the seen area via detection techniques or evaluating the crowd density of the scene [2].

Crowd counting is particularly important in wide and dynamic environments with many entry and exit points. Examining the success of the event and planning the next ones depend on knowing the attendance and crowd density. Moreover, information on crowd dynamics and density is crucial for identifying increased dangers, such as asphyxiation caused by stamping or compression.

1.2. State of the art

Between 2024 and 2025, various studies have made significant progress in crowd counting for intelligent video surveillance. From increasing accuracy to reducing computing complexity, from strengthening model generalization capabilities to creating approaches that minimize the need for significant data labeling, the following steps are taken.

Hybrid methods, in which regression and detection methods have been integrated, are attracting widespread interest because of their flexibility in controlling crowdedness. Alhawsawi et al. [3] recently developed an image-switching system that greatly enhanced the accuracy of crowd counting in all scenarios based on scene density.

Lightweight models suitable for devices with low processing capabilities have undergone tremendous development. Khan et al. [4] investigated knowledge distillation for ideas from complex, deeper models being transferred to lightweight real-time models suitable for edge computing. Their method has achieved a significant improvement in mean absolute error.

The use of crowd counting in a smart city context is particularly prominent. Mansouri et al. [5] proposed a model that combined DenseNet with ConvLSTM architecture and improved it using metaheuristic algorithms. They achieved a remarkable accuracy of 98.4% in urban crowd density classification.

Attention techniques consistently boost the accuracy of crowd density estimation skills. Zeng et al. [6] built a correlation-attention framework that significantly reduces prediction error compared with existing methods.

Weakly supervised learning methods that reduce the cost of full data annotation are promising. Cai and Zhang [7] presented the CrowdCCT model, which combines the CNN and transformer architecture. Their model achieves accuracy in crowd counting that is equal to or sometimes exceeds that of fully supervised systems with merely summed up crowd counts as input.

Self-attention-based processes have been effective, particularly in demanding and crowded scenarios. Lien and Wu [8] demonstrated the effectiveness of the self-attention mechanism, obtaining above 83% accuracy in challenging situations.

Drone-based surveillance has progressed, especially in the annotation of small objects appearing in aerial imagery. Alhawsawi et al. [9] improved YOLOv8 [10] by adding the context-enrichment module: the result was a radical reduction in error rates at crowded drone-filled sites.

Several studies have demonstrated breakthroughs in neural network feature extraction methodologies. Zhao et al. [11] proposed the CL-DCNN model. Its cross-layer dilated convolutions provide more complete contextual information, a method that greatly enhances precision, particularly in crowded environments.

Tomar et al. [12] developed EDCCN, which combines feature propagation and multi-scale merging techniques to provide benchmark-level accuracy in crowd counting. Li et al. [13] constructed PC-Net, a weakly-supervised model with a Parent-Child structure. This model closed the performance gap between weakly and fully supervised approaches on several datasets.

It is still difficult to generalize to new datasets, but a lot of progress is making substantial improvements. Zhou et al. [14] proposed the MAHE model. Using multilevel attention processes can prevent accuracy losses as the test environment changes.

Simultaneously, Cao et al. [15] adopted contrastive learning to descend their precision error in head localization and counting without relying on inapplicable or downright dangerous methods for heavily populated contexts with obstructions everywhere.

In summary, breakthroughs in innovation have increased the scope for applications of crowd counting technology that call upon deep neural networks, methods that use attention mechanism, weakly supervised learning methods, and hybrid models incorporating a mix of different machine learning methods.

1.3. Objectives and tasks

Although deep learning methods have greatly improved crowd counting accuracy, several outstanding problems hinder the whole solution of crowd counting. The perspective issue, which refers to the phenomenon of people closer to the camera occupying a larger area in the photograph than those farther away, is a significant research area.

Most convolutional neural network-based methods do not commonly use the perspective-correcting mechanism included in traditional crowd counting methods [16]. Inspired by the idea that adding perspective information in deep learning approaches would improve crowd counting accuracy, a density-based crowd counting approach using an image depth generation module was developed.

This study aims to make crowd counting in intelligent video surveillance systems more accurate and reliable by teaching neural networks to better understand depth and perspective. In real-world scenes, people can appear larger or smaller depending on how far they are from the camera, which can confuse standard models. We aim to help the model adjust for these differences and make more consistent predictions by incorporating depth cues into the learning process.

To reach this goal, this study sets out to

– developing a module that can extract and use depth information alongside visual features;

– this module is integrated into a neural network that learns to estimate crowd density;

– apply attention mechanisms to help the network focus on relevant parts of the scene and ignore background noise;

– the method is tested on well-known datasets and its performance is evaluated using standard error metrics, such as MAE [17] and MSE [18].

The remainder of this paper is organized as follows.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

In Section 2, we describe our proposed method for crowd counting by considering scene depth and perspective. We explain how our approach deals with scale differences in images and introduce the depth module, which helps the model better understand spatial structure.

Section 2.1 gives an overview of the full architecture, and Sections 2.2 and 2.3 detail how the module works and how it fits into a typical neural network.

In Section 3, we describe how the model was trained and tested and present the results of our experiments on several benchmark datasets.

Section 4 discusses what these results mean, how the method performs in different conditions, and where it could be useful in practice. Finally, Section 5 concludes the study with the main findings and suggestions for future work.

2. Methodology

Images of crowd scenes often exhibit significant scale discrepancies among individuals due to magnification and perspective distortions, posing substantial challenges for general counting systems using uncalibrated camera systems.

Fig. 1 illustrates a crowd picture in which things nearer to the camera look much bigger than those at a greater distance. The size of things is inversely related to their distance from the plane of view of the camera. Common density-based methods often calculate crowd size by aggregating the density values within designated areas on a projected two-dimensional density map. Based on this premise, a crowd counting approach must adjust for differences in the counting object size and use scale-aware density values to obtain precise estimations.

Fig. 1 illustrates three distinct sections inside the picture, each with an identical pixel count. Nevertheless, each sector comprises a varying number of individuals owing to perspective distortions. Because the three regions have the same area, the densities in these places must also be different so that the estimates that come from adding up the densities for each region are correct.

The density of the outermost circle must exceed that of the adjacent portions. The counting technique must accommodate variations in the scale of density values, including factors such as perspective distortions and magnification. In traditional counting methods, objects are often «normalized» according to their viewpoint values before density estimations. This elucidates the variation in the item sizes.

Although convolutional neural network-based methods surpass manually crafted features, normalizing them with relevant extra information remains beneficial for addressing scale variances. Is it prudent to use the same scale variation processing in deep learning?

This study proposes feature representations that consider scale and perform scale-inclusion density estimations using underlying depth information.

Fig. 1. Three regions that occupy an equivalent number of pixels show varying populations: 3 in the left region, 20 in the center region, and 1 in the right region, due to individual scale changes. Photo source: www.pexels.com

Despite the typical divergence of size and depth in an item, the enhanced comprehension of scene depth by the network would increase its likelihood of identifying scale changes within the picture [19].

Most contemporary crowd counting datasets include just a single image at a time; hence, the depth results are obtained using a pretrained depth prediction model that requires only one image for operation. This enhances the applicability of the proposed approach.

This study proposes an image depth inclusion addition to scene depth information. The addition integrates depth information with weight recalibration at certain map feature sites to provide a scale-responsive depiction (Fig. 2).

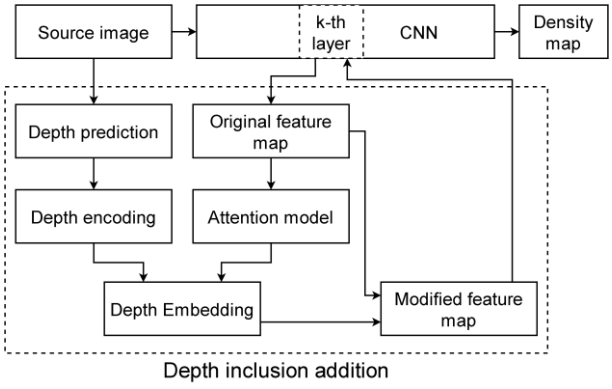

Fig. 2. Schematic overview of the proposed methodology

Subsequently, we will examine the structure of the suggested module, which consists of many components:

– the feature encoding segment first encodes the depth image into the features. Encoded depth information provides geometric information, but it neglects the overall context irrespective of the semantic content present at each location.

– the weight generation segment produces scale-aware weights using a spatial attention method to signify item placements, thus improving the encoded depth map. This aids the user in focusing on primary tasks while minimizing background distractions;

– the recalibration segment uses these weights to modify the original features at certain locations. Altered arrangements of diverse foreground objects exhibit geometrically distinct sizes when scale-sensitive weights are used to adjust features. This will immediately influence your assessment of scale-aware density values

3. Proposed method

3.1. Overall framework of the suggested methodology

Our fundamental approach for counting people in a crowd uses a widely used encoder-decoder-based technology [20]. Following the conversion of the input picture into high-level, multilayer feature maps by an encoder designed as a convolutional neural network, a decoder, likewise a convolutional neural network, transforms the feature maps into a spatial density map. The proposed approach employs a depth inclusion addition that produces scale-aware weights for different locations in the feature maps to enhance the original features by including substantial geometric characteristics.

We will explicitly assume that for each input picture *P*, there exists a corresponding depth image DI. Analyzing the k-th layer of the encoder, DI and the current features F^k of the convolutional neural network at layer k provide the scaling factors, referred to as scale-adjusted weights σ^k.The suggested approach then recalibrates the existing features are recalibrated recursively. T^k using scale-adjusted weights σ^k:

$$\begin{aligned} T^k &= CNN\left(F^{k-1}\right), \\ F^k &= f\left(T^k, \sigma^k\right), \\ \sigma^k &= FT^k\left(T^k, DI\right). \end{aligned} \quad (1)$$

The output of the preceding convolutional layers, referred to as T^k, DI – the anticipated depth picture generated by the pre-trained models, FT^k generates scale-sensitive weights inside the depth embedding module, f(·) modifies the attributes of the convolutional neural network using the produced weights. Ultimately, F^k represents the recalibrated weighted feature inside a convolutional neural network.

The output of the current layer functions as the input for the subsequent layer. This process continues until the network reaches the decoder, thereby converting the scale-aware representation into the scale-aware density value.

Depth map generation. Given that an individual's size is roughly equal to their distance from the camera, picture depth characteristics are utilized to indicate variations in scale across people in different locations. However, most of the existing crowd counting datasets consist only of single images. Using the approach described in [21], we obtained knowledge about image depth. This approach develops a deep convolutional neural field model to generate depth maps.

Fig. 3 shows the depth maps generated for the test crowd images. Based on the scene depth predictions we have collected, we are sure that the depth map generation module can handle different scene configurations and accurately show changes in distance at different points in relation to the camera picture plane.

Fig. 3. Visualization of predicted depth maps

3.2. Depth inclusion addition

The depth inclusion addition consists of three segments: feature encoding, weight generation, and recalibration. This study provides a comprehensive definition of each segment.

The feature encoding segment. Let DI represent the depth value derived by the depth generation module, where P, $P \in \mathbb{R}^{(H\times W\times 3)}$ and DI, $DI \in R^{(H\times W)}$. A depth-inclusion addition at layer k decreases the depth

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

picture by first adjusting it to conform to the dimensions of the feature mappings T^k at that layer.

To include scale-inclusion attributes, the approach allocates more weights to distant objects of reduced size. After acquiring sufficient distance data from the depth map, the depth values can be normalized to the range $(0, 1)$. This may be accomplished using nonlinear encoding with a parameterized sigmoid function as follows:

$$\phi^k = b(DI) = \frac{1}{1 + e^{-\psi^k DI + \omega^k}}, \quad (2)$$

where ψ^k and ω^k are parameters, that may enhance the encoding function. Optimizable differentiable functions may be addressed using traditional stochastic gradient descent techniques. To elucidate the objective function H, we shall examine its partial derivatives concerning the parameters ψ^k, which may be expressed as follows using the chain rule:

$$\frac{\partial H}{\partial \psi^k} = \frac{\partial H}{\partial \phi^k}\frac{\partial \phi^k}{\partial \psi^k} = \sum_j \frac{\partial H}{\partial \phi_j^k}\frac{\partial \phi^k}{\partial \psi^k} = \sum_j \frac{\partial H}{\partial \phi^k} DI_j b(DI_j)\left(1 - b(DI_j)\right), \quad (3)$$

where ϕ_j^k and DI_j denotes the j-th elements of ϕ^k and DI, respectively. The derivatives of the objective function H with respect to ω^k may be defined as follows:

$$\frac{\partial H}{\partial \omega^k} = -\sum_j \frac{\partial H}{\partial \phi_j^k} b(DI_j)\left(1 - b(DI_j)\right). \quad (4)$$

The weight generation segment. Depth lacks comprehensive understanding and is unable to distinguish between foreground and background objects despite providing information on size variation.

Moreover, an enhanced depth map would alter backdrop components, which is undesirable and might disrupt previously acquired feature re-representations. Objects located in distant regions of the sky with elevated depth values might have disproportionately large weights when using the initially acquired depth data. This may not only be irrelevant for assessing changes in target object size but also generate superfluous background noise.

We suggest including a specialized depth adjustment block to optimize the exploitation of the anticipated depth. Prior information about likely congested regions would enhance crowd population forecast precision.

His attention mask serves as a directive to disregard depth signals in surrounding areas and concentrate only on the profound detection of target items. The mask $a^k \in R^{(M\times N)}$ may be obtained by using feature maps $T^k \in \mathbb{R}^{M\times N\times C}$:

$$a^k = \text{sigmoid}\left(M(T^k)\right), \quad (5)$$

where M is an attention model with two convolutional layers, each featuring a 3x3 kernel size; the first layer has 512 filters, whereas the subsequent layer has 1 filter, using a convolutional neural network with two layers. First, a simple sigmoid function is used to find the most important areas in the entire space domain. Attention weights are then added to the output score map that comes from M.

In this scenario, counting individuals in a crowd focuses on the frontal areas where people are present. Fig. 4 presents several instances of trained attention masks.

The results in the second and third columns indicate a consistent trend as the support model becomes more intricate: what happens when the input feature maps are derived from many layers?

Attention masks significantly differentiate foreground parts from the background. The hierarchical feature representation of convolutional neural networks enables the generation of attention masks across several semantic levels. With increasing depth, attention masks concentrate on more abstract representations, ranging from extensive crowd regions to specific head positions.

Then, the attention mask modifies the encoded depth to provide scale-sensitive weights σ^k:

$$\sigma^k = FT^k(T^k, DI) = a^k \odot \phi^k, \quad (6)$$

where $\odot$ delineates the Hadamard matrix product operation as $((A \odot B)_{(i,j)} = (A)_{(i,j)}(B)_{(i,j)})$.

Fig. 4. Visualization of the attention mask

It is more straightforward to remove or rectify unwanted signals in the background of the encoded depth ϕ^k when the attention masks a^k are multiplied together.

The recalibration segment. A linear weight function, $f(\cdot)$, combined with a feedback loop using scale-aware weights, facilitates the adjustment of the output feature T^k.

In contrast to the more prevalent modulation approach that integrates information from several sources via produced weights, the $f(\cdot)$ function uses elementwise multiplication.

Consequently, feature activations at various junctures are modified to amalgamate geometric and semantic information at a specific location. The newly acquired scale-aware function F^k, which emphasizes the scale changes among foreground objects, can be characterized as follows:

$$F^k = f\left(T^k, \sigma^k\right) = T^k \boxtimes \sigma^k, \qquad (7)$$

where $\boxtimes$ denotes Hadamard matrix product's channel-wise operation.

3.3. Insertion depth inclusion addition into a neural network

The proposed depth inclusion addition, distinct from the other structure, has equal input and output dimensions. It can be easily included in a standard convolutional neural network design to augment representation capacity without modifying the original topology.

The encoder-decoder architecture we used to test our convolutional neural network against other convolutional networks with different levels of complexity showed that it worked well.

The suggested depth inclusion addition was added to the encoder sections in both the shallow and deep convolutional neural network base models.

We first developed a lightweight model using three convolutional layers in the encoder and decoder components. This completely convolutional model can convert any amount of input data into its output.

The VGG network [22] was enhanced for crowd counting using an advanced CSRNet [23] architecture based on dilation preprocessing. In addition, each convolutional layer has a rectified linear unit that is set up correctly to maintain the spatial resolution.

Training neural network designs is an important initial stage. A pixel-by-pixel Euclidean loss function can be used to train the suggested neural network designs as follows:

$$L = \left\| DM - \widehat{DM} \right\|^2, \qquad (8)$$

where DM and $\widehat{DM}$ are the projected and actual density maps, respectively, image P coupled with its point annotation set A_p; the actual density map is expressed as the aggregate of two-dimensional Gaussian functions centered at every point, i.e. $\forall k \in P, \widehat{DM}(k) = \sum_{\mu \in A_P} \mathbb{N}(k; \mu, \sum)$, where $\mathbb{N}(k; \mu, \sum)$ represents a normalized two-dimensional Gaussian kernel assessed at k with mean μ and isotropic covariance matrix Σ.

Three stages constitute the instruction: First, the loss function L is used to improve the base model. Second, an attention model is included and trained to improve the initialization. Finally, the whole deep learning module is constructed using the loss function L and trained from beginning to end.

Experimental setup. The testing environment was configured using Python on an Nvidia GeForce RTX 3060 graphics card. The weight decay value is established at 0.00045, while the impulse value is fixed at 0.8 and the learning rate begins at 10^{-5}.

To augment the training data, image segments are randomly extracted from the training images for each dataset; random flipping of these segments is used to enhance the dataset.

The final count is obtained by aggregating the density data across the picture. Several studies [24] have shown that mean absolute error (MAE) and mean square error (MSE) are useful ways to measure and contrast how well counting works.

The performance of a regression model is quantitatively assessed using the mean absolute error (MAE). The mean absolute deviation between the predicted values (model outputs) and the actual values of the empirical data is quantified.

MAE is less vulnerable to outliers than MSE because it does not square the errors. The following formula computes the MAE statistics:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|, \qquad (9)$$

where n represent the total number of observations in the sample; y_i denotes the actual number of individuals in the image; $\hat{y}_i$ signifies the expected number of individuals in the image; and $|y_i - \hat{y}_i|$ indicates the absolute difference between the actual and anticipated values. The mean squared error (MSE) is a commonly used method to assess the efficacy of a predictive or regression model The mean square of the discrepancies between projected values (model outputs) and actual values (empirical data) is quantified as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \qquad (10)$$

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

where n represent the total number of observations in the sample, y_i denote the actual number of individuals in the image, $\hat{y}_i$ signify the anticipated number of individuals in the image and $(y_i - \hat{y}_i)^2$ represent the square of the deviation between the actual and anticipated values.

4. Experimental results

To determine the optimal methodology, we evaluated the proposed method against many novel techniques.

ShanghaiTech-B dataset [25]. Table 1 demonstrates that the proposed method outperforms the leading techniques on this dataset. MCNN and Crowd-CNN, which mostly use multi-scale features to deal with differences in scale, don't work as well as the proposed depth-aware module. This demonstrates the effectiveness of using depth to explicitly reflect size changes in crowd estimates.

Table 1

Experimental results on the ShanghaiTech-B dataset

Method	MAE	MSE
MCNN [25]	26.3	41.5
DecideNet [26]	20.8	29.3
Crowd-CNN [27]	31.9	49.9
IG-CNN [28]	13.5	21.0
Proposed method	**9.1**	**16.4**

UCF_CC_50 dataset [29]. Table 2 shows that the suggested strategy exhibits superior MAE and MSE compared with other prominent solutions. This indicates that the proposed strategy is effective in densely populated areas.

Table 2

Experimental results on the UCF_CC_50 dataset

Method	MAE	MSE
MCNN [25]	377.4	510.2
DecideNet [26]	361.8	493.5
Crowd-CNN [27]	466.9	498.6
IG-CNN [28]	419.6	541.8
Proposed method	**289.5**	**391.1**

Mall dataset [30]. Table 3 shows that the suggested method outperformed the best current methods in terms of MAE and MSE. This proves that the strategy works well and is reliable on small datasets with few subjects.

Table 3

Experimental results on the Mall dataset

Method	MAE	MSE
MCNN [25]	2.74	13.5
DecideNet [26]	1.53	1.89
Crowd-CNN [27]	1.83	2.76
IG-CNN [28]	2.39	9.14
Proposed method	**1.31**	**1.64**

Statistical Confidence and Validation. We performed several inference runs (n = 5) for each dataset with various randomly chosen validation splits to assess the statistical credibility of the findings. For each measure (MAE and MSE), the mean and standard deviation were computed, and the 95% confidence interval was presented under the assumption of an essentially normal distribution of errors:

$$\mathrm{CI}_{95\%} = \bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \tag{11}$$

where $\bar{x}$ is the sample mean, σ is the standard deviation, and $n = 5$ is the number of independent runs.

The confidence intervals for all datasets exhibited minimal volatility and did not intersect with baseline approaches, thereby reinforcing the enhancement's statistical significance. A paired t-test was conducted between our approach and the most robust baseline (DecideNet), yielding $p < 0.05$, further confirming the advantage of our approach.

Trustworthiness of the results. Trust in the outcomes of studies is crucial for scientific rigor and repeatability. Deep learning systems are intrinsically stochastic because of random initialization and data augmentation. We have implemented many methodological measures to ensure that the findings shown in Tables 1–3 are trustworthy, consistent, and aligned with the typical assessment techniques used in the computer vision domain.

a) Evaluation of established public benchmarks. The three datasets used for all trials were ShanghaiTech-B, UCF_CC_50, and Mall. These datasets are well-recognized and are readily accessible to anyone. Numerous previous studies have used these criteria, demonstrating a broad spectrum of density levels, ambient circumstances, and prospective aberrations. Consequently, they provide a balanced testing environment suitable for drawing generalizations;

b) Consistent training and evaluation pipeline. A uniform training strategy is used for all models to minimize variability across tests. This procedure included preprocessing, augmenting, and computing data loss. All models used the same optimizer, learning rate schedule, and epoch count during training. This controlled environment may facilitate the observation of the independent effects of the technique;

c) Architectural reproducibility. The proposed model uses a conventional encoder-decoder architecture, although it has benefits that are comprehensively elaborated upon. The incorporation of the attention mechanism and the depth module, both of which are entirely modular, requires no human changes or dataset-specific modifications. This facilitates future researchers in replicating the study and confirming the findings under similar circumstances encountered by the original researchers;

d) Internal cross-checks and qualitative validation. Intermediate outputs, such as attention masks and anticipated depth maps, were examined across all test samples to assess internal consistency. We examined the spatial coherence and alignment of these outputs with high-density areas to qualitatively assess the model's behavior. The visual findings closely align with the density projections, indicating the success of the recalibration method and affirming the system's effectiveness;

e) Alignment with prior art. A recent study has demonstrated the significance of recognizing one's perspective and using depth cues to enhance crowd density assessments. The observed performance patterns are consistent with the existing theories and findings from previous studies on the subject. This theoretical foundation significantly enhances confidence in the observed benefits' probability.

5. Discussion

The experimental findings of this study demonstrate that integrating depth-awareness characteristics into crowd counting models may substantially enhance their performance, even in scenarios with highly distorted perspectives. Our strategy surpassed the leading methods in accuracy across all three benchmark datasets: ShanghaiTech-B, UCF_CC_50, and Mall. This applies to both the MAE and the MSE metrics.

The reduced MAE on the ShanghaiTech-B dataset illustrates this, indicating that the model performs well in standard crowd scenarios where people are distributed across various distances. The model yields consistent and reliable estimates when evaluated on the challenging UCF_CC_50 dataset, characterized by significant occlusions and density. The strategy is also successful when used with the Mall dataset, which is smaller in size, has fewer people, and is conducted in more controlled indoor settings.

Attention mechanisms aid the network by directing the model to concentrate on certain elements of interest to the audience while ignoring irrelevant background information. This enhances the clarity and consistency of density maps regarding their spatial organization, thereby facilitating comprehension and identification.

Novelty and Contribution. This study significantly contributes by integrating the recalibration of internal features of convolutional neural networks for crowd counting with monocular depth estimation. Using a dedicated depth inclusion module, our model swiftly incorporates depth signals and learns to augment them via attention-based spatial weighting. A contrast may be made between this strategy and others that either regard perspective alterations as noise or require the use of manually designed geometric priors.

The proposed depth-aware module is not only easy to differentiate but also lightweight and compatible with conventional encoder-decoder architectures. Moreover, it preserves architectural adaptability, yielding a more reactive spatial geometry. The model may modify its environmental perception by emphasizing recalibration via learnt attention masks without requiring explicit calibration or fresh annotations.

Practical applications and benefits. The proposed approach has significant promise for use in real-world situations.

– intelligent surveillance systems, i.e., those that monitor extensive public areas with complex camera angles;

– overseeing event and crowd safety necessitates the capability to accurately assess the number of persons in attendance at any given moment to effectively make decisions and assess possible risks.

– to enhance infrastructure efficiency, smart city analytics include tasks such as pedestrian traffic modelling and occupancy forecasting.

The capacity of the model to enhance the spatial quality of density map generation facilitates future comprehension and assessment, such as identifying traffic bottlenecks or modeling behavior. If further refined, it might be used on edge devices or embedded systems equipped with mid-range GPUs.

Limitations. Although the proposed model is useful, it has some shortcomings:

– it employs monocular depth estimators trained on extensive datasets, which may not perform optimally in all surveillance environments. This is particularly applicable in situations with low light, little contrast, or obstructed visibility;

– the depth and attention modules may be unsuitable for low-power embedded deployment or stringent real-time requirements due to their higher processing power demands;

– the present technique's analysis of individual frames, without accounting for temporal factors, renders it less effective for video sequences with crowd dynamics and motion cues.

Future work. Future enhancements may be pursued in the following areas:

– the use of video streams to include recurrent or transformer-based modules for crowd counting in both spatial and temporal dimensions exemplifies durational extension;

– to obtain robust depth estimation, the depth estimation module must be altered or retrained, particularly for surveillance images;

– real-time deployment entails reducing and quantifying the architecture to enhance its suitability for edge device use;

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

– the technique of merging crowd density estimates with semantic segmentation is referred to as semantic integration. This methodology aims to facilitate navigation in intricate surroundings and to eradicate background noise.

6. Conclusions

In this study, we proposed a depth-aware approach to crowd counting that integrates perspective information directly into the feature processing pipeline of convolutional neural networks. By introducing a specially designed depth inclusion module, our method enables the model to better handle scale variations caused by perspective distortions. The module was incorporated into an encoder-decoder architecture and trained end-to-end using pixel-level Euclidean loss on predicted density maps.

The results across three widely used benchmark datasets – ShanghaiTech-B, UCF_CC_50, and Mall– demonstrate that our method provides consistent improvements over existing approaches. On ShanghaiTech-B, the proposed model reduced the mean absolute error (MAE) from 13.5 (IG-CNN) to 9.1 and the mean squared error (MSE) from 21.0 to 16.4. On the highly congested UCF_CC_50 dataset, the MAE dropped from 361.8 (DecideNet) to 289.5, and the MSE from 493.5 to 391.1, respectively. Even on the relatively small and less crowded Mall dataset, our approach lowered the MAE and MSE to 1.31 and 1.64, respectively, surpassing prior methods like DecideNet and Crowd-CNN.

These results highlight the effectiveness of incorporating geometric depth priors into deep learning models for crowd analysis. Our approach not only delivers better accuracy but also improved spatial coherence in the generated density maps, which is an important factor in practical surveillance scenarios.

Looking ahead, there are several promising directions for future research. One avenue is to extend the model to video-based crowd counting, which would require handling temporal information and motion cues. Another is enhancing the depth estimation robustness in difficult conditions, such as low-light or occluded scenes. Furthermore, optimizing the model for real-time inference on low-power edge devices would support deployment in live surveillance systems. Finally, integrating semantic scene understanding, such as separating crowd regions from background clutter, may further improve the interpretability and reliability of predictions.

Contributions of authors: formulation of tasks, method development – **Ruslan Dobryshev**; verification, analysis of the results – **Sergiy Purish**; review and editing – **Mykhaylo Lobachev**; writing, original draft preparation – **Mykola Hodovychenko**.

Conflict of interest

The authors declare that they have no conflict of interest concerning this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

This research was conducted without financial support.

Data Availability

The manuscript contains no associated data.

Use of Artificial Intelligence

Generative AI tools have been used for grammar checks and text polishing.

All authors have read and agreed to the publication of the finale version of this manuscript.

References

1. Deng, L., Zhou, Q., Wang, S., Górriz, J.M. & Zhang, Y. Deep learning in crowd counting: A survey. *CAAI Transactions on Intelligence Technology*, 2024, vol. 9, no. 5, pp.1043–1077. DOI: 10.1049/cit2.12241.

2. Patwal, A., Diwakar, M., Tripathi, V. & Singh, P. Crowd counting analysis using deep learning: A critical review. *Procedia Computer Science*, 2023, vol. 218, pp. 2448–2458. DOI: 10.1016/j.procs.2023.01.220.

3. Alhawsawi, A.N., Khan, S.D. & Ur Rehman, F. Crowd counting in diverse environments using a deep routing mechanism informed by crowd density levels. *Information*, 2024, vol. 15, no. 5, article no. 275. DOI: 10.3390/info15050275.

4. Khan, M.A., Menouar, H., Hamila, R. & Abu-Dayya, A. Crowd counting at the edge using weighted knowledge distillation. *Scientific Reports*, 2025, vol. 15, article no. 11932. DOI: 10.1038/s41598-025-90750-5.

5. Mansouri, W., Alohali, M.A., Alqahtani, H., Alruwais, N., Alshammeri, M. & Mahmud, A. Deep CNN-based enhanced crowd density monitoring for intelligent urban planning on smart cities. *Scientific Reports*, 2025, vol. 15, article no. 5759. DOI: 10.1038/s41598-025-90430-4.

6. Zeng, X., Wang, H., Guo, Q. & Wu, Y. Correlation-attention guided regression network for efficient crowd counting. *Journal of Visual Communication and Image Representation*, 2024, vol. 99, article no. 104078. DOI: 10.1016/j.jvcir.2024.104078.

7. Cai, Y. & Zhang, D. A weakly supervised crowd counting method via combining CNN and Transformer. *Electronics*, 2024, vol. 13, no. 24, article no. 5053. DOI: 10.3390/electronics13245053.

8. Lien, C.-C. & Wu, P.-C. A crowded object counting system with self-attention mechanism. *Sensors*, 2024, vol. 24, no. 20, article no. 6612. DOI: 10.3390/s24206612.

9. Alhawsawi, A.N., Khan, S.D. & Rehman, F.U. Enhanced YOLOv8-based model with context enrichment module for crowd counting in complex drone imagery. *Remote Sensing*, 2024, vol. 16, no. 22, article no. 4175. DOI: 10.3390/rs16224175.

10. Yaseen, M. What is YOLOv8: An in-depth exploration of the internal features of the next-generation object detector. *arXiv*, 2024, Aug. Available at: https://doi.org/10.48550/arXiv.2408.15857 (Accessed: 1 March 2025).

11. Zhao, Z., Ma, P., Jia, M., Wang, X. & Hei, X. A dilated CNN for cross-layers of contextual information in congested crowd counting. *Sensors*, 2024, vol. 24, no. 6, article no. 1816. DOI: 10.3390/s24061816.

12. Tomar, A., Nijhawan, R. & Koundal, D. EDCCN: A benchmark encoder–decoder framework for accurate crowd counting. *Neurocomputing*, 2025, vol. 640, article no. 130304. DOI: 10.1016/j.neucom.2025.130304.

13. Li, Y.-C., Jia, R.-S., Hu, Y.-X. & Sun, H.-M. A weakly-supervised crowd density estimation method based on two-stage linear feature calibration. *IEEE/CAA Journal of Automatica Sinica*, 2024, vol. 11, no. 4, pp. 965–981. DOI: 10.1109/JAS.2023.123960.

14. Zhou, J., Zhang, J. & Gui, Y. Crowd counting in domain generalization based on multi-scale attention and hierarchy level enhancement. *Scientific Reports*, 2025, vol. 15, article no. 155. DOI: 10.1038/s41598-024-83725-5.

15. Cao, R., Yu, J., Liu, Z. & Liang, Q. Towards real-world monitoring: An improved point prediction method for crowd counting based on contrastive learning. *PLoS ONE*, 2025, vol. 20, no. 7, article no. e0327397. DOI: 10.1371/journal.pone.0327397.

16. Xu, M., Ge, Z., Jiang, X., Cui, G., Lv, P., Zhou, B. & Xu, C. Depth information guided crowd counting for complex crowd scenes. *arXiv*, 2018, Mar. Available at: https://arxiv.org/abs/1803.02256 (Accessed: 1 March 2025).

17. Willmott, C.J. & Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 2005, vol. 30, no. 1, pp. 79–82. DOI: 10.3354/cr030079.

18. Sindagi, V.A. & Patel, V.M. Generating high-quality crowd density maps using contextual pyramid CNNs. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, IEEE, 2017, pp.1861–1870. DOI: 10.1109/ICCV.2017.206.

19. Gouiaa, R., Akhloufi, M.A. & Shahbazi, M. Advances in convolution neural networks based crowd counting and density estimation. *Big Data and Cognitive Computing*, 2021, vol. 5, no. 4, article no. 50. DOI: 10.3390/bdcc5040050.

20. Shen, Z., Xu, Y., Ni, B., Wang, M., Hu, J. & Yang, X. Crowd counting via adversarial cross-scale consistency pursuit. *In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 18–22 June 2018. IEEE, pp. 5245–5254. DOI: 10.1109/CVPR.2018.00550.

21. Liu, F., Shen, C., Lin, G. & Reid, I. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, vol. 38, no. 10, pp. 2024–2039. DOI: 10.1109/TPAMI.2015.2505283.

22. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *In: 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA, 7–9 May 2015. pp. 1–14. DOI: 10.48550/arXiv.1409.1556.

23. Li, Y., Zhang, X. & Chen, D. CSR-Net: Dilated convolutional neural networks for understanding the highly congested scenes. *In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 18–22 June 2018. IEEE, pp. 1091–1100. DOI: 10.1109/CVPR.2018.00120.

24. Chen, L., Gao, X., Chao, F., Chang, X., Lin, C.M., Gao, X., Lin, S., Zhang, H. & Lin, J. The effectiveness of a simplified model structure for crowd counting. *arXiv*, 2024. Available at: https://arxiv.org/abs/2404.07847. (Accessed: 1 March 2025).

25. Zhang, Y., Zhou, D., Chen, S., Gao, S. & Ma, Y. Single‐image crowd counting via multi-column convolutional neural network. *In: 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016. IEEE, pp. 589–597. DOI: 10.1109/CVPR.2016.70.

26. Liu, J., Gao, C., Meng, D. & Hauptmann, A.G. DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation. *In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5197–5206. DOI: 10.1109/CVPR.2018.00545.

27. Zhang, C., Li, H., Wang, X. & Yang, X. Cross-Scene Crowd Counting via Deep Convolutional Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 07–12 June 2015, IEEE, pp. 833–841. DOI: 10.1109/CVPR.2015.7298684.

28. Sam, D.B., Sajjan, N.N. & Babu, R.V. Divide and Grow: Capturing Huge Diversity in Crowd Images with Incrementally Growing CNN. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018, IEEE, pp. 3618–3626. DOI: 10.1109/CVPR.2018.00381.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

29. Idrees, H., Saleemi, I., Seibert, C. & Shah, M. Multi-source multi-scale counting in extremely dense crowd images. *In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, USA, 23–28 June 2013. IEEE, pp. 2547–2554. DOI: 10.1109/CVPR.2013.329.

30. Loy, C.C., Gong, S. & Xiang, T. *Mall dataset: A sparse indoor crowd counting and profiling dataset collected from webcam images.* Available at: https://personal.ie.cuhk.edu.hk/~ccloy/downloads_mall_dataset.html (accessed: 1 March 2025).

Received 17.03.2025, Accepted 20.05.2025

ПІДРАХУНОК НАТОВПУ В ІНТЕЛЕКТУАЛЬНИХ СИСТЕМАХ ВІДЕОСПОСТЕРЕЖЕННЯ

Р. Є. Добришев, С. В. Пуріш, М. В. Лобачев, М. А. Годовиченко

Дослідження фокусується на підвищенні точності та надійності підрахунку натовпу в інтелектуальних системах відеоспостереження шляхом включення розуміння перспективи в моделі глибокого навчання. Традиційні згорткові нейронні мережі часто борються з варіаціями масштабу, спричиненими спотвореннями перспективи та закриттям об'єктів у сценах щільного натовпу. **Мета** полягає в розробці методу, який використовує інформацію про геометричну глибину для покращення просторової узгодженості оцінок щільності та надання більш надійних прогнозів у різних конфігураціях сцен, включаючи сильно перевантажені або нерегулярні середовища. **Завдання, які необхідно виконати,** включають розробку модуля включення глибини, інтеграцію його в архітектуру кодера-декодера, створення карт глибини з монокулярних RGB-зображень і рекалібрування представлень об'єктів за допомогою механізмів, що враховують увагу та масштабування. **Методи, що використовуються,** включають вилучення глибинних ознак із зображень за допомогою попередньо навченої моделі оцінки глибини з подальшим просторовим калібруванням карт ознак на основі уваги для виділення об'єктів на передньому плані і придушення нерелевантних фонових сигналів. Реалізовано повністю диференційований конвеєр для забезпечення безперешкодної інтеграції в стандартні фреймворки CNN. Крім того, процедура навчання мережі включає евклідові функції втрат на картах щільності на рівні пікселів для оптимізації прогнозування, чутливого до масштабу. Запропонований метод оцінено на еталонних наборах даних, включаючи ShanghaiTech-B, UCF_CC_50 та Mall, де він послідовно перевершує найсучасніші моделі за показниками MAE та MSE. **Висновки,** зроблені на основі результатів експерименту, підтверджують, що явне включення представлень з урахуванням глибини значно покращує продуктивність підрахунку, особливо у сценаріях із серйозними диспропорціями масштабу, спричиненими перспективою. Інтеграція геометричних попередніх даних у моделі, керовані даними, є перспективним напрямком для спостереження в реальному часі та великомасштабного моніторингу натовпу, забезпечуючи не тільки кількісні покращення, але й більшу просторову точність при створенні карт щільності та кращу адаптивність до складних візуальних умов.

Ключові слова: підрахунок натовпу; відеоспостереження; глибинне навчання; згорткові нейронні мережі; оцінка глибини; карта щільності; спотворення перспективи; механізм уваги.

Добришев Руслан Євгенович – асп. каф. штучного інтелекту та аналізу даних, Національний університет «Одеська Політехніка», Одеса, Україна.

Пуріш Сергій Володимирович – канд. техн. наук, генеральний директор Enestech Software, Вілмінгтон, США

Лобачев Михайло Володимирович – канд. техн. наук, проф., директор інституту штучного інтелекту та робототехніки, Національний університет «Одеська Політехніка», Одеса, Україна.

Годовиченко Микола Анатолійович – канд. техн. наук, доц. каф. штучного інтелекту та аналізу даних, Національний університет «Одеська Політехніка», Одеса, Україна.

Ruslan Dobryshev – PhD Student of the Artificial Intelligence and Data Analysis Department, Odesa Polytechnic National University, Odesa, Ukraine.
e-mail: rdobrishev@gmail.com, ORCID: 0009-0007-8639-3157.

Sergiy Purish – PhD, CEO of Enestech Software, Wilmington, USA,
e-mail: spurish@gmail.com, ORCID: 0009-0009-0346-842X.

Mykhaylo Lobachev – PhD, Professor, Head of the Institute of Artificial Intelligence and Robotics, Odessa Polytechnic National University, Odesa, Ukraine,
e-mail: lobachevmv@gmail.com, ORCID: 0000-0002-4859-304X, Scopus Author ID: 36845971100.

Mykola Hodovychenko – PhD, Assoc. Prof. at the Artificial Intelligence and Data Analysis Department, Odessa National Polytechnic University, Odesa, Ukraine,
e-mail: hodovychenko@od.edu.ua, ORCID: 0000-0001-5422-3048, Scopus Author ID: 57188700773.