Tourist Density Estimation Based On Lightweight Swin-Transformer

Xuxiang Zhang*

School of Finance and Economics, Zhengzhou University of Science and Technology, Zhengzhou, China

** Corresponding author. E-mail: aqiufenga@163.com*

Received: Aug. 24, 2025; Accepted: Oct. 28, 2025

There are two problems in traditional population counting models. (1) The complex heavy-duty counting models have strong counting performance, but they have excessive model parameters and computational costs, thus lacking practicality. (2) The current lightweight models have reduced the complexity of the models, but their counting performance is poor. Therefore, this paper proposes a novel tourist density estimation based on lightweight Swin-Transformer. The proposed method takes advantage of the distinct encoding advantages of Swin-Transformer and convolutional neural network (CNN), effectively capturing the global semantic information and local details of image features, thereby enhancing the model's expressive power. To minimize the loss of feature details during down-sampling, a multi-scale resolution feature pyramid pooling (MFPP) module is designed. By combining features from different dimensions, it acquires more contextual information at different scales and enhances the expression of local details. Various advanced methods are compared on three population datasets. The experimental results show that all the indicators of the proposed framework perform exceptionally well, effectively alleviating the scale differences in tourist counting, generating high-fidelity density maps and enhancing the generalization ability of the model.

Keywords: tourist density estimation, lightweight Swin-Transformer, CNN, multi-scale resolution feature pyramid pooling

http://dx.doi.org/10.6180/jase.202606_29(6).0024

1. Introduction

Crowd counting, as a core direction in image understanding, aims to automatically count the total number of human heads within the surveillance field. With the acceleration of urbanization, large-scale gatherings are becoming increasingly frequent, and the risk of stampedes triggered by dense crowds is also rising [1]. Therefore, the estimation of the number of people in high-density scenarios has become a focus of academic research [2].

Recently, crowd counting based on CNN has attracted more attention from scholars, and many high-quality counting models have been proposed. Wang et al. [3] introduced a context-pyramid CNN that interweaved global context with local details, yielding sharp density maps despite abrupt crowd fluctuations. Oh et al. [4] employed the heavy-duty classic ResNet-50 [5] as the backbone network to extract features, and multiple parallel convolutions to generate density maps. It probabilistically interpreted the model's output through the quantification of uncertainty, thereby improving the quality of predictions and explicitly handling uncertain inputs and special cases. Li et al. [6] employed a complex backbone network for feature extraction and used expansion convolution as the backend to generate density maps. Liu et al. [7] inserted a context-aware module into CSRNet, enabling information about different head sizes to be reflected in the feature maps. Meng et al. [8] devised a teacher-student architecture that harnesses spatial uncertainty to spotlight the most reliable regions. Wang et al. [9] employed a scale tree diversity enhancer and a multi-level auxiliaryer to alleviate the limitations caused by insufficient scale levels in existing methods. However, in order to extract more features, they generally used com-

plex multi-scale convolutions as the backbone. While multi-scale cues are captured, the accompanying parameter surge and computational load remain heavy; indeed, these approaches yield accurate density maps and strong counting results, yet at substantial cost. The parameter quantity and computational cost are very large, consuming a lot of resources during the training process. In today's mobile- and edge-centric landscape, squeezing robust counting accuracy out of tight computational envelopes is nearly as critical as the accuracy itself [10].

Therefore, how to develop an efficient and lightweight crowd counting model has become a hot research topic. In Lin et al. [11], each column used fewer convolutional layers with different-sized convolution kernels to predict the crowd density map. Sam and Babu [12] proposed a top-down feedback convolutional neural network, which used a small number of standard convolutions with two different convolution kernels to generate density maps. However, the features extracted by the few standard convolutional layers were very limited, which leaded to unsatisfactory accuracy of the final count. Gao et al. [13] proposed to encode the intermediate features through foreground/background segmentation (FBS) to separate the foreground and background. Ma et al. [14] devised a lean, end-to-end architecture that first harvested multi-scale cues via a scale-attentive unit and then mapped them to density outputs, while successive small-kernel convolutions curbed model complexity. Liang et al. [15] used the proposed lightweight pyramid convolution module for multi-scale feature extraction. Yi et al. [16] adopted MobileViT as the encoder, trimming both parameters and FLOPs. However, in order to reduce complexity, these models mostly focus on feature extraction while neglecting feature fusion. This inevitably leads to a loss of counting accuracy. In conclusion, the current lightweight crowd counting networks have reduced both the number of parameters and the computational cost. However, to achieve faster running speed, the existing lightweight networks often use fewer standard convolutions or ignore feature fusion, avoiding some similar feature maps and discarding some highly similar features [17]. Although the model complexity has decreased, the counting accuracy has also been affected.

The distribution of the crowd in the image is often uneven and irregular. It may densely gather in certain areas while being sparse in other areas. Meanwhile, the heads in the image often exist at different scales. Using a single-scale feature extractor may lead to a decline in performance when dealing with targets of different scales. Through multi-scale convolution operations, the model simultaneously focuses on global and local information, capturing the distribution of the crowd at different scales from the image, which helps improve the model's accurate understanding of the crowd density distribution and enables it to better handle occlusions and irregular distributions within the crowd. To counter the persistent under-counting in ultra-dense regions and the poor cross-scale generalization caused by extreme size variations, we present a dual-encoder crowd-density estimator that upgrades the Swin-Transformer backbone. By using CNN to learn local detailed features and Swin-Transformer to capture global contextual information, the ability to learn features is effectively enhanced. To curb the spatial-detail erosion inherent in down-sampling, we introduce the Multi-scale Feature Pyramid Pooling (MFPP) module, which harvests contextual cues at diverse resolutions to reinforce robustness. Concurrently, Coordinate Attention (CA) is injected into skip connections to encode precise positional signals, enabling pinpoint crowd localization.

2. Materials and methods

2.1. Swin-Transformer encoding branch

The traditional Transformer calculates the dependencies between each image patch (Token) across the entire image, which can lead to high redundancy and large computational costs. Swin-Transformer slices the image into disjoint windows and restricts self-attention to the patches inside each window, reducing the computational cost from a quadratic to a linear scale [18, 19]. Additionally, it employs a sliding window mechanism to enhance information exchange among features, as illustrated in Fig. 1.

Each Swin-Transformer layer pairs two blocks containing layer-norm (LM), multi-layer perception (MLP), window-based multi-head self-attention (W-MSA) and its shifted variant (SW-MSA). We set the initial dimension of Swin-Transformer to 96, the window size to 8, and the corresponding number of heads in WMSA and SW-MSA to 3, 6, 12, and 24 respectively. For the Swin-Transformer module, the outputs of the l - th layer in W-MSA and SW-MSA are the encoding vectors of Z^l and Z^{l+1} respectively. The specific calculation method is as follows.

$$Z^l = W - MSA\left[LM\left(Z^{l-1}\right)\right] + Z^{l-1} \tag{1}$$

$$Z^{l+1} = MLP\left[LM\left(\hat{Z}^l\right)\right] + \hat{Z}^l \tag{2}$$

$$\hat{Z}^{l+1} = MLP\left[LM\left(Z^{l+1}\right)\right] + Z^{l+1} \tag{3}$$

2.2. Dual-encoding network

The overall dual-encoding network is shown in Fig. 2. For the given image $X \in R^{H\times W\times C}$, after extracting its shallow semantic information of the image, it is encoded in the

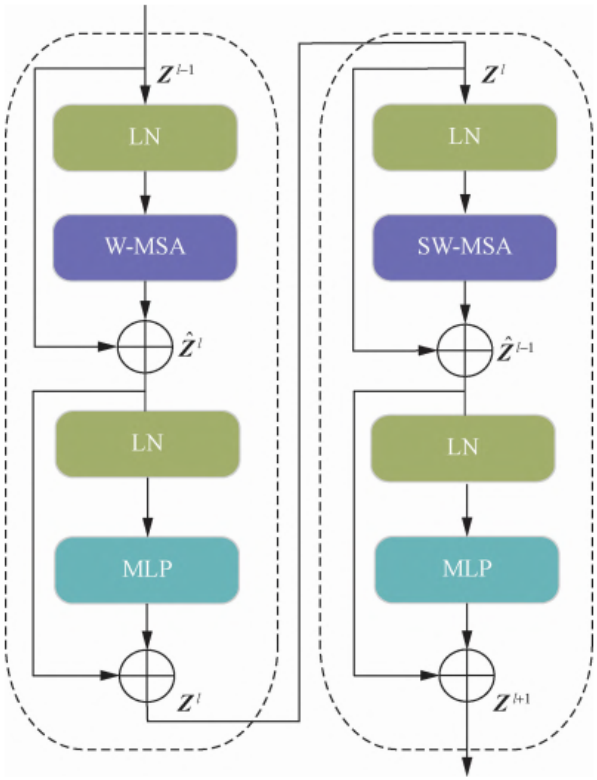

Fig. 1. Swin-Transformer structure

dual-encoding module composed of Swin-Transformer and CNN. Meanwhile, a MFPP module is constructed at the bottom of the model to fuse feature maps of different scales and obtain rich contextual information. After concatenating the feature maps generated by CNN and the vectors produced by Swin-Transformer, they are input into the CA module. The position information is integrated into the feature channels. Finally, through progressive up-sampling, the size of the feature map is restored to obtain accurate segmentation results.

2.3. MFPP module

With deeper down-sampling stages in the architecture, the feature information becomes increasingly abstract, while many local detailed features are lost, resulting in insufficient learning of the semantic information of the features by the network model, and affecting the segmentation accuracy of surgical instruments. To capture and integrate multi-scale semantics while promoting cross-scale feature dialogue. Based on the dilated spatial pooling module in the DeepLabV2 [20] model, a multi-scale resolution feature pyramid pooling (MFPP) module is constructed, as shown in Fig. 3. Firstly, different scales of features are fused and concatenated through convolution operations. Then, an input of Conv 1×1 convolution and four Conv

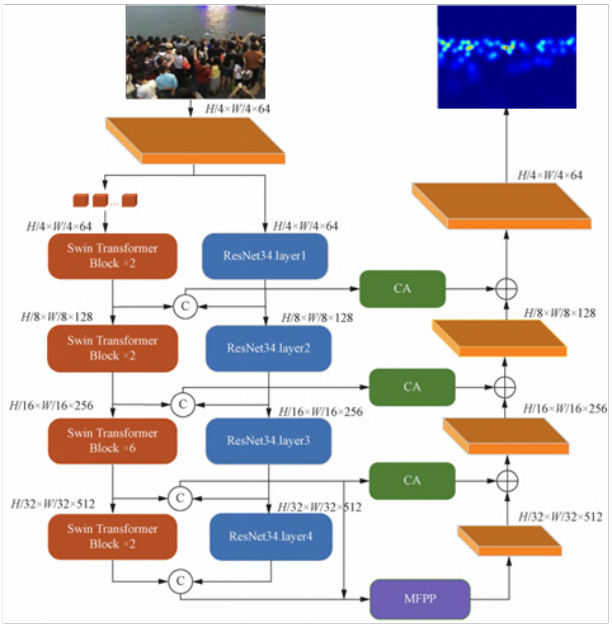

Fig. 2. Crowd counting network structure

3×3 convolutions with different receptive field sizes is provided, where the dilation rates of the Conv 3×3 convolutions are 1, 6, 12, and 18 respectively. By capturing the contextual semantic information in different receptive fields, the model's expression ability and generalization ability are improved. For different dilation rates of features, to learn the channel ratios between different channels, the SE (squeeze-and-excitation) attention module [21] is further introduced. By dynamically adjusting the relationship between different feature channels, the proportion of the interested channels is increased. Finally, the features adjusted by the SE attention module and the features after Conv 1×1 convolution are concatenated, and the specific formula is as follows.

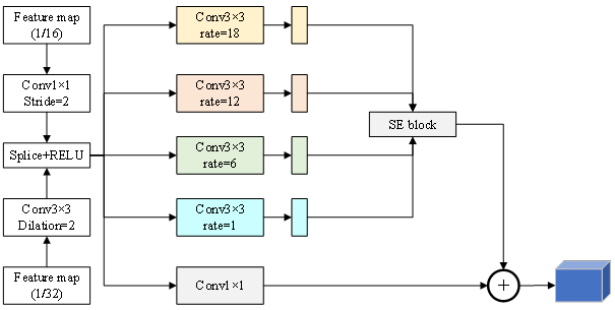

Fig. 3. MFPP module

$$X3 = \text{Concate}\,[X1, \text{Conv}\,2 \times 2(X2)] \tag{4}$$

$$Y1 = f_{se}\,[\text{Concate}\,(\text{Conv}\,v_{dr}(X3, i))] \tag{5}$$

$$Y2 = Y1 + \text{Conv}\,1 \times 1(X3) \tag{6}$$

Where $X1$, $X2$, $X3$ are feature vectors of different scales. Concate represents feature concatenation. Conv_{dr} is the dilated convolution operation. i is the dilation rate, which takes values of 1, 6, 12, and 18 respectively. f_{se} is the function of the SE module. $Y1$ is the feature vector after the channel weights are adjusted by the function f_{se}.$Y2$ is the final output feature vector.

To enable an intricate fusion of the heterogeneous feature streams delivered by ResNet34 and the Swin-Transformer, a sophisticated Coordinate Attention (CA) mechanism is interposed, adaptively aligning spatial-semantic cues across both branches while preserving their complementary discriminative power. The common attention mechanism only focuses on the improvement of channel attention, while ignoring the acquisition of position information. The common attention mechanism only focuses on the improvement of channel attention, while ignoring the acquisition of position information. The coordinate attention module performs global average pooling in two directions on the features, one in the horizontal spatial direction to capture long-range dependencies, and the other in the vertical spatial direction to preserve precise position information, forming a pair of directionally-aware and position-sensitive feature maps. Through operations such as concatenation, convolution, and normalization, the feature representation of the surgical instrument area is enhanced. The coordinate attention module performs global average pooling in two directions on the features, one in the horizontal spatial direction to capture long-range dependencies, and the other in the vertical spatial direction to preserve precise position information, forming a pair of directionally-aware and position-sensitive feature maps. Through operations such as concatenation, convolution, and normalization, the feature representation of the surgical instrument area is enhanced.

2.4. Loss function

During training, the discrepancy between the predicted density distribution and its corresponding ground-truth counterpart is quantified via the L2-norm, with the exact form of the objective expressed in Eq. (7).

$$L_1(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left\|D\left(X_i;\theta\right) - D_g\right\|_2^2 \tag{7}$$

Specifically, θ denotes the ensemble of trainable weights within the architecture, N signifies the cardinality of the training image set, and L_1 quantifies the per-sample discrepancy - measured in L2 space - between the predicted density field and its corresponding ground-truth annotation. X_i is the input image, and D_g is the ground truth density map of the image X_i.$D\left(X_i;\theta\right)$ represents the estimated density map generated by the proposed model.

Since the final result of the model training is to ensure that the predicted number of people is close to the actual number, a loss between the actual number and the estimated number is also proposed. First, the estimated density map $D\left(X_i;\theta\right)$ is summed to obtain the estimated number C_e, and it is defined as shown in Eq. (8).

$$C_e = \sum_{i=0}^{W}\sum_{j=0}^{H} D\left(X_i;\theta\right) \tag{8}$$

Let W and H respectively index the spatial span and vertical extent of the predicted density lattice; then the residual loss L_2, which penalizes the absolute divergence between the ground-truth headcount and its network-derived approximation, is compactly expressed in Eq. (9).

$$L_2(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left\|C_e - C_g\right\|_2^2 \tag{9}$$

L_2 represents the loss between the estimated number and the actual number. C_g represents the actual number, while C_e indicates the estimated number. The final loss function is obtained by weighting and summing the above two loss functions, with the weight factor $\delta = 0.1$, as shown in Eq. (10).

$$L_{loss} = L_1(\theta) + \delta L_2(\theta) \tag{10}$$

3. Results and discussion

3.1. Experimental training process

The experiment selects Python 3.8 as the main programming language, and adopts PyTorch 1.7.0 as the core tool of the deep learning framework. The server operating system used is Ubuntu 8.04. Optimization proceeds via stochastic gradient descent, orchestrated on an NVIDIA GeForce RTX 3090 with CUDA 11.3 as the back-end accelerator; mini-batches of eight samples are ingested per iteration, while the learning rate is anchored at $1 \times \mathrm{e}^{-7}$.

To address the varying density regions in the image, this paper follows the method in Sang et al. [22] and employs an adaptive Gaussian kernel function to generate a true density map D_i^{GT}. The specific generation formula is defined as follows:

$$F(x) = \sum_{i=1}^{N}\delta\left(x - x_i\right) \times G_{\sigma_i}, \sigma_i = \beta\bar{d}^i \tag{11}$$

Here x represents an image with N head annotation points. x_i corresponds to one head annotation point in the image. δ is a pulse function. $\delta\left(x - x_i\right)$ represents the

position of each head in the image. G_{σ_i} is a Gaussian kernel function with variance σ_i. The average distance between head annotation point i and the k surrounding head points is expressed as $\bar{d^i} = \frac{1}{k}\sum_{j=1}^{k} d_j^i$.

3.2. Evaluation indicators

This study adopts two complementary gauges of predictive fidelity: Mean Absolute Error (MAE), which quantifies the expected $\ell 1$-norm deviation between inferred and ground-truth headcounts, thereby foregrounding overall accuracy; and Root Mean Square Error (RMSE), whose L2-norm formulation accentuates the volatility of mis-estimations, thus serving as a sentinel for model stability. Their formal characterizations are subsequently detailed.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|D_i - D_i^{GT}\right| \tag{12}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|D_i - D_i^{GT}\right|^2} \tag{13}$$

Here N represents the number of test samples. D_i and D_i^{GT} are the predicted density map and the true density map of the $i-$ th image respectively.

Experimental validation is carried out on Shanghai Tech, UCF_CC_50 and UCF_QNRF - three reference corpora whose rich annotations and diverse imaging conditions have rendered them de-facto benchmarks for the crowd-density estimation community [23], which can comprehensively verify the performance and robustness of the proposed method. The Shanghai Tech dataset is divided into two parts: Part A and Part B. Relative to Part A, Part B exhibits sparser crowds and markedly less visual clutter. As for UCF_CC_50, evaluation is framed by a five-fold cross-validation protocol: the corpus is randomly partitioned into five disjoint folds, each serving once as the test bed while the remaining four are leveraged for training, ensuring rigorous yet balanced assessment. The image resolutions in the UCF_QNRF dataset have a wide range of variations. This dataset presents diverse scenes, including different perspectives, various lighting conditions, and significant changes in population density.

3.3. Experimental results and discussion

This paper selects three advanced crowd counting methods (in including DHM [24], CCTV [25], DEMP [26]) and conducts comparative experiments with the proposed network on three benchmark datasets. The performance comparison of different counting networks is shown in Table 1. From the comparison results, it can be seen that the proposed method achieves the best MAE index on multiple datasets, and the RMSE index is also quite good, indicating that the proposed network can achieve more accurate counting results and demonstrates superior counting performance.

In Tech-A, the proposed method outperforms the other algorithms in Table 1 in terms of MAE. Compared with the suboptimal algorithm DEMP, the MAE has decreased by 1.18%. In this dataset, the RMSE of the Proposed method also reaches the optimal value with 96.7. This is mainly due to the imbalance of samples in the Tech-A dataset, which causes the model to be more inclined to learn the features of densely populated areas during training, while insufficient learning of features in sparse areas leads to less precise feature extraction. These factors led to the low counting accuracy of the model in sparse areas. In Tech-B, compared with the suboptimal algorithm DEMP, the MAE decreases by 25.64% and the RMSE decreases by 6.03%. In this dataset, it demonstrates the best accuracy and robustness in both MAE and RMSE.

Based on the experimental results on the UCF_CC_50 dataset and the performance comparison with other algorithms, it can be concluded that the proposed method demonstrates significant advantages among the current mainstream algorithms. Its MAE has reached a high accuracy level of 181.3, which proves the excellent accuracy of the new algorithm in this paper. However, the RMSE has not reached the optimal level and is inferior to the DEMP algorithm. This is mainly due to the small scale of the UCF_CC_50 dataset. The sample quantity and diversity are limited, causing the model to easily overfit to the limited training data during the training process and unable to access a sufficient variety of scenarios. As a result, the model has difficulty learning the general characteristics of the crowd, and thus has poor generalization ability when encountering new data. However, the MAE index of this network has achieved the optimal level on this dataset, and the RMSE is also at the leading level, indicating that the network performs well overall.

On the UCF_QNRF dataset, the proposed method achieves the best performance, with all indicators being the optimal. Compared to DEMP, the MAE decreases by 7.03 % and the RMSE decreases by 2.69 %. This clearly demonstrates that the proposed method has excellent predictive performance when dealing with high-resolution multi-scenario situations in the challenging UCF_QNRF dataset.

To present the prediction effect of the proposed model more intuitively, Fig. 4 shows the density plots generated by the proposed method on different datasets. Fig. 4(a), Fig. 4(b), and Fig. 4(c) display the original images of the corresponding samples, the true density plots, and the gen-

Table 1. Evaluation results with different methods on three datasets

Method	Tech-A		Tech-B		UCF_CC_50		UCF_QNRF	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
DHM	72.8	111.7	11.7	16.5	293.8	389.4	121.3	208.6
CCTV	62.5	101.6	9.6	15.1	201.4	309.5	105.5	182.8
DEMP	59.5	100.2	7.9	13.6	182.7	273.8	103.2	177.1
Proposed	58.8	96.7	5.8	14.9	181.3	237.7	93.9	170.3

erated estimated density plots respectively. The common characteristic of these samples is that the scale variation is quite obvious.

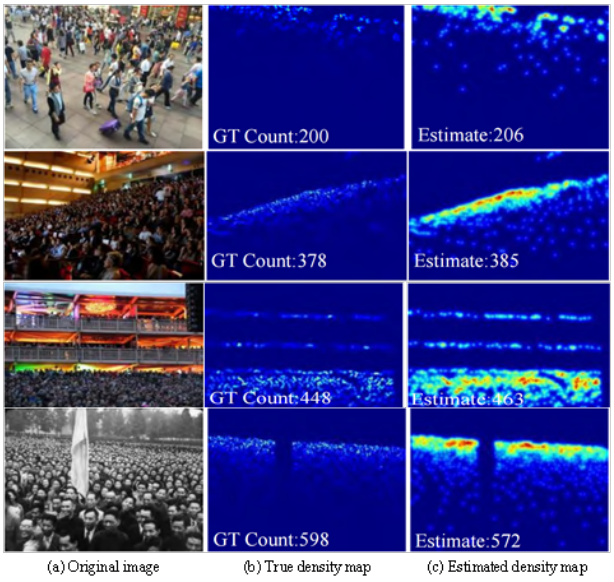

Fig. 4. Density map visualization results

As shown in Fig. 4, compared to the true density map, the estimated density map exhibits higher density values and more concentrated hotspots in densely populated areas. This difference may be due to the fact that the new model can better capture and learn the characteristics of high-density areas during the training process, making the density estimation in these areas more accurate. The estimated density map generated by the proposed model, when regressed to obtain the population count, has a deviation rate of the count results within 5 % compared to the population count in the true density map, which meets the requirements of practical applications. These visualization results indicate that this network can effectively handle the scene counting problem with large-scale variations in scale, and can more reliably estimate the distribution and density of the population.

4. Conclusions

The paper proposes a real-time tourist density estimation framework based on lightweight Swin-Transformer, aiming to resolve the contradiction between the accuracy of crowd counting and the edge deployment in scenic area monitoring. Based on the powerful self-attention mechanism of Swin-Transformer, it overcomes the problem that CNN fails to capture global context information. Through the designed MFPP module, it effectively integrates the feature maps of different encoding layers and extracts more semantic information at different scales. This paper conducts relevant experiments on three mainstream population datasets. Through comparative analysis, it is verified that this new model performs better in handling scale variation issues in the population counting task, thereby obtaining high-quality density estimation maps and enhancing the generalization ability of the model. The paper provides a low-cost density-aware solution that can be easily implemented at the edge for smart tourism, but it does not address extreme occlusion scenarios or cross-domain migration capabilities. Further optimization can be achieved by combining unsupervised domain adaptation in the future.

References

[1] H. Meng, X. Hong, C. Wang, M. Shang, and W. Zuo, (2024) *"Multi-modal crowd counting via a broker modality"*: 231–250. DOI: 10.1007/978-3-031-72904-1_14.

[2] L. Deng, Q. Zhou, S. Wang, J. M. Górriz, and Y. Zhang, (2024) *"Deep learning in crowd counting: A survey"* **CAAI Transactions on Intelligence Technology** ***9***(5): 1043–1077. DOI: 10.1049/cit2.12241.

[3] W. Wang, Q. Liu, and W. Wang, (2022) *"Pyramid-dilated deep convolutional neural network for crowd counting"* **Applied Intelligence** ***52***(2): 1825–1837. DOI: 10.1007/s10489-021-02537-6.

[4] M.-h. Oh, P. Olsen, and K. N. Ramamurthy. "Crowd counting with decomposed uncertainty". In: *Proceedings of the AAAI conference on artificial intelligence*. ***34***. 07. 2020, 11799–11806. DOI: 10.1609/aaai.v34i07.6852.

[5] S. Yin, L. Wang, T. Chen, H. Huang, J. Gao, J. Zhang, M. Liu, P. Li, and C. Xu, (2025) *"LKAFormer: A Lightweight Kolmogorov-Arnold Transformer Model for Image Semantic Segmentation"* **ACM Transactions on Intelligent Systems and Technology**: DOI: 10.1145/3759254.

[6] Y. Li, X. Zhang, and D. Chen. "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, 1091–1100. DOI: 10.1109/CVPR.2018.00120.

[7] W. Liu, M. Salzmann, and P. Fua. "Context-aware crowd counting". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, 5099–5108. DOI: 10.1109/CVPR.2019.00524.

[8] Y. Meng, H. Zhang, Y. Zhao, X. Yang, X. Qian, X. Huang, and Y. Zheng. "Spatial uncertainty-aware semi-supervised crowd counting". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, 15549–15559. DOI: 10.1109/ICCV48922.2021.01526.

[9] M. Wang, H. Cai, X.-F. Han, J. Zhou, and M. Gong, (2022) *"STNet: Scale tree network with multi-level auxiliator for crowd counting"* **IEEE Transactions on Multimedia *25***: 2074–2084. DOI: 10.1109/TMM.2022.3142398.

[10] S. Yin, H. Li, A. A. Laghari, L. Teng, T. R. Gadekallu, and A. Almadhor, (2024) *"FLSN-MVO: edge computing and privacy protection based on federated learning Siamese network with multi-verse optimization algorithm for industry 5.0"* **IEEE Open Journal of the Communications Society *6***: 3443–3458. DOI: 10.1109/OJCOMS.2024.3520562.

[11] H. Lin, Z. Ma, X. Hong, Q. Shangguan, and D. Meng. "Gramformer: Learning crowd counting via graph-modulated transformer". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. ***38***. 4. 2024, 3395–3403. DOI: 10.1609/aaai.v38i4.28126.

[12] D. B. Sam and R. V. Babu. "Top-down feedback for crowd counting convolutional neural network". In: *Proceedings of the AAAI conference on artificial intelligence*. ***32***. 1. 2018. DOI: 10.1609/aaai.v32i1.12290.

[13] J. Gao, Q. Wang, and X. Li, (2019) *"Pcc net: Perspective crowd counting via spatial convolutional network"* **IEEE Transactions on Circuits and Systems for Video Technology *30***(10): 3486–3498. DOI: 10.1109/TCSVT.2019.2919139.

[14] X. Ma, S. Du, and Y. Liu. "A lightweight neural network for crowd analysis of images with congested scenes". In: *2019 IEEE international conference on image processing (ICIP)*. IEEE. 2019, 979–983. DOI: 10.1109/ICIP.2019.8803062.

[15] L. Liang, H. Zhao, F. Zhou, M. Ma, F. Yao, and X. Ji, (2023) *"PDDNet: lightweight congested crowd counting via pyramid depth-wise dilated convolution"* **Applied Intelligence *53***(9): 10472–10484. DOI: 10.1007/s10489-022-03967-6.

[16] J. Yi, Z. Shen, F. Chen, Y. Zhao, S. Xiao, and W. Zhou, (2023) *"A lightweight multiscale feature fusion network for remote sensing object counting"* **IEEE Transactions on Geoscience and Remote Sensing *61***: 1–13. DOI: 10.1109/TGRS.2023.3238185.

[17] S. Yin, L. Wang, and L. Teng, (2024) *"Threshold segmentation based on information fusion for object shadow detection in remote sensing images"* **Computer Science and Information Systems *21***(4): 1221–1241. DOI: 10.2298/CSIS231230023Y.

[18] A. Kumar, S. P. Yadav, and A. Kumar, (2025) *"An improved feature extraction algorithm for robust Swin Transformer model in high-dimensional medical image analysis"* **Computers in biology and medicine *188***: 109822. DOI: 10.1016/j.compbiomed.2025.109822.

[19] Z. Ma, X. Wu, A. Chu, L. Huang, and Z. Wei, (2024) *"SwinFG: A fine-grained recognition scheme based on swin transformer"* **Expert Systems with Applications *244***: 123021. DOI: 10.1016/j.eswa.2023.123021.

[20] S. Yin, H. Li, A. A. Laghari, T. R. Gadekallu, G. A. Sampedro, and A. Almadhor, (2024) *"An anomaly detection model based on deep auto-encoder and capsule graph convolution via sparrow search algorithm in 6G Internet of Everything"* **IEEE Internet of Things Journal *11***(18): 29402–29411. DOI: 10.1109/JIOT.2024.3353337.

[21] M. A. Kizrak and B. Bolat, (2021) *"Crowd density estimation by using attention based capsule network and multi-column CNN"* **IEEE Access *9***: 75435–75445. DOI: 10.1109/ACCESS.2021.3081529.

[22] J. Sang, W. Wu, H. Luo, H. Xiang, and X. Xia, (2019) *"Improved Crowd Counting Method Based on Scale-Adaptive Convolutional Neural Network"* **IEEE Access *7***(99): 24411–24419. DOI: 10.1109/ACCESS.2019.2899939.

[23] Z. Chen, S. Zhang, X. Zheng, X. Zhao, and Y. Kong, (2023) *"Crowd counting based on multiscale spatial guided perception aggregation network"* **IEEE Transactions on Neural Networks and Learning Systems**: DOI: 10.1109/TNNLS.2023.3304348.

[24] J.-a. Cheng, Q. Li, A. Souri, X. Lei, C. Zhang, and M. Gao, (2025) *"Towards trustworthy crowd counting by distillation hierarchical mixture of experts for edge-based cluster computing"* **Cluster computing** *28*(7): 1–15. DOI: 10.1007/s10586-025-05226-y.

[25] K.-H. Kim, T.-K. Ahn, and S. Kim, (2025) *"Estimating Invisible Passenger Count Using CCTV Footage: An Approach Combining Object Detection Models and Machine Learning"* **IEEE Access**: DOI: 0.1109/ACCESS.2025.3597708.

[26] Z. Niu, H. Pi, G. Xiao, S. Yang, Z. Tang, and D. Liu, (2025) *"Low-Light Domain Enhancement and Multi-Domain Progressive Fusion for RGB-T Day-Night Crowd Counting"* **IEEE Internet of Things Journal**: DOI: 10.1109/JIOT.2025.3594227.