Density-Guided Adaptive Patch Learning for Robust Crowd Counting

Abdullah N Alhawsawi
Department of Information and Scientific Services, Custodian of the Two Holy Mosques
Institute for Hajj and Umrah Research, Umm Al-Qura University, Makkah, Saudi Arabia

***Abstract*—Accurate crowd counting in real-world scenes remains challenging due to severe occlusions, perspective distortion, and large intra-scene density variation. Recent deep learning based approaches typically address these challenges using patch-level learning, where images are divided into fixed grids or randomly cropped patches. These approaches then estimate the count in each patch. However, such fixed partitioning strategies often fail to align with the irregular spatial distribution of crowds. This leads to heterogeneous density patterns within patches where the models fail to produce the accurate count. In this study, we propose a simple, yet effective *Density-Guided Adaptive Patch Learning* framework for crowd counting. Instead of relying on fixed-size patches, we first obtain a coarse density estimation to capture the global density structure of a scene. Based on this estimate, the image is dynamically partitioned into density-homogeneous regions, where dense areas are represented using smaller patches and sparse regions using larger patches. Each adaptive patch is then processed independently for density estimation, and the resulting predictions are fused to produce the final crowd density map. The proposed framework is model-agnostic and can be seamlessly integrated with existing crowd counting networks without architectural modification. Extensive experiments on benchmark datasets demonstrate that the proposed adaptive partitioning consistently achieves lower Mean Absolute Error (MAE) in counting accuracy and localization compared to fixed patch-based baselines, particularly in scenes with strong density variation.**

***Keywords*—Computer vision; deep learning; crowd counting**

I. Introduction

In urban areas, large-scale gatherings such as political demonstrations, religious gatherings, and concerts are more prevalent [1]. During such events, security agencies should ensure safety of the people and that the activities of crowd are in control [2]. Automated crowd analysis systems involve counting of people [3]. The goal of the counting system is to determine the number of individuals in a scene and identify their locations [4]. Predicting the correct number of people in crowd can help avoid calamity and crowd disaster beforehand [5].

Despite extensive advancement, computer vision still faces the highly challenging problem of crowd counting [6]. This is due to the reason that the crowd scenes are extremely dense with people obstructing each other, and the point of view and scale vary significantly [7]. In such circumstances, the conventional surveillance and detection techniques cannot be effective. This implies we must apply automated techniques which estimate crowds directly from the image [8]. Recently, deep convolutional neural networks have advanced significantly by learning robust representations specific to specific scenes within crowds [9], [10], [11].

A. Fixed Patch-Based Methods

Most of the existing density-based crowd-counting deep learning methods employ patch-level learning strategies to address scale changes and enhance localization [12], [13], [14]. The methods divide an input image into patches either through fixed grids or random cropping. Then individual patches are individually tackled to determine density. The patch-based learning approach works well, provided the number of people in a patch is roughly equal. In real life, crowds do not necessarily behave in the same way [15]. Areas associated with close proximity can be filled up, whereas spaces that are nearby can remain vacant. Usually the patch boundaries are fixed and cut across these regions, mixing up the density patterns and rendering learned representations less valuable [16]. This degradation is caused by heterogeneous densities in one patch. This introduces regression ambiguity, in which the network is required to predict a single density of the spatially inconsistent distributions of crowds.

Recent works have tried to solve this problem by the use of multi-column architecture [17], [18], [19] or expert selection schemes to manage the scale variance. These techniques stabilize the system when the view varies These methods divide the input image into fixed spaces and then estimate the density in each patch. Thus, the primary problem of patch boundaries being unaligned with the underlying crowd structure remains. This is particularly apparent in the case when the crowd density varies significantly within the image.

A key concept in this study is that crowd density is a natural method of discovering valuable patch boundaries. Areas of similar density characteristics tend to possess equal patterns and scale in their images [20]. Conversely, the density differences are usually intense, implying that a change in the structure of the scene has occurred. On the basis of this observation, we suggest a density-guided strategy that alters patch size as well as shape in accordance with the distribution of the density of the crowd over space.

In this study, we discuss a crowd-counting framework known as Density-Guided Adaptive Patch Partitioning. This is done by first creating a rough density map to provide a sense of the overall density structure of the scene. The image is partitioned into adaptive patches approximately identical in density according to this conjecture. In order to maintain similar context smaller patches are used to depict areas that have much detail and larger patches are used to depict areas that have less detail. To estimate the density, the adaptive

patches are treated individually and provided as input to another convolutional neural network. The network estimates the count in each patch and count from all the patches are combined to produce the final crowd density map.

Experiments on standard benchmark datasets indicate that the adaptive partitioning strategy suggested is always more accurate during counting and localization tasks than the fixed patch-based approaches, particularly in challenging scenes with uneven distributions of crowds.

This work offers the following contributions:

- In this work, we identify the challenges associated with fixed patch-based counting models.
- We address the issues associated with the patch-based method by proposing an adaptive patch counting strategy that utilizes the coarse density information to adjust the size of the patch.
- We perform experiments on a challenging variety of public datasets. From the experiment results, we demonstrate the effectiveness of the proposed method.

II. Related Work

The computer vision community has put a lot of effort in the study of crowd counting, starting with detection-based methods of counting, to regression-based [21], and deep learning-based counting methods [7]. This section will review the most appropriate works and place our approach in the existing literature.

The initial crowd counting methods used individual pedestrians and detected them or tracked them with hand-coded features. Head or body-detection-based methods used Haar wavelets, HOG features, and part-based models to locate faces in images or videos. Tracking-based approaches tried to estimate the number of pedestrians by aggregating trajectories of motion that are extracted using video images [22], [23], [24]. Although these methods are effective in low-density crowds, they have serious performance problems in dense crowds where there is occlusion and clutter.

In order to address the drawbacks of detection and tracking based algorithms, regression-based models were suggested to directly estimate the number of crowds using global or local image features. Initial models used a linear and kernel regression model based on feature of texture, edges, and interest points [25], [26]. A study by Idris et al. [27] was able to use a mixture of handcrafted features, such as HOG-based head detectors and Fourier analysis, to deal with very dense crowds. Despite the strength of these techniques, handcrafted details usually had limited representational power to process complex crowd scenes.

A. Density Map Refinement

Density map estimation also proposed a pixel-based formulation of counting people, which allows to estimate the counts and to localize them spatially at the same time [28]. Later literature continued the concept of random forests and structured regression to enhance efficiency and strength to a better level. The capability to deal with occlusion and the ability to give fine-grained information about the distribution of crowds has made density-based approaches a prevailing paradigm.

Conventional neural networks (CNNs) have become commonly used in counting crowds with the effectiveness of deep learning. Zhang et al. [29] presented a CNN-based cross-scene crowd counting framework based on density and count targets and data-driven fine-tuning. Boominathan et al. [30] presented CrowdNet that is a combination of deep and shallow fully convolutional networks to extract semantic and low-level features across large scale differences.

B. Multi-Scale Approaches

Multi-column CNNs were subsequently developed to explicitly deal with scale variations with multiple branches that have varying receptive fields [18]. Switch-CNN [31] also expanded the concept by dynamically forwarding image patches to image-specific regressors depending on crowd density. Although these methods are highly effective in enhancing performance they are usually based on grid partitioning or a fixed patch size which does not always match the irregular structure of the density of real-world crowds.

Cross-scene crowd counting helps resolve the problem of extrapolating models to unknown settings. A data-driven approach to adaptation suggested by Zhang et al. [29] involves retrieving similar training samples to be fine-tuned to the specific scene. Perspective-aware techniques make use of camera geometry to normalize scale changes, commonly with manually developed perspective maps [32], [29]. Nevertheless, getting the correct perspective information is not always possible and is often a labor-intensive job.

In spite of this breakthrough, the majority of patch-based crowd counting methods use a fixed spatial partitioning approach. These methods combine non-uniform density patches within a patch restricting their utility in images where the density varies strongly across the image. Conversely, our work is about the density structure adaptation of patch boundaries, which is a straightforward and supplementary enhancement of current crowd counting systems.

III. Proposed Method

An overview of the proposed framework is illustrated in Fig. 1. The framework consists of three main modules: 1) coarse density estimation, 2) adaptive patch partitioning based on density homogeneity, and 3) patch-level density refinement and fusion.

A. Coarse Density Estimation

This module estimates the coarse density within each cell. This information is later on used by the other modules to generate adaptive patches and produce refine density map.

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, obtain a coarse density map $D_c \in \mathbb{R}^{H\times W}$ using a lightweight convolutional neural network which is trained using standard density regression loss function and formulated, as in Eq. (1):

$$D_c = f_c(I) \tag{1}$$

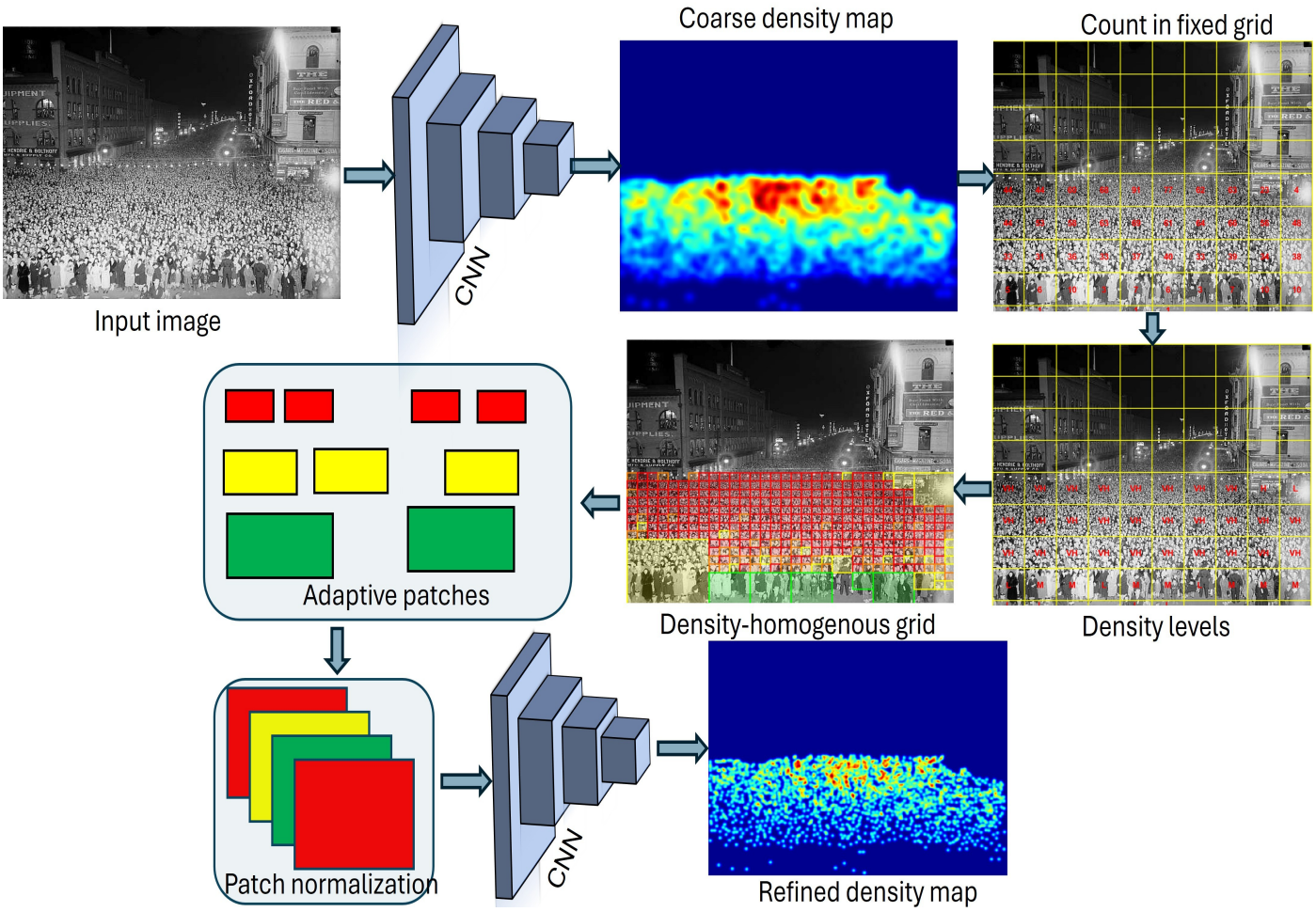

Fig. 1. Overall pipeline of the proposed crowd counting framework.

The primary objective of this module is to accurately estimate the number of people in the image. This module is used to capture the spatial distribution of pedestrians in the scene.

B. Density-Guided Adaptive Patch Partitioning

After obtaining the coarse density map, D_c, , we partition the image into a sequence of adaptive patches, $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$, where P_i and P_j are the neighboring patches and have approximately the same density levels. Algorithm 1 presents the density-guided adaptive patch partitioning.

For obtaining the sequence of adjacent patches, We first divide the coarse density map D_c into non-overlapping base cells of size $s \times s$. For each cell C_i, we compute the the average density value, as in Eq. (2):

$$\bar{d}_i = \frac{1}{|C_i|} \sum_{x \in C_i} D_c(x). \tag{2}$$

Based on the average density value, $\bar{d}_i$, each cell is assigned one of K the density levels using predefined thresholds. In this work, we use four density levels: low, medium, high and very high; therefore, we fix the value of K to 4 levels, as provided in Eq. (3):

$$l_i = \begin{cases} 1, & \bar{d}_i < \tau_1, \\ 2, & \tau_1 \leq \bar{d}_i < \tau_2, \\ \vdots & \\ K, & \bar{d}_i \geq \tau_{K-1}. \end{cases} \tag{3}$$

After assigning a density level to each cell, we then combine the adjacent cells that share the same density levels into a single large cell. In this way, dense regions produce small patches, while sparse regions will produce large patches. This strategy helps us to keep the contextual information intact, which is very crucial for precise crowd counting.

C. Patch-Level Density Refinement

After obtaining the adaptive patches, we then resize the size of patch to a fixed size and provided as input to the density estimtion network.

Let adaptive patch $P_j \in \mathcal{P}$ is process through a density estimation network $f_r(\cdot)$, as formulated in Eq. (4):

Algorithm 1 Density-Guided Adaptive Patch Partitioning

Input: Input image I
Output: Final density map D_f, total count C
Compute coarse density map $D_c = f_c(I)$
Divide D_c into base cells $\{C_i\}$ of size $s \times s$
foreach *cell* C_i **do**
  Compute average density $\bar{d}_i$ Assign density level l_i based on thresholds $\{\tau_k\}$
Merge adjacent cells with identical l_i to form adaptive patches $\mathcal{P}$
foreach *adaptive patch* $P_j \in \mathcal{P}$ **do**
  Resize P_j to fixed input size Predict refined density map $D_j = f_r(P_j)$
Project all D_j back to original image coordinates
Fuse patch density maps to obtain D_f
Compute total count $C = \sum_x D_f(x)$
return D_f, C

$$D_j = f_r(P_j), \tag{4}$$

where, D_j denotes the refined density map for patch P_j.

We can use any Crowd density estimator $f_r(\cdot)$, however, we trained the network using the standard L_2 loss between predicted and ground-truth density maps, as formulated in Eq. (5):

$$\mathcal{L}_{\text{density}} = \frac{1}{N}\sum_{j=1}^{N} \|D_j - D_j^{GT}\|_2^2. \tag{5}$$

D. Density Fusion and Crowd Count Estimation

After computing the crowd density in each patch, we then compute the final density map D_f by projecting each refined patch density map D_j back to its original spatial location by using Eq. (6):

$$D_f(x) = \sum_{j=1}^{N} \mathbb{I}(x \in P_j) \cdot D_j(x), \tag{6}$$

where, $\mathbb{I}(\cdot)$ is an indicator function.

We then compute the final density by aggregating all crowd counts from all patches, as provided in Eq. (7):

$$C = \sum_x D_f(x). \tag{7}$$

IV. Experiments

To evaluate the performance of proposed framework, we use three publicly available dataset, namely ShanghaiTech, UCF_CC_50, and WorldExpo'10. These datasets are widely used to evaluate the performance of crowd counting methods. We provide the details of each dataset as follows:

A. Datasets

1) ShanghaiTech dataset: ShanghaiTech dataset is a popular dataset used for evaluating crowd counting and density estimation methods. This dataset is first collected by Zhang et al. [18]. The dataset includes diverse real-world scenes of high crowd densities. The dataset is categorized into two parts, Part A and Part B depending on the complexity of the scenes.

ShanghaiTech Part A contains 482 images that are collected from diverse scenes. These images are captured from different perspectives and with different spatial resolutions. The images also contain scenes with different counts of people, ranging from 33 to 3,139 people per image, with the average count of people in an image 501.4. Part A has 241,677 annotations, which is enough to test models under extreme conditions, such as occlusion, scale changes, and perspective distortions.

ShanghaiTech Part B, on the other hand, consists of 716 images that were taken on the streets of Shanghai, and all the images have same resolution of 768 x 1024 pixels. This part of the dataset comparatively contains low-density crowded scenes, where the count of people per image ranges from 9 to 578 people with an average of 123.6 people. Part B contains 88,488 annotations. This dataset is more suitable for assessing the performance of the models in realistic applications in surveillance-like tasks.

2) UCF_CC_50 dataset: The UCF_CC_50 is a conventional standard that is widely employed to test crowd counting and density estimation models. The data is first curated by Idrees et al. [27] The data set consists of diverse images of various scenes and resolutions. The number of individuals varies significantly in the images with an average number of people per image is 1,279.5 with a range of between 94 and 4,543 individuals. There are 63,974 annotations in the dataset. Since the dataset is complex in nature, this dataset is usually employed to test the effectiveness of the models in different challenging scenes.

3) WorldExpo'10 dataset: WorldExpo'10 is a large dataset for crowd counting employed to test the crowd counting and density estimation methods in various real-world scenarios, with a relatively huge volume of data. The dataset consists of 3,980 images with a constant resolution of 576 × 720. The range of the crowd counts is between 1 individual frame and 253 per image, with a mean of 50.2 people per frame. The dataset provides 199,923 annotated individuals. Due to the large number of samples with moderate crowd density and diversity, WorldExpo.10 tends to be used to test the generalization capability and reliability of a model across different scenes.

To evaluate the performance of proposed framework, we use Mean Absolute Error (MAE) and Mean Squared Error (MSE) as evaluation metrics formulated in Eq. (8) and Eq. (9). These metrics are widely used to quantify the performance of the models.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |C_i - C_i^{GT}|, \tag{8}$$

$$\text{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i - C_i^{GT})^2}, \tag{9}$$

where, C_i and C_i^{GT} denote the estimated and ground-truth counts, respectively.

B. Implementation Details

We implemented the framework in Python with the PyTorch library. We trained the coarse density estimation network independently using standard density maps. These density maps are generated by first projecting the point annotations onto the zero-valued image. The size of zero-valued image is same as the size of the original image. We then employ a 2D Gaussian kernel to generate a density map. We utilize these density map to train the coarse density estimation network using L_2 loss.

Adaptive patch partitioning thresholds are set according to levels of density of crowds by following the convention in [33]. Particularly, four levels of density are taken into account. These levels include very low, low, medium, and high. The areas (cells) with a count of less than 8 persons is classified as very low, areas with 9 to 16 persons are classified as low, medium-density areas have 17 to 21 persons, and high-density areas have more than 21 persons. After classifying each cell into density levels, we then merged the neighboring cells that have the same density level to form adaptive patches. Due to this strategy, dense regions, which are inherently compact regions, will produce small patches, while very low and low-density regions, which are generally smooth and uniform, are merged into larger patches.

We trained all the models using stochastic gradient descent with a learning rate of 10^{-6} and momentum of 0.9.

C. Results on ShanghaiTech

We report the performance comparison of the proposed method and other baseline methods in Table I on the ShanghaiTech dataset. We reported the results on both parts of the dataset in terms of MAE and MSE.

TABLE I. PERFORMANCE COMPARISON ON THE SHANGHAITECH DATASET.

	Part A		Part B	
Method	MAE	MSE	MAE	MSE
MCNN	110.2	173.2	26.4	41.3
Switch-CNN	90.4	135.0	21.6	33.4
CrowdNet	88.6	138.7	20.1	31.8
Proposed	**86.3**	**131.9**	**19.2**	**29.6**

From the table, it is obvious that proposed framework beats all other reference methods by achieving low MAE and MSE values. Further analysis revealed that models perform well on Part B than Part A. This implies that the models perform well on low-density areas, which also supports our motivation that fixed grid partitioning does not work well in scenes where the crowd distribution is not uniform. On the other hand, the proposed adaptive partitioning strategy better captures the local density structure, which leads to improved performance.

D. Results on UCF_CC_50

A comparative analysis of various methods on the challenging UCF_CC_50 dataset is provided in Table II. In general, the findings in Table II indicate that the proposed method has achieved good performance compared to all other approaches, producing the smallest values of MAE and MSE.

TABLE II. PERFORMANCE COMPARISON ON THE UCF_CC_50 DATASET

Method	MAE	MSE
Idrees et al.	419.5	487.1
MCNN	377.6	509.1
Switch-CNN	318.1	439.2
CrowdNet	452.5	–
Proposed	**305.7**	**421.4**

This implies that the proposed framework does not only enhance counting accuracy but also generates more predictive stability with high accuracy. Other previous approaches like Idrees et al. and MCNN are more error-prone, indicating the challenge of dealing with large-scale density changes in highly crowded scenes. Switch-CNN demonstrates a significant improvement because of the density-sensitive switching mechanisms; nevertheless, the proposed method remains in the lead. CrowdNet also reports a relatively high MAE and fails to give a comprehensive evaluation of its strength. The improved performance of the proposed method underscores its better ability to learn complex distributions of crowds in densely populated and sparsely populated regions that is of great concern in UCF_CC_50 where there is a lack of sufficient training samples, as well as a large number of crowds.

E. Results on WorldExpo'10

1) WorldExpo'10 dataset: Results on the WorldExpo'10 dataset are summarized in Table III. The proposed method achieves the lowest average MAE across all five test scenes, indicating improved cross-scene generalization. Unlike approaches that rely on scene-specific fine-tuning or perspective maps, the proposed framework adapts patch sizes dynamically based on density structure, without requiring additional annotations. Notably, consistent improvements are observed across scenes with varying crowd densities and camera viewpoints. These results suggest that density-guided adaptive partitioning effectively mitigates domain shifts between scenes by preventing inappropriate aggregation of sparse and dense regions within fixed patches.

F. Ablation Studies

1) Effect of adaptive patch partitioning: To evaluate the impact and importance of the adaptive patch partitioning method and to compare it with fixed grid and random patch strategies, we report the result in Table IV. From Table IV, it is demonstrated that adaptive patch partitioning achieves better performance compare to baseline strategies. This is due to the reason that the adaptive patch partitioning strategy adopts the density-guided method rather than fixed and random patch sizes. Fixed grid partitioning places artificial boundaries often cutting across dense areas whereas random cropping is spatially inconsistent.

TABLE III. MAE COMPARISON ON THE WORLDEXPO'10 DATASET

Method	S1	S2	S3	S4	S5	Avg.
MCNN	3.4	20.6	12.9	13.0	8.1	11.6
Switch-CNN	4.2	14.9	14.2	18.7	4.3	11.2
Cross-Scene CNN	9.8	14.1	14.3	22.2	3.7	12.9
Proposed	**3.1**	**13.7**	**11.8**	**11.9**	**4.0**	**8.9**

TABLE IV. EFFECT OF ADAPTIVE PATCH PARTITIONING ON SHANGHAITECH PART A.

Method Variant	MAE
Fixed grid patches	90.8
Random patches	92.1
Adaptive partitioning (ours)	**86.3**

TABLE V. EFFECT OF COARSE DENSITY ESTIMATION QUALITY ON UCF_CC_50.

Coarse Density Source	MAE
Uniform partitioning	331.4
Weak CNN (ours)	**305.7**
Heavy CNN backbone	304.9

TABLE VI. SENSITIVITY TO BASE CELL SIZE s ON SHANGHAITECH PART A.

Base Cell Size s	MAE
16×16	87.5
32×32	**86.3**
64×64	88.1

2) Effect of coarse density estimation: We also measure the influence of different coarse density methods on the overall performance in Table V. Table V evaluates the impact of the quality of coarse density estimation on the performance on UCF_CC_50 dataset. Uniform partitioning gives maximum error which shows that neglecting density information gives poor performance. The addition of a weak CNN to estimate the coarse density can be used to achieve much lower MAE, and it demonstrates that a simple density-aware model can be successfully used to guide the partitioning process. A slight improvement is made by replacing the weak CNN with a heavier backbone. This implies that coarse density estimation accuracy does not matter much at this point and a light and efficient model will suffice with the adaptive partitioning.

3) Patch size sensitivity: The impact of different base cell sizes s on performance over the ShanghaiTech Part A dataset are reported in Table VI. As the results indicate, the moderate cell size of 32×32 obtains the lowest MAE (86.3) which performs better compared to the smaller (16×16, MAE = 87.5) and larger (64×64, MAE = 88.1) configurations. The difference between performance with optimal settings and the values of all tested cases is in a relatively small range of 1.8 MAE, which means that the variance is small in relation to this parameter. Such small change indicates that the proposed technique is not very sensitive and statistically stable to the selection of the base cell size. The smaller cells can lead to noisy partitions, and the bigger cells can limit adaptive flexibility.

G. Discussion

From the above experiment results on three datasets, including, ShanghaiTech, UCF_CC_50 and WorldExpo 10, it is demonstrated that the proposed density-guided adaptive patch partitioning scheme achieved consistent performance in all these datasets. Although these datasets are diverse with varying densities, the complexity of the scene, and resolution, the proposed approach yields consistent and significant gains over fixed patch-based ones.

In the ShanghaiTech dataset, the proposed approach exhibits more gains on Part A than on Part B. In Part A, crowd density varies significantly within individual images. Fixed grid partitioning frequently divided the dense regions or combined tthe areas with distinct density characteristics, which degrades the performance. On the other hand, the proposed method is more effective at preserving the local structure of the crowd since it aligns patch boundaries with regions of homogeneity in density, which leads to better density estimates. The proposed framework also achieve good performance on Part B which contains varying sparse density scenarios. This demonstrate that the proposed framework work well in both high and sparse density crowded scenes.

On the ShanghaiTech dataset, which contains both highly congested and relatively sparse scenes, the proposed method shows stronger gains on Part A than on Part B. This behavior aligns well with the design motivation of the framework. Part A images exhibit significant intra-image density variation, where fixed grid partitioning frequently fragments dense regions or mixes heterogeneous density patterns. By aligning patch boundaries with density-homogeneous regions, the proposed approach better preserves local crowd structure, resulting in improved density estimation. In contrast, Part B scenes are more uniformly distributed, and while the performance gains are smaller, the proposed method maintains competitive accuracy without degradation, indicating good robustness in simpler scenarios.

The UCF_CC_50 dataset is even more of a challenge because of its restricted size and a large variety of crowd numbers. The high performance on this dataset demonstrates that the proposed adaptive partitioning achieves good performance compared to the fixed and random patch techniques. The capability to produce smaller patches in high-density areas enables the refinement network to concentrate on fine-grained crowd patterns which is essential in situations of extreme density and extreme occlusion. The experiments on this dataset indicates that adaptive partitioning is especially useful when the training data is very sparse and has large density variations.

On the WorldExpo-10 dataset, which focuses on cross-scene generalization, the proposed method has the lowest average error on all the test scenes. As compared to methods that make use of scene-specific fine-tuning or perspective maps, the suggested architecture dynamically adjusts patch sizes based on density cues only. This allows managing domain shifts in various camera angles and crowd configurations without any extra annotations or scene-specific calibration. The trend of steady enhancement in all the test scenes suggests that density-guided partitioning offers a strong cross-scene crowd counting mechanism.

The visualization of different stages of proposed framework on different sample images is illustrated in Fig. 2.

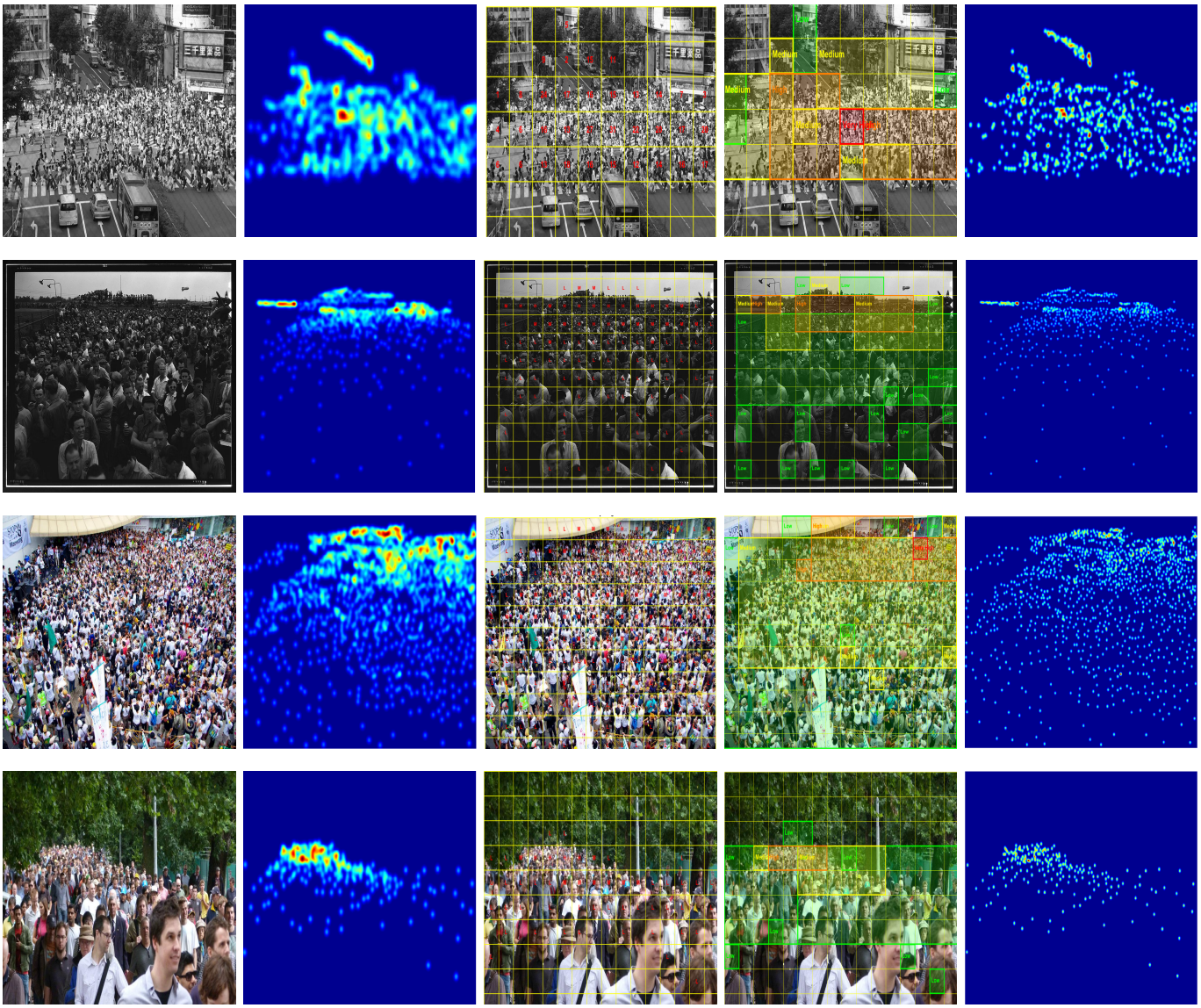

Fig. 2. Visualization of different stages of the proposed framework. The first two rows are the sample frames from the UCF_CC_50 dataset, while the last two rows represent the sample image from the ShanghaiTech dataset. The first column represents the input sample frames. Second column represents the coarse density. Third column represents the density level cells while the fourth and fifth column represent the adaptive patches and refined density map, respectively.

V. Conclusion

In this study, we presented a density-guided adaptive patch partitioning framework for crowd counting. The proposed approach addresses a key limitation of existing patch-based methods, which rely on fixed spatial partitioning and often fail to align with the irregular density distribution of real-world crowds. By leveraging a coarse density estimation to guide adaptive patch generation, the proposed framework produces density-homogeneous regions that are better suited for patch-level density refinement.

The proposed method is simple, lightweight, and model-agnostic, and can be seamlessly integrated with existing crowd counting networks without architectural modification or additional supervision. Extensive experiments on three widely used benchmarks demonstrate consistent improvements over fixed patch-based baselines, particularly in scenes with strong intra-image density variation and complex crowd layouts.

Future work may explore extending the adaptive partitioning strategy to temporal crowd analysis in videos and integrating it with more advanced density estimation models while preserving the simplicity and efficiency of the proposed framework. Furthermore, we will explore to employ adaptive partitioning strategy to existing modular architectures. Also, the density-guided adaptive partitioning principle can potentially be used in other dense prediction and spatial analysis problems besides counting crowds, and future studies can explore its application to similar dense prediction and spatial analysis problems like object localization, congestion detection, traffic flow estimation, and count cells in biomedical images. Such extensions may also establish the generalizability and extended applicability of the concept of adaptive spatial decomposition

in addition to maintaining efficiency and modularity, as seen in the present study.

REFERENCES

[1] E. Felemban, S. D. Khan, A. Naseer, F. Ur Rehman, and S. Basalamah, "Deep trajectory classification model for congestion detection in human crowds." *Computers, Materials & Continua*, vol. 68, no. 1, 2021.

[2] S. Hall, W. E. Cooper, L. Marciani, and J. M. McGee, *Security management for sports and special events: An interagency approach to creating safe facilities*. Human kinetics, 2011.

[3] A. N. Alhawsawi, S. D. Khan, and F. U. Rehman, "Enhanced yolov8-based model with context enrichment module for crowd counting in complex drone imagery," *Remote Sensing*, vol. 16, no. 22, p. 4175, 2024.

[4] C. Mccarthy, H. Ghaderi, F. Marti, P. Jayaraman, and H. Dia, "Video-based automatic people counting for public transport: On-bus versus off-bus deployment," *Computers in Industry*, vol. 164, p. 104195, 2025.

[5] B. Anzengruber, D. Pianini, J. Nieminen, and A. Ferscha, "Predicting social density in mass events to prevent crowd disasters," in *International Conference on Social Informatics*. Springer, 2013, pp. 206–215.

[6] A. N. Alhawsawi, S. D. Khan, and F. Ur Rehman, "Crowd counting in diverse environments using a deep routing mechanism informed by crowd density levels," *Information*, vol. 15, no. 5, p. 275, 2024.

[7] M. Hassan, F. Hussain, S. D. Khan, M. Ullah, M. Yamin, and H. Ullah, "Crowd counting using deep learning based head detection," *Electronic Imaging*, vol. 35, pp. 293–1, 2023.

[8] T. Jin, X.-W. Ye, W.-M. Que, and M.-Y. Wang, "Automatic detection, localization and quantification of structural cracks combining computer vision and crowd sensing technologies," *Construction and Building Materials*, vol. 476, p. 141150, 2025.

[9] Y. Wu, L. Qiu, J. Wang, and S. Feng, "The use of convolutional neural networks for abnormal behavior recognition in crowd scenes," *Information Processing & Management*, vol. 62, no. 1, p. 103880, 2025.

[10] W. Mansouri, M. A. Alohali, H. Alqahtani, N. Alruwais, M. Alshammeri, and A. Mahmud, "Deep convolutional neural network-based enhanced crowd density monitoring for intelligent urban planning on smart cities," *Scientific Reports*, vol. 15, no. 1, p. 5759, 2025.

[11] K. Zhao, C. He, S. Peng, and T. Lu, "A hybrid multi-scale transformer-cnn unet for crowd counting," *Sensors*, vol. 26, no. 1, p. 333, 2026.

[12] Y. Zhou, J. Yang, H. Li, T. Cao, and S.-Y. Kung, "Adversarial learning for multiscale crowd counting under complex scenes," *IEEE transactions on cybernetics*, vol. 51, no. 11, pp. 5423–5432, 2020.

[13] S. D. Khan, Y. Salih, B. Zafar, and A. Noorwali, "A deep-fusion network for crowd counting in high-density crowded scenes," *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, p. 168, 2021.

[14] L. Dong, H. Zhang, Y. Ji, and Y. Ding, "Crowd counting by using multi-level density-based spatial information: A multi-scale cnn framework," *Information Sciences*, vol. 528, pp. 79–91, 2020.

[15] A. J. ALZAHRANI and S. D. Khan, "Characterization of different crowd behaviors using novel deep learningframework," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 29, no. 1, pp. 169–185, 2021.

[16] S. A. M. Saleh, S. A. Suandi, and H. Ibrahim, "Recent survey on crowd density estimation and counting for visual surveillance," *Engineering Applications of Artificial Intelligence*, vol. 41, pp. 103–114, 2015.

[17] M. Saqib, S. D. Khan, N. Sharma, and M. Blumenstein, "Person head detection in multiple scales using deep convolutional neural networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–7.

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

[19] Y. Yang, G. Li, D. Du, Q. Huang, and N. Sebe, "Embedding perspective analysis into multi-column convolutional neural network for crowd counting," *IEEE Transactions on Image Processing*, vol. 30, pp. 1395–1407, 2020.

[20] S. Basalamah, S. D. Khan, and H. Ullah, "Scale driven convolutional neural network model for people counting and localization in crowd scenes," *IEEE Access*, vol. 7, pp. 71 576–71 584, 2019.

[21] D. Ryan, S. Denman, S. Sridharan, and C. Fookes, "An evaluation of crowd counting methods, features and regression models," *Computer Vision and Image Understanding*, vol. 130, pp. 1–17, 2015.

[22] C. Yan-yan, C. Ning, Z. Yu-yang, W. Ke-han, and Z. Wei-wei, "Pedestrian detection and tracking for counting applications in metro station," *Discrete dynamics in nature and society*, vol. 2014, no. 1, p. 712041, 2014.

[23] O. Sidla, Y. Lypetskyy, N. Brandle, and S. Seer, "Pedestrian detection and tracking for counting applications in crowded situations," in *2006 IEEE International Conference on Video and Signal Based Surveillance*. IEEE, 2006, pp. 70–70.

[24] Z. Sun, J. Chen, L. Chao, W. Ruan, and M. Mukherjee, "A survey of multiple pedestrian tracking based on tracking-by-detection framework," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1819–1833, 2020.

[25] A. B. Chan and N. Vasconcelos, "Bayesian poisson regression for crowd counting," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 545–551.

[26] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, "Density map regression guided detection network for rgb-d crowd counting and localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1821–1830.

[27] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2547–2554.

[28] V. Lempitsky and A. Zisserman, "Learning to count objects in images," *Advances in neural information processing systems*, vol. 23, 2010.

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

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

[31] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2017, pp. 4031–4039.

[32] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–7.

[33] M. Saqib, S. D. Khan, and M. Blumenstein, "Texture-based feature mining for crowd density estimation: A study," in *2016 International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, 2016, pp. 1–6.