Scale-Aware Crowd Counting Network With Annotation Error Modeling

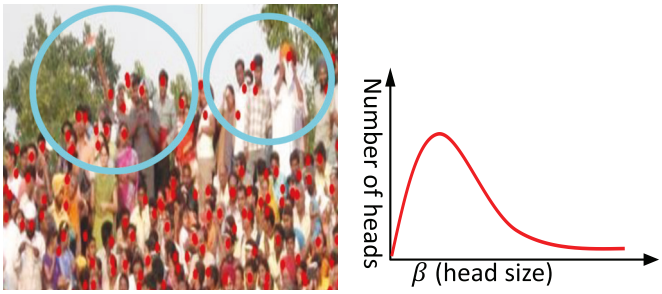
Yi-Kuan Hsieh, Jun-Wei Hsieh, *Senior Member, IEEE*, Xin Li, *Fellow, IEEE*, Yu-Ming Zhang, Yu-Chee Tseng, *Fellow, IEEE*, and Ming-Ching Chang, *Senior Member, IEEE*

***Abstract*—Traditional crowd-counting networks suffer from information loss when feature maps are reduced by pooling layers, leading to inaccuracies in counting crowds at a distance. Existing methods often assume correct annotations during training, disregarding the impact of noisy annotations, especially in crowded scenes. Furthermore, using a fixed Gaussian density model does not account for the varying pixel distribution of the camera distance. To overcome these challenges, we propose a Scale-Aware Crowd Counting Network (SACC-Net) that introduces a scale-aware loss function with error-compensation capabilities of noisy annotations. For the first time, we simultaneously model labeling errors (mean) and scale variations (variance) by spatially varying Gaussian distributions to produce fine-grained density maps for crowd counting. Furthermore, the proposed scale-aware Gaussian density model can be dynamically approximated with a low-rank approximation, leading to improved convergence efficiency with comparable accuracy. To create a smoother scale-aware feature space, this paper proposes a novel Synthetic Fusion Module (SFM) and an Intra-block Fusion Module (IFM) to generate fine-grained heat maps for better crowd counting. The lightweight version of our model, named SACC-LW, enhances the computational efficiency while retaining accuracy. The superiority and generalization properties of scale-aware loss function are extensively evaluated for different backbone architectures and performance metrics on six public datasets: UCF-QNRF, UCF CC 50, NWPU, ShanghaiTech A, ShanghaiTech B, and JHU. Experimental results also demonstrate that SACC-Net outperforms all state-of-the-art methods, validating its effectiveness in achieving superior crowd-counting accuracy. The source code is available at https://github.com/Naughty725.**

***Index Terms*—Annotation error modeling, scale-aware crowd counting network (SACC-Net), density map generation, low-rank approximation.**

I. Introduction

CROWD counting is an increasingly important technique in computer vision with applications in public safety and crowd behavior analysis [1], [2]. Over the years, many CNN-based crowd-counting methods have been developed to predict crowd density maps from a given image [3], [4], [5], [6], [7], [8], [9], [10], [11]. The number of people in the image is then calculated by adding up the predicted values on the density map. In past methods, the image was passed directly through a backbone network, where the last layer was used to predict the density map. Most existing methods did not adequately account for the scale problem when viewing people in the 3D space: people at the far end tend to look smaller than those close to the camera. Existing counting methods have difficulty generating fine-grained density maps to accurately count people at the far end of an input image after it passes through the pooling layer.



(a) (b)

Fig. 1. Modeling uncertainty (including annotation errors) for crowd counting. (a) Inaccurate annotations lead to biased mean (red dots deviate from the center of human faces). (b) Different camera distances lead to a positively skewed distribution of head sizes β, characterizing the change in variance.

Moreover, many existing methods require precise annotations from which a density map can be constructed using the L2-norm [3], [12], [13] or Bayesian Loss (BL) [6]. Unfortunately, even for human annotators, labeling errors are inevitable because ground-truth labeling might vary from subject to subject. As illustrated in Fig. 1, accurately pinpointing the center of each individual's head in an image is not a trivial task, and the process can pose technical challenges, particularly for people who appear small at a distance. As the crowd size increases, the distance from a person to the camera is not constant: individuals far away might only occupy a few pixels, or even less than a pixel in the image, rendering annotation more challenging and unreliable. Therefore, treating all pixels equally in Bayesian Loss [6] will likely affect the accuracy of crowd counting. How to handle scale variations and annotation errors in crowd counting remains an open problem [14], to the best of our knowledge.

The main motivation for this work is to improve crowd-counting accuracy by addressing the scaling truncation problem (caused by the pooling operations) as well as the

Received 23 April 2024; revised 4 November 2024; accepted 5 February 2025. Date of publication 24 April 2025; date of current version 9 May 2025. This work was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST 113-2221-E-A49-126-MY3, Grant 112-2221-E-A49-098-MY3, and Grant 113-2622-E-A49-008. The associate editor coordinating the review of this article and approving it for publication was Dr. Ioannis Pratikakis. *(Corresponding author: Jun-Wei Hsieh.)*

Yi-Kuan Hsieh, Jun-Wei Hsieh, and Yu-Chee Tseng are with the College of Artificial Intelligence, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: jwhsieh@nycu.edu.tw).

Xin Li and Ming-Ching Chang are with the Computer Science Department, University at Albany, SUNY, Albany, NY 12222 USA.

Yu-Ming Zhang is with the Computer Science Information Engineering Department, National Central University, Taoyuan 32001, Taiwan.

Digital Object Identifier 10.1109/TIP.2025.3555116

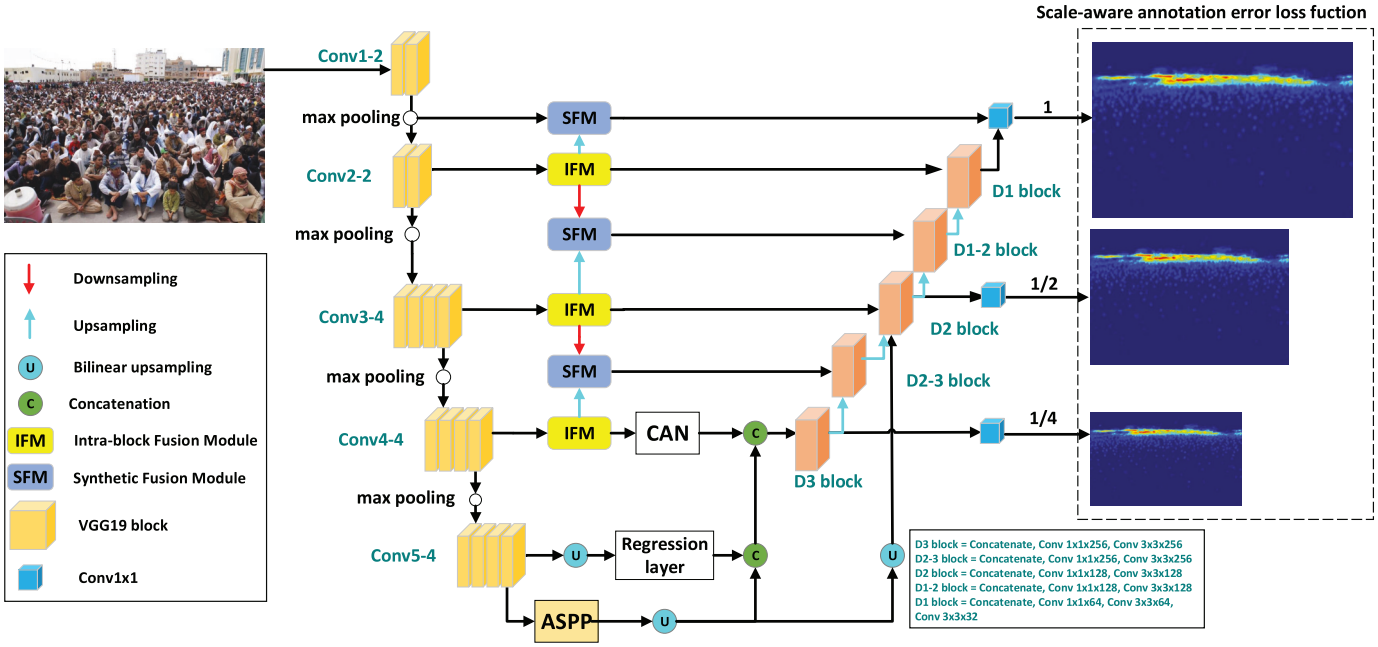


Fig. 2. Details of the proposed **Scale-Aware Crowd Counting Network (SACC-Net)** architecture for scale-aware crowd counting. The VGG-19 backbone, together with two newly designed feature fusion modules, namely the **Synthetic Fusion Module (SFM)** and **Intra-Block Fusion Module (IFM)** are incorporated into the training using a new scale-aware loss function. CAN and ASPP are integrated into the regression model to obtain multiscale features.

issue of annotation errors across scales. The Feature Pyramid (FP) can capture the visual features of objects from coarse to fine scales, and FP has become the standard component for most State-of-The-Art (SoTA) object counting frameworks [3], [4], [5], [6], [7], [8], [9], [10], [11], [15]. However, the adopted pooling operations in standard FP methods construct feature maps into $\frac{1}{2}$, $\frac{1}{4}$, or $\frac{1}{8}$ of the input size, where the scale truncation makes small objects disappearing. To address this issue, we propose a novel **Synthetic Fusion Module** (SFM) that constructs feature maps with finer scales of $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{6}$, *i.e.*, not just the half-sizing intervals. A smoother scale space can be obtained this way, to fit the ground truth whose scale in fact changes continuously. We further design an **Intra-block Fusion Module** (IFM) to fuse all feature layers within the same convolution block, so that more fine-grained information can be sent to the decoder for effective crowd counting. Finally, most existing crowd-counting architectures [3], [4], [5], [7], [8], [9], [16], [17], [18] do not meet the requirement in operating speed for real-time crowd-counting. To this end, our architecture can be easily converted to a lightweight version that brings real-time efficiency and comparable accuracy.

To address the problem of annotation errors, we propose a novel **scale-aware loss function** that simultaneously considers the annotation noise, head-to-head correlation, and adjustment for variances at different scales. In [19], a multivariate Gaussian distribution was used to handle this annotation problem; however, their model is fixed and not scale-aware for all objects of different sizes. In real images, the sizes of human heads vary at different positions; see Fig. 1(b). Thus, we argue that annotation error modeling should be scale-aware, and capable of adapting to changes in head size. We derive a multivariate Gaussian distribution with a full covariance matrix of different scales to model the correlation between pixels at different scales. To speed up computation, we adopt a low-rank approximation method. Finally, our scale-aware loss function is designed to compensate for human annotation errors and thus greatly improves the trained model performance. Our new architecture, **Scale-Aware Crowd Counting Network (SACC-Net)** as described in Fig. 2, is integrated into VGG-19 and trained by a new loss function with scale-aware annotation error modeling. SACC-Net achieves SoTA performance on five popular crowd-counting datasets.

Our work presents the following key contributions:

- We design a scale-invariant loss function for crowd-counting networks. Observing the typically skewed distribution of head sizes in images, we develop a novel scale-aware density model to effectively manage to accommodate annotation errors and scale variations. The introduction of new scale-aware loss function allows concurrent handling of scale variations and annotation errors, resulting in fine-grained crowd-counting density maps.
- We propose the SACC-Net as the new SoTA for crowd-counting. Our method integrates information across layers and compensates for annotation errors across scales. A synthetic fusion module (SFM) is proposed to generate a smoother scale space to address the scale truncation problem. An intra-block fusion module (IFM) is designed to fuse all feature layers within the same convolution block to generate finer-grained information for effective crowd counting.
- We report comprehensive experimental results to justify the superiority and generalization properties of the proposed scale-aware loss function. When combined with recently developed STEERER [20] architecture, our scale-aware loss function notably improves the Mean Absolute Error (MAE)/Mean Square Error (MSE)

performance on both backbones of VGG19 and HRNet. The proposed SACC-Net outperforms all SoTA methods on six popular crowd-counting datasets.

- A light-weight version of SACC-Net named SACC-LW, employs a bifurcation design that divides the feature map processing into two routes (a VGG block and a Simple Convolution Block), improving the running FPS by more than twice while retaining accuracy. In practical applications, SACC-LW supports real-time tracking and counting such as early warning of stampede accidents in social gatherings.

II. Related Works

We survey literature covering various aspects of image-based crowd counting.

A. Scale Variations

A key challenge in crowd counting methods relying on summing density maps is the scale variation arising from different distances between cameras and targets. A CNN with a switching strategy is employed in [21] to address this, by optimizing between density and count estimation for enhanced generalizability. In [22], a multi-column CNN utilizes varied convolution kernels for extracting multi-scale features, but [3] observes redundancy in feature learning, hindering efficient training with deeper layers. To tackle this, [3] adopts VGG16 for obtaining multi-scale features using convolutions with different dilation rates. Alternatively, a multi-branch strategy is employed in [8] to select fixed-size convolution filters in each layer for consistent multi-scale feature extraction. The Pan-Density Network in [16] effectively captures global and local contextual features to count crowds with varying density. To avoid redundant convolutional feature computations, multi-resolution feature maps are generated in [7] by dividing dense regions into sub-regions. To handle scale variations, multi-scale contextual information is encoded into a regression model [23]. In [9], a density attention network generates various attention masks to focus on a particular scale. Multi-scale information is maintained using a densely connected architecture in [24]. Multiple kernels are adopted in [15] to generate various density maps for a given image to count crowds more accurately in a semi-supervised way. In [10], both CAN [23] and ASPP [25] are integrated into the regression model to obtain multiscale features. In [18], a hierarchical mixture of density experts merges multi-scale density maps to overcome problems such as perspective distortions and crowd variations in crowd counting.

B. Annotation Deviation

Crowd-counting datasets commonly utilize *point-wise* or *dotted* annotations to denote individual objects in an image. Unlike bounding-box annotations, the lack of size information leads to variations in subsequent deviation and performance assessment. To address this issue, the average distance from each head to its three neighbors is calculated in [22] to estimate the head size using Gaussian standard deviation. The effect of body structure on crowd counting was studied in [26]. Locally connected Gaussian kernels are introduced in [27], replacing convolution filters to relax pixel-level spatial invariance for object counting. In [28], a decoupled two-stage crowd counting was proposed to partially alleviate this problem. A recent study in [29] explores location-agnostic crowd counting.

C. Loss Function

Traditionally, density-based crowd-counting approaches employ pixel-wise Mean Square Error (MSE) loss for training. However, recent developments have overcome the limitations of MSE loss. For instance, [30] introduced a combinatorial loss that incorporates spatial abstraction and correlation terms to effectively reduce annotation deviation. The Bayesian Loss (BL) in [6] leverages a density contribution probability model to address deviation impact, although it struggles to reduce false positives. The Density Map (DM)-count loss in [31] gauges the similarity between predicted and ground-truth density maps. The multivariate Gaussian distribution-based loss in [19] considers annotation noise and correlation but lacks scale awareness in its design. In this paper, we recognize that annotation pixel errors can significantly degrade the counting of small objects (*i.e.*, people far away). We thus derive a scale-aware loss function to rectify such annotation error.

D. Network Architecture

The latest survey [14] indicates that CNNs remain the dominant choice for crowd counting. However, recent literature explores alternative designs, such as the adoption of a Multi-Layer Perceptron (MLP) in CrowdMLP [32]. CrowdMLP focuses on modeling global dependencies of embeddings and regress total counts using a multi-granularity MLP regressor. Attention-based enhancements have also become popular, with works like Hierarchical Attention [17], Context Attention Fusion Network [33], Dual Attention Network [34], and Feature Pyramid Attention [35] being notable examples. Additionally, the influence of transformer architecture is evident, with recent studies like [36] and [37] exploring the application of vision transformers in crowd counting. Finally, an interesting recent development is the integration of vision-language models into crowd counting, as seen in [38]. Knowledge distillation has also been developed for efficient crowd counting in [39].

E. Multi-Scale Feature Extraction

Recent advancements in crowd and object counting tackle common challenges such as scale variation, background interference, and efficiency, particularly in dense and dynamic environments. Several models have emerged with unique solutions: the Ghost Attention Pyramid Network (GAP-Net) [40] and Attentive Hierarchy ConvNet (AHNet) [41] employ lightweight designs with attention mechanisms to improve multi-scale feature extraction and operational efficiency, making them well-suited for smart city applications. The Scale-Context Perceptive Network (SCPNet) [42] and Scale Region Recognition Network (SRRNet) [43] improve both counting and localization accuracy by integrating context and scale recognition modules. Meanwhile, the Group

and Graph Attention Network (GGANet) [44] and Group-Split Attention Network (GSANet) [45] leverage attention mechanisms to minimize background noise. Additionally, the Deep Spatial Prior Interaction (DSPI) [46] network extends capabilities to zero-shot counting, enabling adaptability in diverse applications. Together, these innovations signify substantial progress in object counting, combining accuracy, adaptability, and computational efficiency across various real-world scenarios.

The methods discussed above mainly emphasize improving accuracy, often with limited focus on addressing annotation errors across scales and efficiency. In contrast, our approach makes a novel contribution by enhancing both accuracy and efficiency. We propose a scale-aware loss function that tackles annotation noise, head-to-head correlations, and variance adjustments across different scales to boost accuracy. Additionally, this paper introduces a bifurcation design that splits feature map processing into two parallel paths, doubling the running speed in frames per second (FPS) while maintaining accuracy.

III. Method

This section describes the proposed network architecture and how it is trained. § III-A starts with the basic formulation of generating a density map for crowd counting. § III-B introduces our Scale-Aware Crowd Counting Network (SACC-Net), which generates the density map from an input image. § III-C explains how we model uncertainty or noise introduced by manual ground-truth annotation. § III-D elucidates the representation of the scale-aware crowd density function as a Gaussian normal distribution. Due to the large covariance matrix in this Gaussian representation, § III-E outlines the computation of a low-rank approximation using SVD. In § III-F, we describe how this low-rank approximation is employed to define the final loss function, incorporating regularization. Finally, § III-G shows a lightweight version called SACC-LW, with improved efficiency and only a slight degradation of counting accuracy.

A. Crowd Counting Density Map Generation

Traditional methods treat the counting task as a density regression problem [19], [47], [48]. Given an image $\mathcal{I}$ with N people to be counted, let $\mathbf{H}_i$ denote the true position of the head of the i^{th} person. For any pixel location x in the image $\mathcal{I}$, the crowd density y at x is modeled as a Gaussian kernel centered at each annotation point. Let β denote the annotation variance of the Gaussian kernel, and let $\sum_{i=1}^{N}\mathcal{N}(x|\mu,\Sigma)$ denote the Probability Density Function (PDF) for a multivariate Gaussian with mean μ and covariance matrix Σ. We calculate the squared Mahalanobis distance as $\|x\|_{\Sigma}^2 = X^T\Sigma^{-1}X$, where X is the feature vector of x extracted from a network backbone. The crowd density y at position x is modeled as:

$$y(x)=\sum_{i=1}^{N}\mathcal{N}(x|\mathbf{H}_i,\beta\mathbf{I})=\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi\beta}}\exp\left(-\frac{\|x-\mathbf{H}_i\|_{\beta\mathbf{I}}^2}{2}\right). \quad (1)$$

From data-driven learning, the density map y for all annotated head positions $\mathbf{H}_i$ in the image $\mathcal{I}$ is estimated by a regressor $f(\mathcal{I})$, where the learning objective is typically defined using the L_2 loss $\mathcal{L}(y, f(\mathcal{I})) = |y - f(\mathcal{I})|^2$ or a Bayesian loss [6]. The crowd count is obtained by summing up the values of the density map y across all pixels in $\mathcal{I}$.

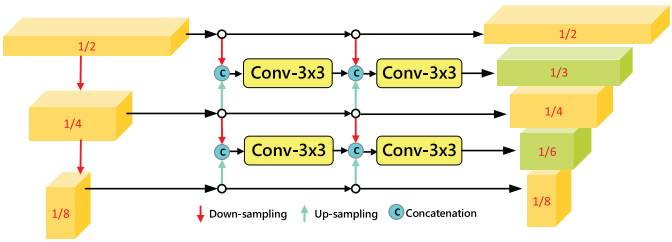


Fig. 3. The **Synthetic Fusion Module (SFM)** produces a smoother scaling space for crowd counting density map generation.

B. Scale-Aware Crowd Counting Network (SACC-Net)

Our scale-aware crowd-counting network features several newly designed modules, leading to improved crowd-counting density map estimation. Fig. 2 overviews our SACC-Net architecture. First, the *Synthetic Fusion Module (SFM)* is introduced to produce a refined feature map interpolation across scales, which effectively overcomes issues of uneven feature fusion resulting from the typical stride-2 down-sampling. Secondly, the *Intra-block Fusion Module (IFM)* effectively fuses all feature layers within the same convolution block, such that more fine-grained information [49] can be sent to the decoder for crowd counting. Finally, the ASPP [25] and CAN [23] modules are adopted at the end of SACC-Net, to leverage *atrous* convolutions with different rates to extract multiscale features for accurate counting. More precisely, the ASPP module is added to the last layer (Conv5-4) of the used backbone and the CAN module follows the last IFM module (see Fig. 2). Details of SFM and IFM are described as follows.

1) Synthetic Fusion Module (SFM): Unlike typical CNN backbones that employ down-sampling through pooling or stride-2 convolution to create feature maps of $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$ intervals, our newly designed Synthetic Fusion Module (SFM) enhances scale-aware crowd counting by offering denser scale space samples. This is more aligned with the *continuous* scale changes observed in reality. SFM achieves this by generating *synthetic layers* between original layers, resulting in refined density scales at $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{6}$, $\frac{1}{8}$, and beyond.

The input configuration of SFM varies depending on its position within SACC-Net, as illustrated by the brown blocks in Fig. 2. SFM can take two or three inputs in generating the synthetic layers. Fig. 3 depicts how SFM works, which involves down-sampling and up-sampling of feature scale space. SFM initially performs linear scaling of the inputs, followed by merging through a 1×1 convolution. The results are further fused using a 3×3 convolution. This process synthesizes a new feature layer from the two original adjacent layers, contributing to a smoother scaling space for crowd counting.

a) Down-Sampling: As shown in Fig. 4, the input feature map of size $W \times H$ is first disassembled into regular 4×4 patches. Then, a convolution operation with kernel 2×2 and stride 1 is applied to obtain a new feature patch with size

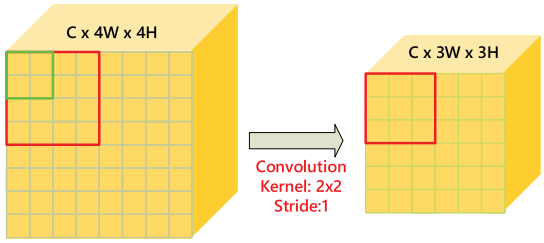


Fig. 4. The Down-sampling process in SFM. A convolution kernel 2×2 (denoted by green) with stride 1 is used to convert a 4×4 patch to a new 3×3 feature map.

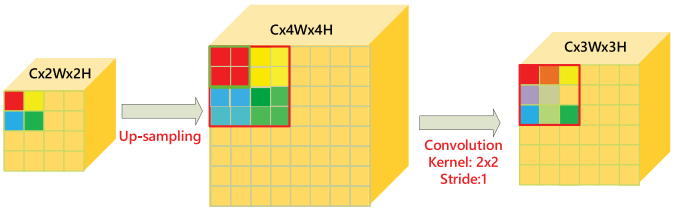


Fig. 5. The Up-sampling process in SFM. Each 2×2 feature patch is up-sampled to 4×4, followed by a convolution operation using a 2×2 kernel and a stride of 1, resulting in a new 3×3 feature map.

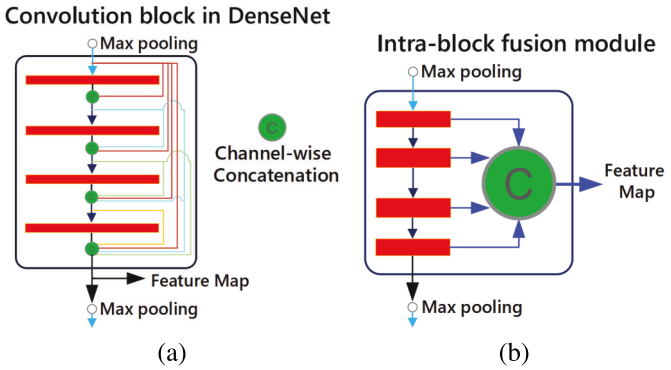


Fig. 6. Comparison between (a) convolution block in DenseNet [50] and (b) our **Intra-block Fusion Module (IFM)**.

3×3. This process reduces the feature map dimensions from $C \times 4W \times 4H$ to $C \times 3W \times 3H$.

b) Up-Sampling: As shown in Fig. 5, each 2×2 feature patch is first up-sampled to 4×4. Then, a convolution operation with kernel 2×2 and stride 1 is applied to the 4×4 patch to obtain a new 3×3 feature map. After that, the feature map with dimension $2W \times 2H$ is enlarged to dimension $3W \times 3H$.

2) Intra-Block Fusion Module (IFM): In traditional CNN architectures like VGG, features are extracted through sequential convolutions, with only the feature maps from the last layer of a convolution block transmitted to the next module. Our newly designed Intra-block Fusion Module (IFM) diverges from this approach by allowing all layers within the convolution block to contribute fine-grained features for precise density map generation. In contrast to the DenseNet structure [50] depicted in Fig. 6(a), which employs a fully connected structure linking all layers, potentially causing training challenges and inefficiencies, our IFM design, shown in Fig. 6(b), utilizes fewer connections than DenseNet and ensures efficient generation of required feature maps. IFM offers three advantages over DenseNet: (1) IFM requires less memory usage, as it uses 1×1 convolution to directly obtain the output; (2) IFM obtains more representative features by aggregating information from all layers within the convolution block; (3) IFM contains fewer parameters, making it more efficient compared to DenseNet.

C. Annotation Noise Modeling

We address uncertainty in manual annotations of the human head position as the example shown in Fig. 1(a). These point-wise annotation errors lead to inaccuracies in training the data-driven model for image-based crowd density estimation y defined in Eq. (1). We next derive a solution to address this issue. Let $\tilde{\mathbf{H}}_i$ denote the annotated head position of the i^{th} person with potential annotation error, and ε_i denote its annotation noise, $\tilde{\mathbf{H}}_i = \mathbf{H}_i + \varepsilon_i$. We assume the annotation noise is independent and identically distributed (i.i.d), $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \alpha\mathbf{I})$, where α is an annotation variance parameter. Recall that β is the Gaussian annotation variance defined in Eq. (1). Let $q_i = x - \mathbf{H}_i$ denote the position difference between the i^{th} annotation and position x. Let ϕ_i denote a Gaussian kernel for the i^{th} annotation. Considering the annotation noise, we model the density $\mathbb{D}(x)$ at location x as the sum of the individual Gaussian kernels:

$$\begin{aligned}\mathbb{D}(x) &= \sum_{i=1}^{N} \mathcal{N}(x|\tilde{\mathbf{H}}_i, \beta\mathbf{I}) = \sum_{i=1}^{N} \mathcal{N}(x|\mathbf{H}_i + \varepsilon_i, \beta\mathbf{I}) \\ &= \sum_{i=1}^{N} \mathcal{N}(q_i|\varepsilon_i, \beta\mathbf{I}) \cong \sum_{i=1}^{N} \phi_i. \end{aligned} \tag{2}$$

In the literature, [19] did not distinguish the range of annotation errors between small and large objects. It is a fixed-scale model using the NoiseCC loss to rectify the annotation noise. A common limitation in all state-of-the-art methods [19], [47], [48] regarding Eq. (2) is the use of a fixed β with constant value to model $\mathbb{D}(x)$ of the crowd density around each head position. As mentioned earlier, head sizes vary based on their distances relative to the camera. Therefore, our approach makes β scale-aware and adaptive to the size of each head present in the image.

D. Scale-Aware Gaussian Density Function

We observed a positively skewed distribution of people's head size β in the crowd counting datasets, with smaller heads being more frequent, as shown in Fig. 1(b). We introduce a scale-aware annotation error using mixed Gaussian distributions for density map generation. Assuming there are only S scales utilized in modeling a person's head with annotation errors, we represent the density of a head as a mixed Gaussian model. Eq. (2) can be rewritten as:

$$\mathbb{D}(x) = \sum_{i=1}^{N} \sum_{s=1}^{S} w_s \mathcal{N}(q_i|\varepsilon_i, \beta_s\mathbf{I}) \cong \sum_{s=1}^{S} w_s \sum_{i=1}^{N} \phi_i^s, \tag{3}$$

where $\{w_s\}$ are weights and $\sum_{s=1}^{S} w_s = 1$. In addition, $\phi_i^s = \mathcal{N}(q_i|\varepsilon_i, \beta_s\mathbf{I})$, *i.e.*, the Gaussian kernel placed in the ith annotation at the scale s and parameterized with the annotation error ε_i and the variance β_s. Let $\mathbb{D}_s = w_s \sum_{i=1}^{N} \phi_i^s$. Then, Eq.(3) can be rewritten as

$$\mathbb{D}(x) = \sum_{s=1}^{S} w_s \mathbb{D}_s(x). \tag{4}$$

Let $\mathcal{I}_s$ and J_s be the scaled-down version of $\mathcal{I}$ on scale s and the number of pixels in $\mathcal{I}_s$, respectively. For all pixels x_j in $\mathcal{I}_s$, a multivariate random variable for the density map $\mathbb{D}_s(x)$ is constructed as

$$\mathbb{D}_s = [\mathbb{D}_s(x_1), \cdots, \mathbb{D}_s(x_j), \cdots, \mathbb{D}_s(x_{J_s})]. \tag{5}$$

1) Scale-Aware Probability Distribution: To calculate $\mathbb{D}_s$ in a closed form, we approximate $\mathbb{D}_s$ as a Gaussian function using the scale-aware mean μ_s and the variance Σ_s^2 as $\hat{p}(\mathbb{D}_s) \sim \mathcal{N}(\mathbb{D}_s|\mu_s, \Sigma_s^2)$. The mean μ_s is calculated as:

$$\begin{aligned}\mu_s = \mathbb{E}[\mathbb{D}_s] &= \mathbb{E}\left[w_s \sum_{i=1}^{N} \mathcal{N}(q_i|\varepsilon_i, \beta_s \mathbf{I})\right] \\ &= w_s \sum_{i=1}^{N} \mathcal{N}(q_i|0, (\alpha+\beta_s)\mathbf{I}) = \sum_{i=1}^{N} \mu_i^s,\end{aligned} \tag{6}$$

where $\mu_i^s = w_s \mathcal{N}(q_i|0, (\alpha+\beta_s)\mathbf{I})$ and the annotation error $\varepsilon_i \sim \mathcal{N}(0|0, \alpha\mathbf{I})$. The variance Σ_s^2 is calculated by:

$$\begin{aligned}\Sigma_s^2 &= \text{var}(\mathbb{D}_s) = \mathbb{E}[\mathbb{D}_s^2] - \mathbb{E}[\mathbb{D}_s]^2 \\ &\cong \sum_{i=1}^{N} \left[\frac{w_s^2}{4\pi\beta_s} \mathcal{N}\left(q_i|0, (\beta_s/2+\alpha)\mathbf{I}\right) - \left(\mu_i^s\right)^2\right].\end{aligned} \tag{7}$$

2) Gaussian Approximation to Scale-Aware Joint Likelihood $\boldsymbol{D}_s$: We next calculate the covariance $\text{Cov}(\mathbb{D}_s(x_j), \mathbb{D}_s(x_k))$ between locations x_j and x_k. This term is modeled using a multivariate Gaussian approximation of the joint likelihood $\mathbf{D}_s$ at scale s. Let $q_i(x_j) = x_j - \tilde{\mathbf{H}}_i$ be the difference between the spatial location of the i^{th} annotation and the pixel x_j location. The density $\mathbb{D}_s(x_j)$ is calculated using Eq. (3) as:

$$\mathbb{D}_s(x_j) = w_s \sum_{i=1}^{N} \mathcal{N}(q_i(x_j)|\varepsilon_i, \beta_s\mathbf{I}) = w_s \sum_{i=1}^{N} \phi_i^s(x_j), \tag{8}$$

where $\phi_i^s(x_j) = \mathcal{N}(q_i(x_j)|\varepsilon_i, \beta_s\mathbf{I})$ and the annotation noise ε_i is the same random variable across all $\phi_i^s(x_j)$. Define the Gaussian approximation to $\mathbf{D}_s$ as $\hat{p}(\mathbf{D}_s) = \mathcal{N}(\mathbf{D}_s|\mu_s, \Sigma_s)$, where μ_s and Σ_s are defined in Eqs. (6) and (7). From Eq. (6), the j^{th} entry in μ_s is $\mathbb{E}[\mathbb{D}_s(x_j)] = \sum_{i=1}^{N} \mu_i^s(x_j)$. The diagonal of the scale-aware covariance matrix is calculated as $\boldsymbol{\Sigma}^s_{x_j,x_j} = \text{Var}(\mathbb{D}_s(x_j))$. The covariance term is then:

$$\begin{aligned}\boldsymbol{\Sigma}^s_{x_j,x_k} &= \text{Cov}(\mathbb{D}_s(x_j), \mathbb{D}_s(x_k)) \\ &= \sum_{i=1}^{N} \left[w_s^2 \Omega_i^s(x_j, x_k) - \mu_i^s(x_j)\mu_i^s(x_k)\right],\end{aligned} \tag{9}$$

where $\Omega_i^s(x_j, x_k) = \mathbb{E}[\phi_i^s(x_j)\phi_i^s(x_k)]$.

E. Low-Rank Approximation Using SVD

Due to the vast dimension of $\boldsymbol{\Sigma}^s_{x_j,x_k}$ that is $J_s \times J_s$, we derive a low-rank approximation with non-zero rows and columns for efficiency improvement. Let $\hat{\boldsymbol{\Sigma}}^s$ denote the approximation to $\boldsymbol{\Sigma}^s$ using Singular Value Decomposition (SVD), which is calculated as:

$$\hat{\boldsymbol{\Sigma}}^s \cong \boldsymbol{\Sigma}^s = \mathbf{U}^s \mathbf{C}_L^s \mathbf{V}^T, \tag{10}$$

where $\mathbf{U}^s$ is a $J_s \times J_s$ orthogonal matrix, $\mathbf{C}_L^s$ is a non-negative $J_s \times J_s$ diagonal matrix with diagonal entries sorted from high to low, and $\mathbf{V}^T$ is a $J_s \times J_s$ orthogonal matrix. Let $v_j^s = \boldsymbol{\Sigma}^s_{x_j,x_j}$. To obtain this low-rank approximation, each pixel x_j is first ordered by v_j^s. Then, the top-M pixels whose percentages of variance are larger than a threshold γ are selected from $\mathcal{I}_s$ for this low-rank approximation:

$$\frac{\sum_{j=1}^{M} v_j^s}{\sum_{i=1}^{J} v_j^s} > \gamma. \tag{11}$$

Let the set of indices of the top M pixels be denoted by L, *i.e.*, $L = \{l_1, l_2, \ldots, l_m, \ldots, l_M\}$. Then, only the elements in L are selected to approximate $\boldsymbol{\Sigma}^s$. We ensure that the low-rank approximation retains the majority of the informative structure in the data while discarding less significant components. This trade-off between dimensionality reduction and accuracy preserves most of the useful information, aligning with our goal of enhancing computational efficiency without compromising the model's performance in crowd density estimation, as illustrated in Eq. (11). Table III shows that optimal performance is achieved when the threshold γ is set to 0.8. Approximation of the matrix $\boldsymbol{\Sigma}^s$ by a rank-M matrix requires a representation of $\boldsymbol{\Sigma}^s$ as the sum of several terms ordered by their importance. SVD achieves this by transforming $\boldsymbol{\Sigma}^s$ into the sum of rank-1 matrices that is weighted by the corresponding singular values. Namely, $\boldsymbol{\Sigma}^s = \mathbf{U}^s \mathbf{C}_L^s \mathbf{V}^{sT}$ is equivalent to:

$$\boldsymbol{\Sigma}^s = \sum_{i=1}^{J_s} c_i^s \cdot \mathbf{u}_i^s \mathbf{v}_i^{sT}, \tag{12}$$

where the scale $s = 1, \ldots, S$, c_i^s is the i^{th} singular value, and $\mathbf{u}_i^s$ and $\mathbf{v}_i^{sT}$ are the corresponding left and right singular vectors

TABLE I
DETAILED PARAMETERS USED FOR TRAINING

Dataset	learning rate	batch size	crop size
UCF-QRNF	1e-5	12	512×512
UCF CC 50	1e-5	10	512×512
NWPU	1e-5	8	512×512
ShanghaiTech	1e-4	12	512×512
JHU	1e-4	10	512×512

TABLE II
ABLATION STUDY OF THE GAUSSIAN ANNOTATION VARIANCE β_1. THE BEST AND SECOND-BEST RESULTS ARE SHOWN IN RED AND BLUE

β_1	2	4	6	8	10	12	14	16
MAE	187.4	179.2	174.5	157.3	168.3	177.6	189.2	198.1
MSE	273.1	265.4	253.7	235.9	246.6	255.3	257.7	260.4

TABLE III
PERFORMANCE ANALYSES OF THE THRESHOLD γ IN EQ.(11) AMONG DIFFERENT DATASET

γ	UCF-QNRF		NWPN		JHU		S.H.Tech-A	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
0.2	95.6	177.2	91.2	277.5	82.1	270.9	79.5	99.3
0.4	91.7	155.8	87.6	266.7	74.3	259.5	70.4	93.7
0.6	82.8	138.9	79.4	246.3	63.5	223.4	61.9	89.4
0.7	79.4	133.1	76.2	223.7	60.9	227.5	59.7	86.5
0.75	75.8	128.5	73.6	216.4	57.7	213.7	55.8	81.3
0.8	73.9	121.7	70.0	211.4	53.6	201.5	52.1	76.6
0.85	74.7	126.3	73.2	214.7	55.9	209.9	53.8	80.4
0.9	75.1	132.2	73.9	219.2	56.7	215.4	54.4	85.9

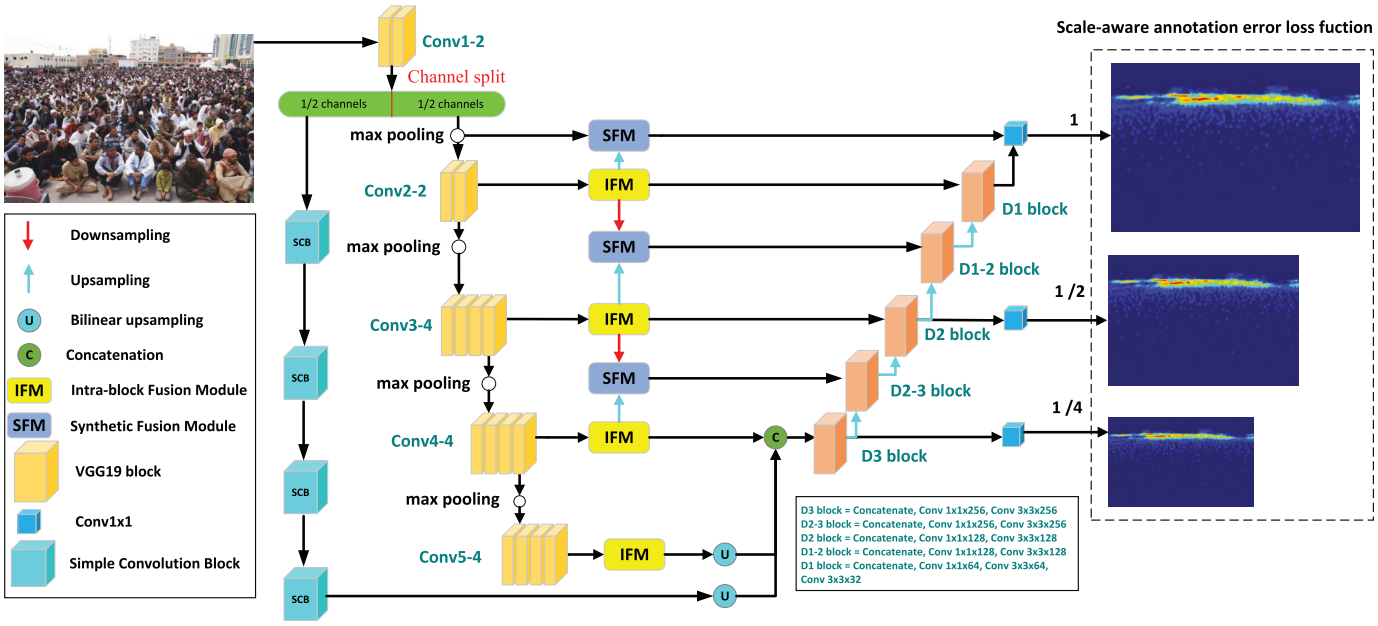


Fig. 7. Architecture of the lightweight SACC-LW model. With the two-way feature branching approach, only half of the channels go through the VGG-19 backbone compared to the full version.

at the scale s, respectively. A straightforward idea is to keep only the top M terms on the right-hand side of Eq. (12). For $\mathbf{\Sigma}^s$ in Eq. (12) and a target rank M, the proposed rank-M approximation is then:

$$\hat{\mathbf{\Sigma}}^s \cong \sum_{i=1}^{M} c_i^s \cdot \mathbf{u}_i^s \mathbf{v}_i^{sT}, \tag{13}$$

where the singular values c_i^s are sorted as $c_1^s \geq c_2^s \geq \cdots \geq c_{J_s}^s \geq 0$. With $\hat{\mathbf{\Sigma}}^s$, the rank-M approximate negative log-likelihood function is:

$$-\log\ \hat{p}(\mathbf{D}_s) = -\log \mathcal{N}(\mathbf{D}_s|\mu_s, \hat{\mathbf{\Sigma}}^s) \propto \|\mathbf{D}_s - \mu_s\|^2_{\hat{\mathbf{\Sigma}}^s}. \tag{14}$$

The time complexity to calculate the right-hand sides of Eqs. (13) and (14) take $O(M^3)$ and $O(M^2)$, in contrast to $O(J_s^3)$ and $O(J_s^2)$ that are required to calculate the original matrix $\mathbf{\Sigma}^s$ and the distance $\|\mathbf{D}_s - \mu_s\|^2_{\mathbf{\Sigma}^s}$, respectively.

F. Regularization and the Final Loss Term

We calculate the Gaussian approximation to $\mathbf{D}_s$ based on Eqs. (13) and (14). To ensure that the predicted density map near each annotation satisfies the condition of density values summing to 1, we define the regularizer $\mathcal{R}_i^s$ for the i-th annotation point as:

$$\mathcal{R}_i^s = \left| \sum_j \mathbb{D}_s(x_j) \frac{\phi_i^s(x_j)}{\sum_{i=1}^{N} \phi_i^s(x_j)} - 1 \right|, \tag{15}$$

where $\mathbb{D}_s(x_j)$ is the j^{th} term of $\mathbf{D}_s$. Let $\bar{\mathbf{D}}_s = \mathbf{D}_s - \mu_s$. The final loss function is:

$$\mathcal{L} = \sum_{s=1}^{S} \bar{\mathbf{D}}_s^T (\hat{\mathbf{\Sigma}}^s)^{-1} \bar{\mathbf{D}}_s + \sum_{s=1}^{S} \sum_{i=1}^{N} \mathcal{R}_i^s, \tag{16}$$

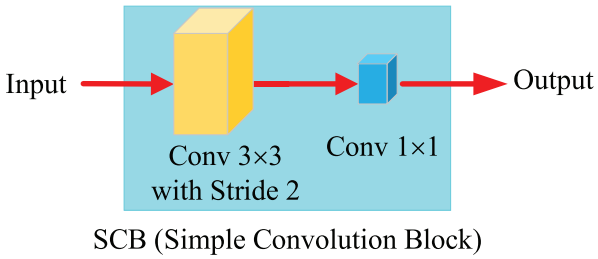


Fig. 8. Convolutions in the Simple Convolution Block (SCB).

where the first term is the low-rank approximation from Eq. (14) and the second term is the accumulation of regularization terms from Eq. (15).

G. Light-Weight Version (SACC-LW)

Density-based crowd-counting methods demonstrate notable counting accuracy, yet their efficiency for real-time applications remains a challenge. Note that the time complexities for both SFM and IFM are $O(W \times H \times C)$, where C is the number of channels used, and W and H are the input width and height. Such high computational complexity might become a roadblock to support real-time applications such as the early warning of stampede accidents in social gatherings.

To improve the efficiency of the original SACC-Net, we employ a bifurcation design to effectively balance the computation load across layers and reduce memory demands. We divide the feature map processing of the `Conv-1-2` block of Fig. 2 into two routes. The new light-weight architecture, SACC-LW, has one processing branch going through a VGG block and the other branch passing directly through the Simple Convolution Block (SCB), as shown in Fig. 7. This design is motivated by our observation that additional parameters in heavier models often lead to the learning of redundant

TABLE IV
ACCURACY COMPARISONS AMONG DIFFERENT LOSS FUNCTIONS WITH VARIOUS BACKBONES ON UCF-QNRF

	VGG19		CSRNet		MCNN	
	MAE	MSE	MAE	MSE	MAE	MSE
L2	98.7	176.1	110.6	190.1	186.4	283.6
BL [6]	88.8	154.8	107.5	184.3	190.6	272.3
NoiseCC [19]	85.8	150.6	96.5	163.3	177.4	259.0
DM-count [31]	85.6	148.3	103.6	180.6	176.1	263.3
Gen-loss [56]	84.3	147.5	92.0	165.7	142.8	227.9
Ours	**73.91**	**121.7**	**90.83**	**150.67**	**134.52**	**213.71**

TABLE V
ACCURACY COMPARISONS AMONG DIFFERENT LOSS FUNCTIONS WITH VARIOUS BACKBONES ON SHANGHAITECH PART-A

	VGG19		CSRNet		MCNN	
	MAE	MSE	MAE	MSE	MAE	MSE
L2	71.4	136.5	80.61	149.12	147.8	201.6
BL [6]	62.8	101.8	68.2	115.0	110.2	173.2
NoiseCC [19]	61.9	99.6	67.28	109.31	105.5	169.7
DM-count [31]	59.7	95.7	65.71	105.53	103.8	165.4
Gen-loss [56]	61.3	95.4	63.42	102.51	102.3	162.9
Ours	**52.19**	**76.63**	**60.39**	**96.83**	**95.7**	**160.1**

TABLE VI
ACCURACY COMPARISONS AMONG DIFFERENT LOSS FUNCTIONS WITH VARIOUS BACKBONES ON SHANGHAITECH PART-B

	VGG19		CSRNet		MCNN	
	MAE	MSE	MAE	MSE	MAE	MSE
L2	9.1	13.9	11.63	17.57	29.17	52.33
BL [6]	8.62	13.56	10.6	16.0	26.4	41.3
NoiseCC [19]	8.37	13.13	10.47	15.69	25.33	41.04
DM-count [31]	7.4	11.8	9.76	13.82	23.91	35.49
Gen-loss [56]	7.3	11.7	9.51	13.66	22.89	33.77
Ours	**6.16**	**9.71**	**9.38**	**13.23**	**20.51**	**31.26**

TABLE VII
ACCURACY COMPARISONS AMONG DIFFERENT LOSS FUNCTIONS WITH VARIOUS BACKBONES ON JHU

	VGG19		CSRNet		MCNN	
	MAE	MSE	MAE	MSE	MAE	MSE
L2	85.9	354.1	90.7	388.4	104.3	412.5
BL [6]	75.0	299.9	82.3	364.2	91.4	399.6
NoiseCC [19]	67.7	258.5	72.5	334.1	83.4	360.9
DM-count [31]	68.4	283.3	75.6	356.9	90.4	378.2
Gen-loss [56]	59.9	259.5	71.4	291.7	85.7	357.2
Ours	**53.6**	**201.5**	**58.6**	**231.4**	**70.3**	**262.7**

or unnecessary features, which do not significantly enhance model accuracy. Consequently, we developed the architecture illustrated in Fig. 7.

The SCB consists of two simple convolutions as in Fig. 8. The separation can balance the computation load of each layer, as well as reduce the memory traffic load. In contrast to the original convolution on all channels, only half of the channels are sent to the next block, providing efficiency improvement. This design significantly reduces the model parameters without compromising accuracy. As shown in Table X later, SACC-LW can achieve nearly real-time counting ($> 25fps$) for 2K-resolution videos.

IV. EXPERIMENTAL RESULTS

We evaluated our crowd-counting models (SACC-Net and SACC-LW) and compared them with 20 State-of-The-Art (SoTA) methods across six public datasets, namely UCF-QNRF [51], UCF CC 50 [52], NWPU-Crowd [53], ShanghaiTech Parts-A [22], ShanghaiTech Parts-B [22], and JHU-CROWD++ [54].

A. Model Training Parameters

Our method was pre-trained on ImageNet [55] using the Adam optimizer. Given the varying image dimensions in the datasets, we crop patches of a fixed size at random locations and then augment the data by random horizontal flipping with a probability of 0.5. The learning rates during training are set to $1e^{-5}$, $1e^{-5}$, $1e^{-5}$, $1e^{-4}$, and $1e^{-4}$ for the UCF-QNRF, UCF CC 50, NWPU, ShanghaiTech, and JHU datasets, respectively. To stabilize the training loss change, we use batch sizes of 12, 10, 8, 12, and 10, respectively. The learning rates and batch sizes were selected according to the complexity of each dataset evaluated. For more complex datasets, larger updates to the model's weights are needed to ensure convergence during training. In contrast, simpler datasets require smaller learning rates to avoid overshooting the optimal point during gradient descent. For the ShanghaiTech and JHU datasets, higher learning rates were chosen due to their greater complexities and diversities. All training stage parameters are listed in Table I. Similar to other state-of-the-art methods [3], [4], [5], [7], [8], [9], [10], [11], performance is evaluated using mean absolute error (MAE) and mean squared error (MSE).

B. Parameter Settings for β_s, w_s, α, and γ

The head size distribution $P_{head}(h)$ shown in Fig. 1(a) is positive-skewed and can be derived by aggregating training data. The variance β in Eq. (1) should be proportional to the head size h. One option is to set the mean of h as the initial value of β_1, *i.e.*, $\beta_1 = \sum_h hP_{head}(h)$. However, a more favorable approach is to treat α and β_1 as manual hyperparameters. Learning them from data requires a large amount of data and introduces dependence on the training set, potentially impacting generalization capability. On the contrary, treating them as hyperparameters makes the method adaptable to different scenarios.

In a CNN backbone like VGG19, the pooling operation reduces the feature map size by half, consequently decreasing the head size in the feature map. Given β_s, the value of β_{s+1} can be recursively obtained as $\beta_{s+1} = \beta_s/2$. This subsampling operation also leads to the eventual disappearance of small heads. Subsequently, w_s is set to $P_{head}(\beta_{(S+1-s)})$, where S is the largest scale used to model $\mathbb{D}(x)$ in Eq. (3), and we set $S = 3$. After normalization, we ensure $\sum_{s=1}^{S} w_s = 1$.

We conducted an experiment to examine the impact of annotation variance α and annotation error variance β_1 on feature map generation. As illustrated in Fig. 9, an increase in β_1 results in a decrease in MAE. However, when $\beta_1 \geq 8$, the MAE begins to increase instead. Similarly, with small values of α, the MAE is large, but when $\alpha \geq 8$, the MAE decreases and tends to stabilize. Consequently, we set $\alpha = 8$ and $\beta_1 = 8$. In Table II presents the results of an ablation study on the Gaussian annotation variance parameter β_1. As β_1 increases from 2 to 10, both the mean absolute error (MAE) and mean squared error (MSE) decrease, reaching their lowest values at

TABLE VIII

PERFORMANCE COMPARISONS BETWEEN OUR METHOD AND STEERER [20] WITH/WITHOUT OUR LOSS FUNCTION. THE BEST AND SECOND-BEST RESULTS ARE SHOWN IN RED AND BLUE

Methods	Backbone	UCF-QNRF		NWPU		S. H. Tech-A		S. H. Tech-B		JHU	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
STEERER [20]	VGG19	76.7	135.1	68.3	318.4	55.6	87.3	6.8	10.7	55.4	221.4
STEERER + Our Loss		74.8	127.5	**65.7**	234.7	53.9	83.4	6.2	9.5	54.6	207.1
SACC-Net + Our Loss		**73.9**	**121.7**	70.01	**211.43**	**52.19**	**76.63**	**6.16**	**9.71**	**150.66**	**201.5**
STEERER [20]	HRNet [57]	74.3	128.3	63.7	309.8	54.5	86.9	5.8	8.5	54.3	238.3
STEERER + Our Loss		73.3	118.7	62.4	275.8	51.9	74.2	5.6	7.3	53.1	194.8
SACC-Net + Our Loss		**73.2**	**115.4**	**61.5**	**204.1**	**51.4**	**72.4**	**5.3**	**7.4**	**141.7**	**194.3**

TABLE IX

PERFORMANCE COMPARISONS AMONG THE SOTA CROWD COUNTING METHODS. THE BEST AND SECOND-BEST RESULTS ARE SHOWN IN RED AND BLUE

Methods	Venue	backbone	UCF-QNRF		NWPU		S. H. Tech-A		S. H. Tech-B		UCF CC 50		JHU	
			MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CSRNet [3]	CVPR'18	VGG16	-	-	121.3	522.7	68.2	115.0	10.3	16.0	266.1	397.5	121.3	387.8
CAN [23]	CVPR'19	VGG16	107	183	-	-	62.3	100.0	7.8	12.2	212.2	243.7	-	-
S-DCNet [7]	ICCV'19	VGG16	104.4	176.1	-	-	58.3	95.0	6.7	10.7	204.2	301.3	90.2	370.5
SANet [13]	ECCV'18	MCNN	-	-	190.6	491.4	67.0	104.5	8.4	13.6	258.4	334.9	190.6	491.4
BL [6]	ICCV'19	VGG19	88.7	154.8	105.4	454.2	62.8	101.8	7.7	12.7	229.3	308.2	-	-
SFANet [11]	Arxiv'19	VGG16	100.8	174.5	-	-	59.8	99.3	6.9	10.9	-	-	-	-
DM-Count [31]	NeurIPS'20	VGG19	85.6	148.3	88.4	498.0	59.7	95.7	7.4	11.8	211.0	291.5	88.4	388.6
RPnet [21]	CVPR'15	VGG16	-	-	-	-	61.2	96.9	8.1	11.6	-	-	-	-
AMSNet [58]	ECCV'20	VGG19	101.8	163.2	-	-	56.7	93.4	6.7	10.2	208.4	297.3	-	-
M-SFANet [10]	ICPR'21	VGG19	85.6	151.2	-	-	59.6	95.6	6.3	10.2	162.3	276.7	-	-
TEDnet [30]	CVPR'19	VGG19	113.0	188.0	-	-	64.2	109.1	8.2	12.8	249.4	354.5	-	-
P2PNet [59]	ICCV'21	VGG16	85.3	154.5	77.4	362	52.7	85.0	6.2	9.9	172.7	256.1	-	-
GauNet [27]	CVPR'22	ResNet50	81.6	153.7	-	-	54.8	89.1	6.2	9.9	186.3	256.5	-	-
MAN [60]	CVPR'22	VGG19	77.3/83.4*	131.5/146*	76.5/76.6*	323.0/465.4*	56.8	90.3	-	-	-	-	-	-
HA-CCN [17]	TIP'19	VGG16	118.1	180.4	-	-	62.9	94.9	8.1	13.4	256.2	348.4	-	-
PaDNet [21]	TIP'19	VGG19	96.5	170.2	-	-	59.2	98.1	8.1	12.2	185.8	278.3	-	-
HMoDE+REL [18]	TIP'22	VGG19	81.6	153.7	73.4	331.8	54.4	87.4	6.2	9.8	159.6	211.2	55.7	214.6
ADM [61]	TIP'23	ResNet50	74.5	149.7	70.1	266.9	76.7	127.3	-	-	-	-	72.9	279.7
MRL [15]	TIP'23	VGG19	126.7	209.7	97.0	413.5	68.3	111.9	11.0	17.6	-	-	72.9	279.7
GGANet [44]	TNNLS'23	-	91.9	158.6	-	-	62.0	110.7	7.4	13.1	189.0	288.7	69.6	277.4
GAPNet [40]	FGCS'23		118.5	217.2	174.1	514.7	67.1	110.4	9.8	15.2	202.8	246.9	-	-
SRRNet [43]	TITS'23	HRNet [57]	89.5	162.9	-	-	60.8	103.0	7.4	13.6	172.9	256.3	62.4	254.6
SCPNet [42]	IoT'23	HRNet [57]	93.7	164.3	-	-	57.3	102.1	7.5	13.8	**132.0**	295.0	66.2	251.0
DKD [39]	TIP'24	-	91.7	150.1	97.0	413.5	64.4	103.0	7.4	12.7	210.3	283.8	-	-
SACC-Net + Our Loss	-	VGG16	77.3	142.9	75.2	254.1	52.4	78.7	6.1	9.8	156.4	208.7	56.2	204.1
SACC-Net + BL Loss	-	VGG19	85.4	145.4	86.7	442.9	55.28	90.3	6.5	10.6	167.4	235.4	-	-
SACC-Net + Our Loss	-	VGG19	**73.9**	**121.7**	**70.0**	**211.4**	**52.1**	**76.6**	**6.1**	**9.7**	150.6	**187.8**	**53.6**	**201.5**
SACC-LW + Our Loss	-	VGG16	83.8	149.2	88.2	304.8	54.3	90.8	6.2	10.7	167.5	231.6	58.7	253.1
SACC-LW + BL Loss	-	VGG19	90.2	175.1	99.3	490.7	63.5	103.7	7.6	11.5	175.1	253.4	-	-
SACC-LW + Our Loss	-	VGG19	81.4	144.5	85.3	288.4	53.7	88.9	6.2	10.1	157.1	203.6	56.4	232.1

Symbol * denotes scores produced by running the original source codes provided by the authors.

TABLE X

EFFICIENCY COMPARISON OF OUR SCAA-NET AND SCAA-LW AGAINST SOTA METHODS ON A SINGLE NVIDIA 2080TI GPU

Methods	Frames per second (FPS)		
	512 × 384	512 × 512	1280 × 720
CAN [23]	41.56	33.42	13.05
M-SFANet [10]	42.28	31.45	12.45
SFANet [11]	39.71	30.54	11.16
ADM [61]	43.96	33.14	12.91
MAN [60]	40.16	31.82	11.30
SACC-Net	25.24	20.61	8.19
SACC-LW	**57.37**	**45.16**	**25.07**

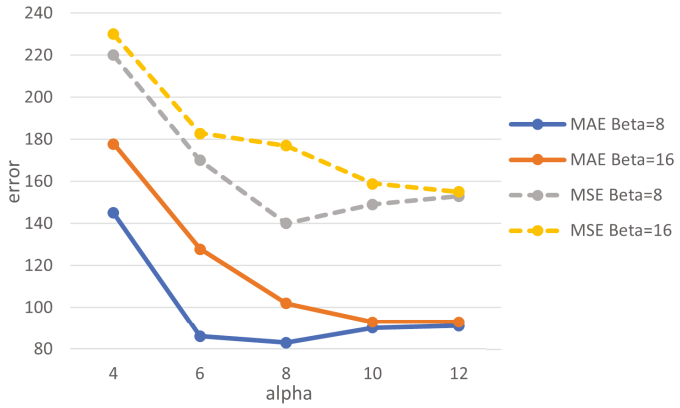


Fig. 9. MAE varies with different values of annotation variance (α) and Gaussian annotation variance variance (β_1). The lowest MAE is observed when initial values are set to $\beta_1 = 8$ and $\alpha = 8$.

$\beta_1 = 8$, with MAE at 157.3 and MSE at 235.9. This indicates that setting $\beta_1 = 8$ achieves the best performance. However, as β_1 continues to increase beyond 8, both MAE and MSE start to rise again, suggesting that larger values of β_1 may lead to decreased accuracy. This trend demonstrates that an optimal choice of β_1 is crucial for minimizing errors in the model.

Regarding the parameter γ, we conducted an additional experiment to evaluate its effects on accuracy. In Table III, setting γ to 0.8 yields optimal performance across multiple datasets, including UCF-QNRF, NWPU, JHU, and S.H.Tech-A. Specifically, in the QNRF and NWPU datasets, the MAE

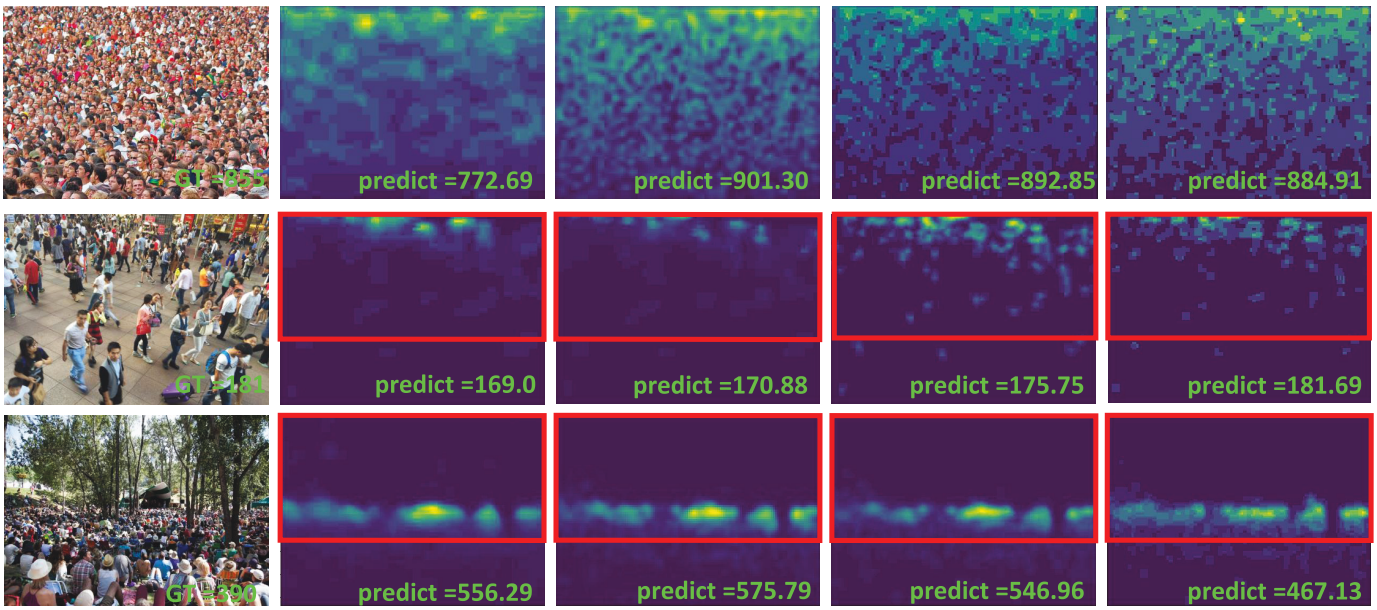


Fig. 10. Visualizations of the crowd counting heatmap generated using different loss functions on ShanghaiTech Part-A and Part-B. (a) The input image with ground truth. (b-d) show the heaptmaps generated using (b) MSE loss, (c) Bayesian loss, (d) NoiseCC loss, and (e) our scale-aware loss.

and MSE reach their lowest values of 73.9 and 121.7 (QNRF) and 70.0 and 211.4 (NWPU), respectively. Conversely, when γ is set to 0.2, the MAE is the highest (worst) across several datasets. As γ increases, the MAE gradually decreases, reaching its lowest point at $\gamma = 0.8$, after which it begins to rise again. This trend indicates that while increasing γ initially improves performance, excessively high values ($\gamma > 0.8$) may lead to a slight degradation in accuracy.

C. Performance Comparisons w.R.T. Loss Functions and Backbones

We assess the effectiveness of our proposed loss function by comparing it with L2, BL [6], NoiseCC [19], DM-count [31], and the generalized loss [56] using different backbones on the UCF-QNRF dataset. The results in Table IV demonstrate that our proposed scale-aware loss function consistently outperforms other state-of-the-art loss functions across various backbones. Recognizing the variability in human head sizes, our scale-aware approach effectively addresses scaling issues, an aspect not covered by NoiseCC. This leads to superior performance on the UCF-QNRF dataset compared to other loss functions. The comparisons on the SHANGHAITECH PART-A and PART-B datasets among different backbones are presented in Table V and Table VI, respectively, showcasing the effectiveness of our proposed loss function across different datasets. Table VII shows the accuracy comparisons among different loss functions and backbones on the JHU dataset. Clearly, our method outperforms other loss functions across various backbones.

In addition to the backbones mentioned above, STEERER [20] utilized another dense backbone, HRNet [57], for their performance evaluations. Table VIII shows comparisons between STEERER and our SACC-Net using the same backbones, VGG19 and HRNet. This table indicates that our architecture shows a clear advantage over STEERER across five datasets. Furthermore, applying our proposed loss function to STEERER results in a significant performance improvement. This finding further validates that our loss function not only enhances the effectiveness of our model but also demonstrates versatility by improving the accuracy of other models, such as STEERER.

D. Comparisons With SoTA Methods

To comprehensively evaluate the performance of our proposed method, we compare it against twenty state-of-the-art methods: CSRNet [3], CAN [23], S-DCNet [7], SANet [13], BL [6], SFANet [11], DM-Count [31], RPnet [21], AMSNet [58], M-SFANet [10], TEDnet [30], P2PNet [59], GauNet [27], MAN [60], HA-CCN [18], PaDNet [21], HMoDE+REL [18], ADM [61], MRL [15], and DKD [39]. Table IX presents the comparative results across five benchmark datasets. Our method consistently achieves the best MAE on all datasets, particularly excelling on large-scale datasets such as UCF-QNRF, NWPU-Crowd, and ShanghaiTech Part-A. In terms of the MSE metric, our method outperforms all state-of-the-art methods.

As shown in Table IX, VGG16 and VGG19 are two widely used backbones for evaluating the performance of most state-of-the-art (SoTA) methods. To ensure a fair comparison, we adopted them in this ablation study as well. Our SACC-net model consistently outperforms other models, particularly on the ShanghaiTech Parts A and B datasets. With the VGG16 backbone, our model performs competitively with ADM [61] on the UCF-CC50 dataset, even though ADM [61] employs a more advanced backbone. Using VGG19, our model achieves a notable performance advantage over other models, with the exception of SCPNet [42]. While SCPNet [42] attains a slightly lower MAE on UCF-CC50, due to its more sophis-

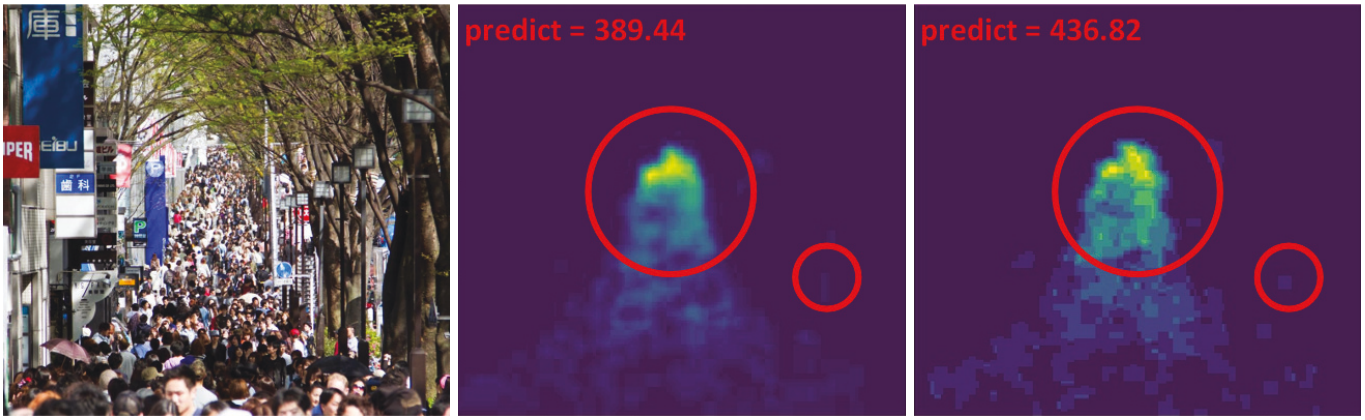


Fig. 11. Visualization of the SACC-Net crowd counting heatmap on ShanghaiTech Part-A: (a) Input image with ground-truth 429 heads. (b-c) show the generated heatmap, with (b) lacking IBF and (c) incorporating IBF. It is evident that (c) is visually and objectively superior to (b).

TABLE XI

COMPARISONS REGARDING THE PARAMETER SIZE, MAC AND FLOPS OF OUR SACC-LW MODEL WITH VGG16 AND CAN ON THE UCF-QNRF DATASET WITH INPUT DIMENSION 224 × 224

Methods	Parameters (M)	MAC (G)	FLOPS (G)
VGG16	7.89	15.47	7.73
CAN [23]	18.1	21.99	10.99
M-SFANet [10]	28.62	25.08	12.5
SFANet [11]	17	19.94	9.9
ADM [61]	16.14	13.05	6.82
MAN [60]	30.9	58.2	29.0
SACC-Net (Light)	**1.86**	**6.17**	**3.0**

TABLE XII

ABLATION STUDY ON THE IMPACT OF VARIOUS FUSION MODULES ON OUR MODEL WITH INPUT DIMENSION 512 × 512. SFM REFERS TO THE SYNTHETIC FUSION MODULE AND IFM TO THE INTRA-BLOCK FUSION MODULE

Methods	SFM	IFM	UCF-QNRF		S.H.Tech-A		S H.Tech-B		Params.
			MAE	MSE	MAE	MSE	MAE	MSE	
SACC-Net	✓	✓	73.91	121.70	52.19	76.63	6.16	9.71	51.24 M
	×	✓	81.47	132.58	54.30	83.71	6.22	9.85	44.04 M
	✓	×	82.81	137.62	54.83	90.39	6.28	9.93	32.61 M
	×	×	84.16	149.81	57.50	98.12	6.35	10.05	28.61 M

TABLE XIII

ABLATION STUDY OF THE SACC-NET RUNNING SFM+IFM AT DIFFERENT DENSITY SCALES ON UCF-QNRF

SFM+IFM	Scale1	Scale2	Scale3	UCF-QNRF	
				MAE	MSE
	✓			85.45	145.74
	✓	✓		84.07	135.63
	✓	✓	✓	82.42	130.04
	✓			83.81	140.19
✓	✓	✓		82.71	130.29
	✓	✓	✓	**73.91**	**121.7**

ticated HRNet backbone, our SACC-net achieves substantial improvements in the MSE metric across all cases.

E. Additional Experimental Results

Table X illustrates the efficiency comparisons among different backbones evaluated on a single 2080Ti GPU. Remarkably, with comparable accuracies, the efficiency of our lightweight version is double that of CAN [23], M-SFANet [10], and SFANet [11].

Table XI shows the ablation study for model parameter size, multiply accumulate (MAC), and floating point operations per second (FLOPS) among our light-weight architecture and other SoTA methods, respectively, where the input size is $3 \times 224 \times 224$. For fair comparisons, VGG16 is used as a baseline. The parameter size of our lightweight model is only *one-tenth* of other SoTA methods with comparable accuracies.

Visualization of Crowd Counting Heat Maps: To validate the effectiveness of our proposed loss function, Fig. 10 presents three visualization examples, demonstrating how our method generates a detailed heat map for counting small objects with enhanced accuracy in crowd counting. Ground truth head counts are shown in Fig. 10(a), while heat maps generated by MSE loss, Bayesian loss [6], and NoiseCC [19] are displayed in (b) to (d), respectively, along with their prediction results. Among these, the MSE loss function performs the worst. Bayesian loss [6] improves on MSE but does not address annotation errors. While NoiseCC [19] tackles annotation errors, it does not consider variations in head size due to distance and camera angles. In Fig. 10(e), the results predicted by our proposed loss function show a clear improvement in accuracy compared to other loss functions. The red boxes highlight our approach's effectiveness in capturing extremely small heads in distant regions.

F. Ablation Studies

We conducted ablation studies to analyze how the introduction of our synthetic and intra-block fusion approaches as well as the number of scales used in the process can affect crowd counting accuracy.

Impacts of SFM and IFM: Table XII presents the results of the ablation study on the effects of the synthetic and intra-block fusion approaches. It is evident that incorporating fusion modules significantly improves performance. Furthermore, IFM contributes more to counting accuracy improvement than SFM. However, the combination of both fusion modules results in the highest accuracy. For instance, our SACC-Net with these modules reduces error rates significantly from 84.16

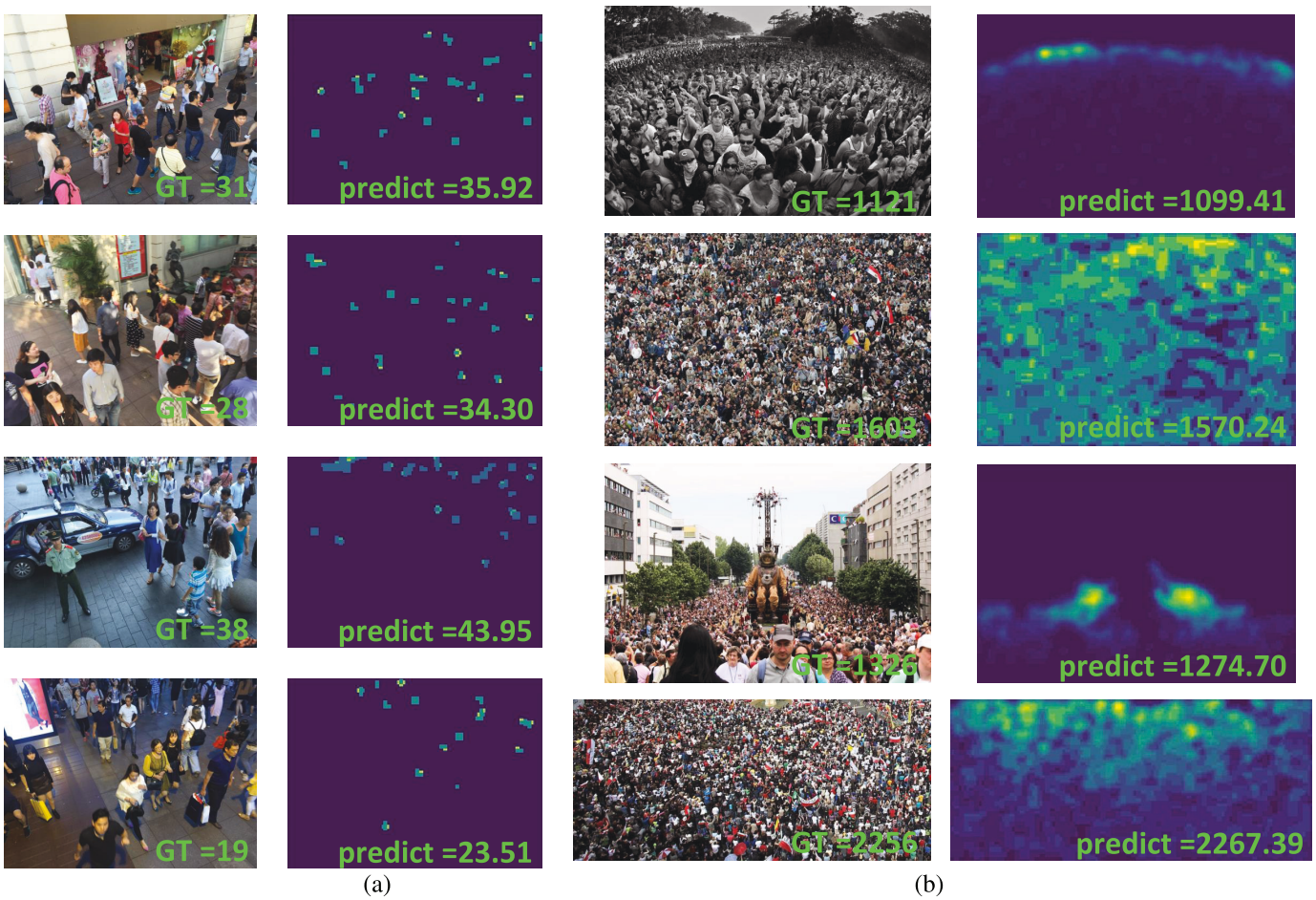


Fig. 12. Visualizations of crowd counting from extremely sparse and dense cases. (a) Sparse cases. (b) Dense cases.

to 73.91 in MAE and from 149.81 to 121.7 in MSE for the UCF-QNRF dataset.

Impact of Scale Numbers: We next evaluate the impact of the number of scales on enhancing crowd count accuracy. VGG19 incorporates five pooling layers, reducing the original image to a $1/32 \times 1.32$ ratio. The feature map in the last layer lacks sufficient information to calculate the required covariance matrix, and the first layer is too basic for crowd counting. Since three layers yield optimal performance, we set S to three in Eq. (3). Table XIII presents accuracy comparisons among three combinations of three scales (corresponding to layer 2, layer 3, and layer 4). The three-scale scale-aware loss function significantly enhances crowd-counting accuracy on the UCF-QNRF dataset, particularly in the MAE metric.

Scale-Aware Loss Function: Table XIV presents the ablation study of Bayesian loss and NoiseCC loss using our scale-aware loss function on SACC-Net, with an input dimension of 512×512 under various training epochs. We observe that both BL and NoiseCC perform better with our proposed scale-aware loss function compared to not using it. When the number of epochs increases, the MAE and MES metrics decrease more.

Fig. 11 provides visualizations generated by our method with and without the IFM module. The finer heat map details for smaller heads in (c) result in more accurate crowd counting, demonstrating the effectiveness of IFM. Furthermore, the SMF module synthesizes multiple layers to create better density maps for crowd counting. Finally, Fig. 12 illustrates the results of crowd counting under extremely sparse and dense conditions. Fig. 11(a) shows visualizations for sparse cases, while Fig. 11(b) displays those for dense cases. Even under these challenging conditions, our SACC-Net performs reliably, demonstrating its robustness.

Failure Cases of Crowd Counting: Backlighting often obscures or causes head features to disappear, creating challenges for crowd counting. Fig. 13 illustrates failure cases caused by backlighting conditions. Similarly, Fig. 14 presents

TABLE XIV

ABLATION STUDY ON THE BAYESIAN LOSS (BL) AND NOISECC LOSS (NOISECC) USING THE SCALE-AWARE LOSS (SAL) ON SACC-NET, WITH INPUT DIMENSION 512×512

Methods	SAL	Epoch	UCF-QNRF		S. H. Tech-A		S. H. Tech-B		UCF-CC50	
			MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
BL	✓	300	157.2	227.5	153.0	206.4	86.7	129.3	232.9	309.9
		400	145.4	208.9	137.8	181.3	71.2	109.9	198.6	273.5
		1000	107.2	154.3	97.1	138.7	16.1	37.8	166.4	243.5
	×	300	181.8	252.9	174.3	254.0	126.2	179.8	290.6	345.0
		400	169.8	239.4	162.8	238.6	119.7	158.9	263.4	317.8
		1000	138.7	193.1	128.6	186.3	31.4	52.7	226.1	273.1
NoiseCC	✓	300	142.9	187.1	134.8	190.1	69.2	99.9	219.6	279.4
		400	127.6	146.7	118.5	147.3	63.8	87.2	186.9	255.1
		1000	86.2	113.5	67.3	129.6	9.4	19.7	178.2	224.7
	×	300	164.7	229.9	157.5	238.8	96.3	150.7	284.8	310.4
		400	154.6	196.3	142.7	213.7	79.5	137.0	250.9	284.4
		1000	117.8	153.1	96.8	165.8	18.7	43.5	192.6	247.3

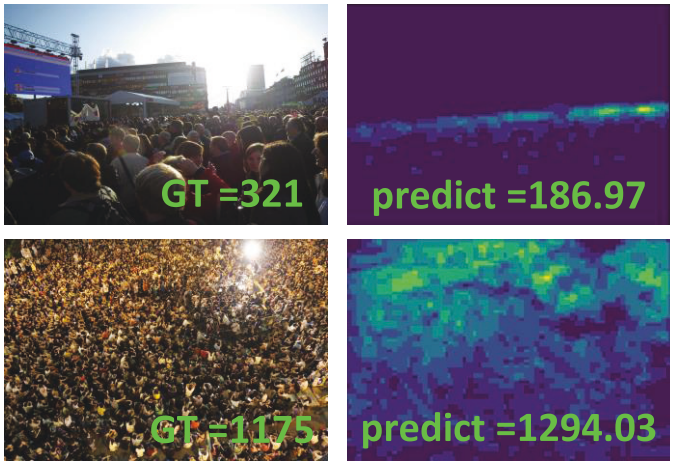


Fig. 13. Failure cases of crowd counting when the back-lighting was cast.

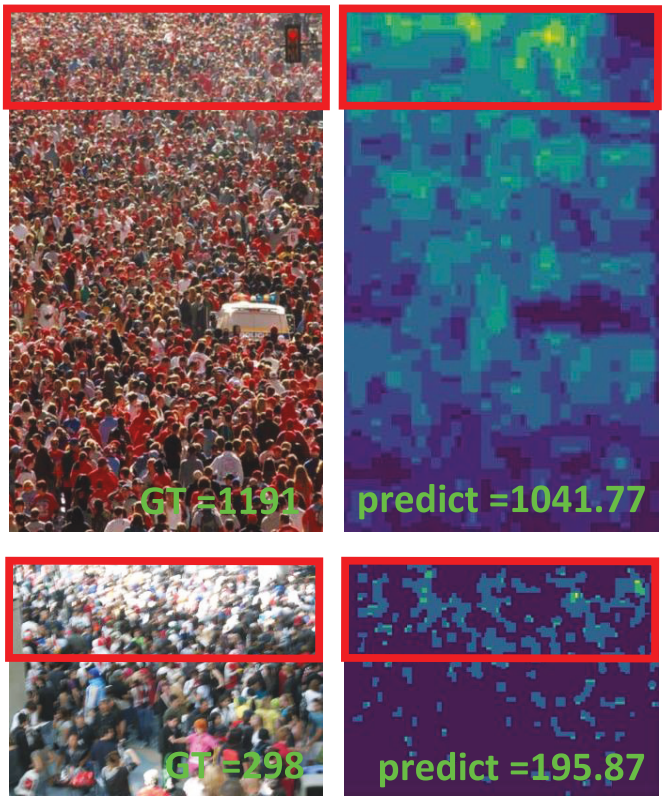


Fig. 14. Failure cases of due to blurred heads. The heads far from the cameras (denoted by red rectangle) were seriously blurred.

another type of failure due to blurring. When heads are far from the camera, their features become blurred, complicating extraction and leading to errors in crowd counting.

V. Conclusion

We presented a scale-aware crowd-counting network named SACC-Net, together with a new loss function that addresses the annotation noise *w.r.t.* scale for improving crowd counting. To overcome the scale truncation issue, our proposed SFM efficiently handles scale truncation problems, generating a smoother scale space for accurate counting of large objects. The IFM is developed to fuse feature layers within the same convolution block, enhancing information granularity for precise counting of small objects. The lightweight version of SACC-Net, SACC-LW, is both efficient and accurate. We evaluated the impacts of annotation variance α, the threshold γ, and annotation error variance β_1 on Mean Absolute Error (MAE) and Mean Squared Error (MSE) metrics. SACC-Net outperforms all SoTA methods on six datasets. Furthermore, using the same architecture, our scale-aware loss function surpasses several competing loss functions, including BL, NoiseCC, DM-count, and Gen-loss, utilized in existing methods.

Future Work: Future endeavors include exploring automatic parameter selection for α and β_s through data-driven learning. The underexplored domain of transfer learning or domain adaptation in crowd counting [62] presents an avenue for investigation. Further lightweight enhancements can enable the deployment of SCAA-lw for direct operation on drones. Addressing the challenge of crowd counting under adversarial conditions, such as inclement weather [63], represents an intriguing extension. Additionally, extending this line of research from human-centric crowd counting to the automatic counting of other visually similar objects (*e.g.*, vehicles, fruits, fishes, birds) would contribute to the broader field of image processing and computer vision. Generalizing the visual counting problem from a human-centric to a nature-centric context remains an unexplored research area with promising prospects. While our approach effectively addresses annotation errors, it still faces challenges with severe blurring and backlighting, which are common in real-world environments. In the future, we plan to incorporate noise reduction and diffusion techniques to enhance image clarity. Additionally, exploring scale and illumination invariance in crowd counting presents another promising direction for further research.

Acknowledgment

The authors would like to thank National Center for High-performance Computing (NCHC) for providing computational and storage.

References

[1] B. Li, H. Huang, A. Zhang, P. Liu, and C. Liu, "Approaches on crowd counting and density estimation: A review," *Pattern Anal. Appl.*, vol. 24, no. 3, pp. 853–874, Aug. 2021.

[2] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, "CNN-based density estimation and crowd counting: A survey," 2020, *arXiv:2003.12783*.

[3] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100.

[4] C. Xu et al., "AutoScale: Learning to scale for crowd counting and localization," 2019, *arXiv:1912.09632*.

[5] S. Bai, Z. He, Y. Qiao, H. Hu, W. Wu, and J. Yan, "Adaptive dilated network with self-correction supervision for counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4594–4603.

[6] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, South Korea, Oct. 2019, pp. 6142–6151.

[7] H. Xiong, H. Lu, C. Liu, L. Liu, Z. Cao, and C. Shen, "From open set to closed set: Counting objects by spatial divide-and-conquer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8362–8371.

[8] R. Rama Varior, B. Shuai, J. Tighe, and D. Modolo, "Multi-scale attention network for crowd counting," 2019, *arXiv:1901.06026*.

[9] X. Jiang et al., "Attention scaling for crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4706–4715.

[10] P. Thanasutives, K.-I. Fukui, M. Numao, and B. Kijsirikul, "Encoder–decoder based convolutional neural networks with multi-scale-aware modules for crowd counting," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2382–2389.

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

[12] J. Wan and A. Chan, "Adaptive density map generation for crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1130–1139.

[13] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[14] M. A. Khan, H. Menouar, and R. Hamila, "Revisiting crowd counting: State-of-the-art, trends, and future perspectives," *Image Vis. Comput.*, vol. 129, Jan. 2023, Art. no. 104597.

[15] X. Wei, Y. Qiu, Z. Ma, X. Hong, and Y. Gong, "Semi-supervised crowd counting via multiple representation learning," *IEEE Trans. Image Process.*, vol. 32, pp. 5220–5230, 2023.

[16] Y. Tian, Y. Lei, J. Zhang, and J. Z. Wang, "PaDNet: Pan-density crowd counting," *IEEE Trans. Image Process.*, vol. 29, pp. 2714–2727, 2020.

[17] V. A. Sindagi and V. M. Patel, "HA-CCN: Hierarchical attention-based crowd counting network," *IEEE Trans. Image Process.*, vol. 29, pp. 323–335, 2020.

[18] Z. Du, M. Shi, J. Deng, and S. Zafeiriou, "Redesigning multi-scale neural network for crowd counting," *IEEE Trans. Image Process.*, vol. 32, pp. 3664–3678, 2023.

[19] J. Wan and A. B. Chan, "Modeling noisy annotations for crowd counting," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 3386–3396.

[20] T. Han, L. Bai, L. Liu, and W. Ouyang, "STEERER: Resolving scale variations for counting and localization via selective inheritance learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 21791–21802.

[21] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 833–841.

[22] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 589–597.

[23] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5099–5108.

[24] Y. Miao, Z. Lin, G. Ding, and J. Han, "Shallow feature based dense attention network for crowd counting," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, 2020, pp. 11765–11772.

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[26] S. Huang et al., "Body structure aware deep crowd counting," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1049–1059, Mar. 2018.

[27] Z.-Q. Cheng, Q. Dai, H. Li, J. Song, X. Wu, and A. G. Hauptmann, "Rethinking spatial invariance of convolutional networks for object counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 19638–19648.

[28] J. Cheng, H. Xiong, Z. Cao, and H. Lu, "Decoupled two-stage crowd counting and beyond," *IEEE Trans. Image Process.*, vol. 30, pp. 2862–2875, 2021.

[29] M. Wang, H. Cai, Y. Dai, and M. Gong, "Dynamic mixture of counter network for location-agnostic crowd counting," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 167–177.

[30] X. Jiang et al., "Crowd counting and density estimation by trellis encoder–decoder networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6133–6142.

[31] B. Wang, H. Liu, D. Samaras, and M. Hoai, "Distribution matching for crowd counting," 2020, *arXiv:2009.13077*.

[32] M. Wang, J. Zhou, H. Cai, and M. Gong, "CrowdMLP: Weakly-supervised crowd counting via multi-granularity MLP," *Pattern Recognit.*, vol. 144, Jul. 2023, Art. no. 109830.

[33] T. Wang, T. Zhang, K. Zhang, H. Wang, M. Li, and J. Lu, "Context attention fusion network for crowd counting," *Knowledge-Based Syst.*, vol. 271, Jul. 2023, Art. no. 110541.

[34] W. Zhai et al., "DA^{2}net: A dual attention-aware network for robust crowd counting," *Multimedia Syst.*, vol. 29, no. 5, pp. 3027–3040, Oct. 2023.

[35] W. Zhai, M. Gao, Q. Li, G. Jeon, and M. Anisetti, "FPANet: Feature pyramid attention network for crowd counting," *Int. J. Speech Technol.*, vol. 53, no. 16, pp. 19199–19216, Aug. 2023.

[36] S. S. Savner and V. Kanhangad, "CrowdFormer: Weakly-supervised crowd counting with improved generalizability," *J. Vis. Commun. Image Represent.*, vol. 94, Jun. 2023, Art. no. 103853.

[37] B. Li, Y. Zhang, H. Xu, and B. Yin, "CCST: Crowd counting with Swin transformer," *Vis. Comput.*, vol. 39, no. 7, pp. 2671–2682, Jul. 2023.

[38] D. Liang, J. Xie, Z. Zou, X. Ye, W. Xu, and X. Bai, "CrowdClip: Unsupervised crowd counting via vision-language model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 2893–2903.

[39] R. Wang et al., "Efficient crowd counting via dual knowledge distillation," *IEEE Trans. Image Process.*, vol. 33, pp. 569–583, 2024.

[40] X. Guo, K. Song, M. Gao, W. Zhai, Q. Li, and G. Jeon, "Crowd counting in smart city via lightweight ghost attention pyramid network," *Future Gener. Comput. Syst.*, vol. 147, pp. 328–338, Oct. 2023.

[41] W. Zhai et al., "An attentive hierarchy ConvNet for crowd counting in smart city," *Cluster Comput.*, vol. 26, no. 2, pp. 1099–1111, 2023.

[42] W. Zhai, M. Gao, X. Guo, Q. Li, and G. Jeon, "Scale-context perceptive network for crowd counting and localization in smart city system," *IEEE Internet Things J.*, vol. 10, no. 21, pp. 18930–18940, Nov. 2023.

[43] X. Guo, M. Gao, W. Zhai, Q. Li, and G. Jeon, "Scale region recognition network for object counting in intelligent transportation system," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 15920–15929, Dec. 2023.

[44] X. Guo, M. Gao, G. Zou, A. Bruno, A. Chehri, and G. Jeon, "Object counting via group and graph attention network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 9, pp. 11884–11895, Sep. 2024.

[45] W. Zhai, M. Gao, M. Anisetti, Q. Li, S. Jeon, and J. Pan, "Group-split attention network for crowd counting," *J. Electron. Imag.*, vol. 31, no. 4, Jun. 2022, Art. no. 041214.

[46] J. Chen, Q. Li, M. Gao, W. Zhai, G. Jeon, and D. Camacho, "Towards zero-shot object counting via deep spatial prior cross-modality fusion," *Inf. Fusion*, vol. 111, Nov. 2024, Art. no. 102537.

[47] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. NIPS*, 2010, pp. 1324–1332.

[48] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1879–1888.

[49] J. Wan, N. S. Kumar, and A. B. Chan, "Fine-grained crowd counting," *IEEE Trans. Image Process.*, vol. 30, pp. 2114–2126, 2021.

[50] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[51] H. Idrees et al., "Composition loss for counting, density map estimation and localization in dense crowds," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 532–546.

[52] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2547–2554.

[53] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-crowd: A large-scale benchmark for crowd counting and localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2141–2149, Jun. 2021.

[54] V. Sindagi, R. Yasarla, and V. Patel, "Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1221–1231.

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[56] J. Wan, Z. Liu, and A. B. Chan, "A generalized loss function for crowd counting and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1974–1983.

[57] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2020.

[58] Y. Hu et al., "NAS-count: Counting-by-density with neural architecture search," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2020, pp. 747–766.

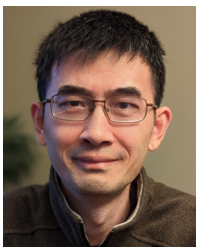
[59] Q. Song et al., "Rethinking counting and localization in crowds: A purely point-based framework," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3365–3374.

[60] H. Lin, Z. Ma, R. Ji, Y. Wang, and X. Hong, "Boosting crowd counting via multifaceted attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19628–19637.

[61] Z. Zhao and X. Li, "Deformable density estimation via adaptive representation," *IEEE Trans. Image Process.*, vol. 32, pp. 1134–1144, 2023.
[62] D. Khan and I. W.-H. Ho, "CrossCount: Efficient device-free crowd counting by leveraging transfer learning," *IEEE Internet Things J.*, vol. 10, no. 5, pp. 4049–4058, Mar. 2023.
[63] Z.-K. Huang, W.-T. Chen, Y.-C. Chiang, S.-Y. Kuo, and M.-H. Yang, "Counting crowds in bad weather," 2023, *arXiv:2306.01209*.

Yi-Kuan Hsieh is currently pursuing the Ph.D. degree in artificial intelligence with National Yang Ming Chiao Tung University (NYCU), Taiwan. His research interests include low-power computer vision, image processing, pattern recognition, and deep learning. His recent work focuses on small object counting, such as shrimp larvae estimation, few-shot learning, change detection, EfficientViT, and token pruning. His related research has been published in top conferences and journals, including AAAI, IEEE ICIP, and IEEE INTERNET OF THINGS JOURNAL.

Jun-Wei Hsieh (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the National Central University, Taiwan, in 1995. He was an Associate Professor with the Department of Electrical Engineering, Yuan Ze University, and a Visiting Researcher with the MIT AI Laboratory. Since August 2009, he has been a Professor and the Dean of the Department of Computer Engineering, National Taiwan Ocean University. After August 2019, he has been a Professor with the College of AI, National Yang Ming Chiao Tung University. He hosted or co-hosted many large-scale AI projects from different companies and governments in the past. He has many successful experiences in industrial-academic cooperation and technology transferring, especially in ITS. He has authored more than 150 peer-reviewed journal and conference publications and 20 U.S./Taiwan patents. His research interests include AI, deep learning, smart farming, video surveillance, intelligent transportation systems, image and video processing, object recognition, machine learning, 3D printing, medical image analysis, and computer vision. In May 2019, he received the First Prize of the Ministry of Science and Technology Best Display Award, and the Third Place of the AI Investment Potential Award. Due to his contributions in traffic flow estimation, he helped Elan Company to receive the Gold Award from Taipei International Computer Show, in 2019. He also received the Outstanding Research Award of National Taiwan Ocean University, in 2012, 2016, 2017, and 2019, and the Outstanding Research Award of Yuan Ze University, in 2006, 2007, and 2008. He and his students received the Silver Medal of the 2019 National College Software Creation Competition, the Silver Medal of 2018 National Microcomputer Competition, the Best Paper Award of Information Technology and Applications in Outlying Islands Conference, in 2013, 2014, 2016, 2017, 2018, 2021, and 2022, respectively, the Best Paper Award of Tanet 2017, the Best Paper Award of NCWIA 2020, 2021, and 2022, respectively, and the Best Paper Awards of IS3C 2020. He also received the Best Paper Award of CVGIP Conference, in 1999, 2003, 2005, 2007, 2014, 2017, 2018, and 2022, the Best Paper Award of DMS Conference, in 2011, the Best Paper Award of IIHMSP 2010, and the Best Patent Award of Institute of Industrial Technology Research, in 2009 and 2010, respectively. He serves as the Program Chair for the Conference on Multimedia Modeling 2011 and the IEEE Advanced Video and Signal-Based Surveillance (AVSS) 2019.



Xin Li (Fellow, IEEE) received the B.S. degree from the University of Science and Technology of China, Hefei, in 1996, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2000. He was a Member of Technical Staff with Sharp Laboratories of America, Camas, WA, USA, from August 2000 to December 2002. From January 2003 to August 2023, he was a Faculty Member with the Lane Department of Computer Science and Electrical Engineering, West Virginia University. In 2023, he joined the Department of Computer Science, University at Albany. His current research interests include computer vision, NeuroAI, multimodal AI, foundation models, and autism research. He was elected as a fellow of IEEE in 2017 for his contributions to image interpolation, restoration, and compression.

Yu-Ming Zhang is currently pursuing the Ph.D. degree in computer science with National Central University (NCU), Taoyuan, Taiwan. His research interests include low-power object detection and efficient neural architecture search, with related work published in NeurIPS, IEEE ICIP, IEEE ISCAS, IEEE AVSS, and other international conferences. He achieved First Place in the Hackathon organized by Taiwan's Ministry of Education in 2017, along with the Office Award with Taiwan AI CUP the same year. In 2020, he earned the Bronze Award in Kaggle's M5 competition. He has also been recognized with the Best Paper Award with the NCWIA Conference in both 2022 and 2023, and at TANET 2023 and ITAOI 2024. He served as a reviewer for BMVC 2023.

Yu-Chee Tseng (Fellow, IEEE) received the B.S. degree in computer science from National Taiwan University, Taipei. Taiwan, in 1985, the M.S. degree in computer science from National Tsing-Hua University, Hsinchu, Taiwan, in 1987, and the Ph.D. degree in computer and information science from Ohio State University, in January 1994. He was the Chairperson, from 2005 to 2009, and the Dean, from 2011 to 2017, with the College of Computer Science, National Chiao Tung University (NCTU), Taiwan. He has been the NCTU Chair Professor, since 2011, and the Y. Z. Hsu Scientific Chair Professor, from 2012 to 2013. In 2021, NCTU merged with Yang Ming University and was subsequently renamed National Yang Ming Chiao Tung University (NYCU). He is currently the lifetime Chair Professor of NYCU. His H-index is more than 60. His research interests include mobile computing, wireless communication, and the Internet of Things. He received the Outstanding Research Award (National Science Council, in 2001, 2003, and 2009), the Academic Award (Ministry of Education), the Best Paper Awards (ICPP 2003, iThings 2014, APNOMS 2015, and IoTaaS 2017), the Elite I. T. Award, in 2004, the Distinguished Alumnus Award (The Ohio State University, in 2005), the Y. Z. Hsu Scientific Paper Award, in 2009, the TWAS Prize, in 2018, and the National Chair Professorship, from 2020 to 2023. He served/serves on the editorial boards for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, and IEEE INTERNET OF THINGS JOURNAL.

Ming-Ching Chang (Senior Member, IEEE) is currently an Associate Professor and the Co-Director of the CVML Laboratory, Department of Computer Science, College of Nanotechnology, Science, and Engineering (CNSE), University at Albany, State University of New York (SUNY). He has authored more than 157 peer-reviewed journal and conference publications, seven U.S. patents and 15 disclosures. His research has been founded by DARPA, IARPA, NIJ, VA, GE Global Research, and Inventec Corporation. He has rich experience in leveraging expertise from multiple domains to accomplish multi-discipline programs and projects. He receives multiple paper awards from international conferences, including the ECCV 2024 Beyond Enclidean Hyperbolic and Hyperspherical Learning for Computer Vision Workshop Best Paper Award, the IEEE MIPR 2023 Best Student Paper Award, the AI City Challenge 2017 Honorary Mention Award, the IEEE WACV 2012 Best Student Paper Award, and the IEEE AVSS 2011 Best Paper Award—Runner-Up. He frequently serves as the program chair, the area chair, and a referee for leading journals and conferences. He is the core organizer of the AI City Challenge, a multi-year (2017–2023) IEEE CVPR Workshops. He serves as the Program Co-Chair for the IEEE ICME 2025 Conference, the General Chair (2025) and the Program Chair (2024 and 2019) for the IEEE AVSS Conference, and the General Chair (2024) and the TPC Chair Lead (2022) for the IEEE MIPR Conference. He was the Area Chair of the IEEE ICIP conferences (2017 and 2019–2025) and the Outstanding Area Chair of the ICME 2021 Conference. He chairs the steering committee of the IEEE AVSS Conference since 2022. He is an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA.