Size-Adaptive Density Estimation for Accurate Object Counting in Dense Environments

Phu Nguyen*, An Nguyen Duc†, Bao Bui Quoc*
*Department of Electronics, Ho Chi Minh University of Technology (HCMUT), Vietnam National University-Ho Chi Minh City (VNU-HCM), Ho Chi Minh City, Vietnam
†..., ..., ...

***Abstract*—Object counting in images with significant size variations and dense object clusters remains a challenge in computer vision. Existing methods often struggle with accuracy due to limitations in handling these complexities. This paper proposes an approach based on a Size-Adjustable Gaussian Kernel for density map generation, addressing the issue of object size heterogeneity. Our method dynamically adjusts the kernel parameters based on individual object dimensions, ensuring accurate representation of both spatial distribution and size variations. We utilize a modified VGG16 architecture with a Bi-directional Feature Pyramid Network (BiFPN [1]) and attention mechanisms (ECA [2] and MSCA [3]) for robust feature extraction and object recognition across scales. Extensive experiments on a challenging shrimp larvae image dataset demonstrate the superior performance of our method, achieving significant improvements in counting accuracy compared to state-of-the-art approaches. This work paves the way for more precise and reliable object counting in various applications with diverse object sizes and densities.**

***Index Terms*—Object counting, density estimation, size-adjustable Gaussian kernel, deep learning, convolutional neural networks, attention mechanisms, shrimp larvae, aquaculture.**

I. Introduction

Object counting within images forms a crucial task in various domains, from crowd analysis and traffic monitoring to ecological surveys and industrial quality control. However, achieving accurate counts, especially when objects exhibit significant size variations and appear in dense clusters, presents a persistent challenge in computer vision. Existing methods often fall short due to inherent limitations in handling these complexities.

Detection-based approaches [5]–[12], while effective for well-defined and separated objects, struggle with accuracy in dense scenarios where objects overlap. The reliance on precise bounding boxes becomes problematic, leading to miscounts and inaccurate localization. Similarly, regression-based methods [13]–[18], while offering a direct mapping from image features to object counts, often overlook crucial spatial information, resulting in a loss of detail regarding object distribution and arrangement within the scene.

Density map-based methods have emerged as a promising alternative, leveraging spatial information to generate density maps that reflect the distribution of objects within an image. However, conventional approaches often employ fixed Gaussian filters with constant kernel size and variance, neglecting the inherent size variations among objects. This leads to suboptimal performance, particularly when dealing with populations exhibiting diverse sizes. While some methods attempt to address this by using adaptive Gaussian filters based on average distances between neighboring objects, they still lack the precision achieved by incorporating individual object size information. To overcome these limitations, this paper presents a novel approach that tackles the challenges of object size variation and density head-on. Our proposed method introduces a Size-Adjustable Gaussian Kernel for generating density maps. This kernel dynamically adjusts its parameters based on the individual dimensions of each object, ensuring that the generated density map accurately reflects both the spatial distribution and the size variations within the object population.

We leverage the power of Convolutional Neural Networks (CNNs) for feature extraction, employing a modified VGG16 architecture as the backbone of our model. This architecture is further enhanced by the integration of a Bi-directional Feature Pyramid Network (BiFPN [1]) to effectively capture multi-scale features, crucial for recognizing objects of varying sizes. Additionally, we incorporate attention mechanisms, such as Efficient Channel Attention (ECA [2]) and Multi-Scale Conv Attention (MSCA [2]), to refine the model's focus on relevant features and improve object recognition across different scales. This comprehensive approach ensures that our model effectively captures the intricacies of object size and distribution within the image.

We validate the effectiveness of our proposed method through extensive experiments on a challenging dataset of shrimp larvae images. The inherent size heterogeneity within this population, coupled with the dense clustering of individuals, makes it an ideal testbed for evaluating our approach. The results demonstrate that incorporating object size information into the density map generation process leads to significant improvements in counting accuracy compared to existing methods. Our Size-Adjustable Gaussian Kernel, in conjunction with the powerful feature extraction and attention mechanisms of our CNN architecture, effectively addresses the challenges posed by diverse object sizes and densities, paving the way for more precise and reliable object counting in various applications.

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.
Manuscript received April 19, 2021; revised August 16, 2021.

II. Related Works

A. Density Map Generation

Counting the number of objects in an environment where objects are in dense and overlap each other remains a topic that has been extensively researched, and there are numerous positive solutions proposed to address it. According to Loy *et al*. [4], they have categorized three primary approaches for tackling this issue: detection-based methods [5]–[12], regression-based methods [13]–[18], and density estimation-based methods [19]–[31].

Primarily, detection-based methodologies stand out in the quest for object localization within images. These approaches leverage sophisticated object detection frameworks, with a prevailing emphasis on their deployment in early researches. The efficacy of these methods hinges on meticulously trained detection models capable of discerning and extracting comprehensive object features. Prominent instances include the YOLO (You Only Look Once) model [5]–[7], as well as the R-CNN (Region-based Convolutional Neural Network) model [10]–[12]. These models, meticulously trained to identify objects, employ bounding boxes to enumerate objects in images. Nevertheless, their precision exhibits a proclivity to diminish in scenarios where objects are densely clustered or overlap. This phenomenon persists despite the models' adeptness in assimilating information about the objects. As a result, the effectiveness of this approach is most pronounced when the objects to be counted display clear demarcations, introducing challenges in accurately determining bounding box coordinates, especially within densely populated and overlapping object scenarios. [32].

To overcome the constraints inherent in detection-based methodologies, regression-based approaches provide a direct mapping from crowd images to numerical object quantities by learning the relationship between extracted features and the actual object count. Despite their success in mitigating challenges related to congestion and overlap, these methods tend to neglect the spatial distribution of objects—an essential factor for precise counting.

The density map-based approach strategically exploits spatial information to formulate density maps for individual training samples, facilitating the computation of the final object count through summation across the density map. This methodology effectively addresses the limitations inherent in regression-based techniques by preserving pertinent spatial details of enumerated objects. Pioneering this paradigm, Lempitsky *et al.* advocate for the conversion of point annotations into density maps via Gaussian kernel filters [33]. Their subsequent model training involves labels with the density maps generated from these Gaussian filters. The density map-based approach garnering widespread adoption due to its promising outcomes. However, the judicious selection of Gaussian filter parameters, such as kernel size and variance, remains a pivotal consideration.

While conventional density map methods [19], [33] often apply a fixed Gaussian filter with the same kernel size and variance parameters across the entire input image, leading to a loss of object size information, advanced strategies strive to overcome this limitation. They employ adaptive Gaussian filters [20], [22]–[28], [30] that dynamically adjust based on the average distance between neighboring objects, thereby mimicking object sizes within intricate geometric contexts. Nevertheless, these approaches still compromise on the accuracy achieved due to the lack of object size information. Remarkably, models trained using these methods employ loss functions that supervise pixel-to-pixel transitions from input images to density map outputs, emphasizing the crucial role in producing high-quality density maps for optimal performance. Thus, in addition to preservation of spatial information remains a hallmark advantage of density map-based methodologies, the integration of object size information is necessary for crafting superior density maps, thereby augmenting overall model efficacy.

B. CNN-based methods

Convolutional Neural Networks (CNNs) have emerged as a powerful paradigm, showcasing remarkable efficacy across diverse computer vision applications. Among these, the task of crowd counting in images or videos has particularly benefited from CNN-based approaches. Various strategies within the CNN framework have been explored, including but not limited to multi-column CNN [20], [22], [23], [27], Dilated CNN [24], [34]–[36], and Pyramid Pooling [37]–[39], all geared towards optimizing the extraction and synthesis of input features.

One notable advancement in the realm of multi-column CNNs is the MCNN (Multi-column CNN) model proposed by Zhang *et al.* [22]. This innovative model excels in its ability to extract a spectrum of diverse data features at different scales. Each column within the architecture is adept at learning distinct features, fostering a comprehensive synthesis of information. However, the inherent complexity introduced by employing multiple columns, each dedicated to extracting diverse data, poses a challenge.

In the dilated CNN approach, the CSRNet [24] model has made significant strides by incorporating dilated kernels in various branches. This strategic integration allows for the creation of larger receptive fields, enabling the network to capture more extensive contextual information without imposing an undue burden on system resources. Another noteworthy approach is exemplified by FPANet model (Zhai *et al.*, 2023) [39], which strategically amalgamates multi-scale features from a pyramid architecture, effectively addressing challenges arising from scale variations. The amalgamation of multi-column designs, multi-branch CNN architectures, and pyramid structures in these methods underscores their efficiency in feature extraction across diverse hierarchical levels. Building upon these precedents, our research endeavors to integrate the BiFPN (Bi-directional Feature Pyramid Network) [1] architecture into our neural network. This augmentation is poised to introduce a nuanced, multi-level contextual understanding, specifically tailored to tackle challenges associated with object scale variations—a focal point within the purview of our investigation.

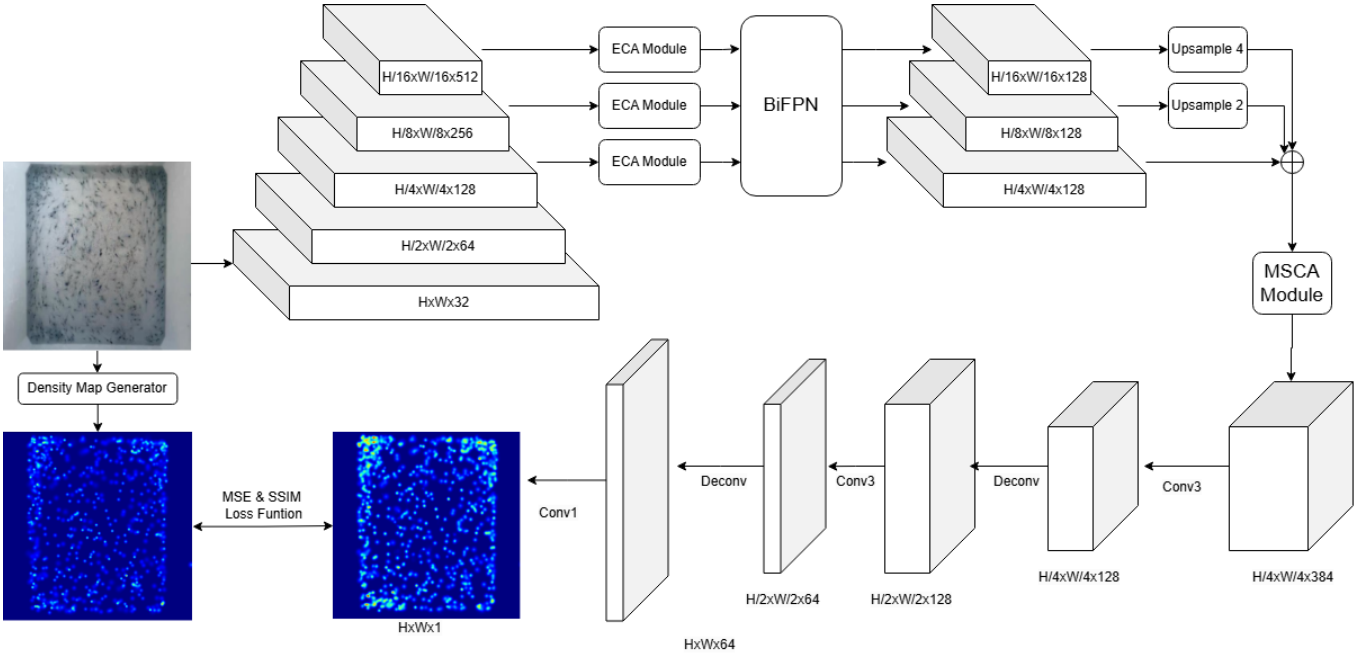

Fig. 1. **Proposed Model**: Modified VGG16 [40] as backbone and BiFPN [1] to connect feature maps

III. Proposed Method

A. Bi-directional Multi-scale Feature Pyramid Network Architecture (BMFPN)

Addressing the crowd counting challenge, we employ VGG16 as the backbone of our model. The combination of Efficient Channel Attention (ECA) and Bidirectional Feature Pyramid Network (BiFPN) elevates deep feature representation and attention focus, effectively handling intricate scenarios in crowded settings. Enhancing adaptability across scales, the model undergoes additional refinement with Multi-Scale Conv Attention (MSCA). This novel fusion results in a potent and versatile model, positioned to reshape the landscape of crowd analysis in real-world situations.

*1) **Feature Map Extraction (FME)**:* We opted to employ VGG16 [40] as the backbone for our model, leveraging its strong feature extraction capabilities from images. VGG16, with its deep-layered architecture, consistently exhibits superior performance across various computer vision tasks. The flexibility of VGG16, especially in intermediate and deep layers, makes it an ideal candidate for addressing challenges related to complex feature extraction in crowd counting. However, due to the substantial size of VGG16's architecture, requiring significant memory and computational resources for training, we strategically decided to halve the number of channels in each VGG16 layer. Subsequent testing and evaluation confirmed that this adjustment had no significant impact on the model's ability to detect and extract features from crowds. Instead, it resulted in a lighter and faster model.

Layer 3, with high resolution, provides detailed information and small features, facilitating the detection of small and distant objects. Its primary function is to identify fine details and integrate them to create an overall description of the crowd. Layer 4, situated in the middle, optimally combines detailed and generalized information about the crowd, enhancing understanding of its structure, identifying medium-sized objects, and presenting a general representation of object relationships. Layer 5, characterized by low resolution, contributes to gathering overall information about the crowd and its prominent features, aiding the model in identifying important features related to the distribution and relationships between large object groups. Thus, in this study, we use layers 3, 4, and 5 as a feature map for our model.

Exploiting the features from layers 3, 4, and 5 of the VGG16 model with feature representation capabilities at different size levels, we recognized the necessity of integrating multiple attention modules, such as the Efficient Channel Attention (ECA) module, into our architecture. This integration serves several critical purposes.

Firstly, attention mechanisms enhance the model's capability to discern and prioritize relevant features, particularly important in complex tasks like crowd counting. The architecture of ECA focuses on efficiently capturing important features within each channel. It operates by computing attention weights for each channel independently, thus reducing computational overhead compared to methods that involve interactions between spatial and channel dimensions. ECA achieves this by utilizing a lightweight mechanism that applies a 1D convolutional operation along the channel dimension, followed by a sigmoid activation to generate attention weights.

Secondly, the combination of different attention mechanisms allows us to leverage their complementary strengths and mitigate individual weaknesses. In our architecture, the integration of ECA with Bidirectional Feature Pyramid Network (BiFPN) provides a powerful synergy. While BiFPN constructs a multi-level pyramid system to consolidate information from different layers, ECA optimizes attention within each channel,

ensuring that the model focuses on crucial features at every stage of the feature pyramid.

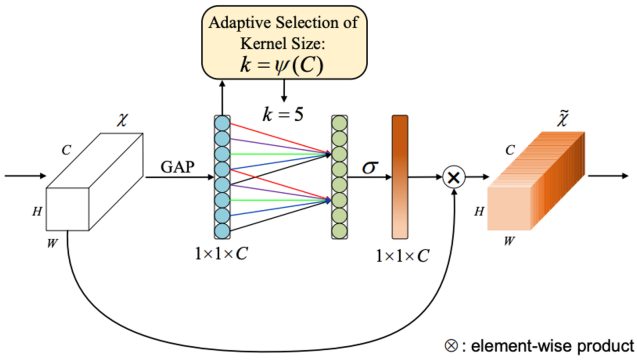

Fig. 2. Efficient channel attention (ECA) module

This integration not only enhances the model's ability to capture and utilize multi-scale features effectively but also optimizes the allocation of computational resources. The resulting synergy between attention mechanisms enhances the model's capacity to discriminate and count objects in crowds, making it suitable for diverse scenarios and challenging real-world conditions. By applying an ECA block to each of the layers 3, 4, and 5, we create a robust, flexible, and efficient crowd counting model that excels in various scenarios.

Following the synthesis of features using Bidirectional Feature Pyramid Network (BiFPN), we proceed by transmitting information through the Multi-Scale Conv Attention (MSCA) module. The MSCA module plays a critical role in enhancing the attention mechanism within the model, particularly focusing on features at different scales within the image space. This module consists of convolutional layers that operate at multiple scales simultaneously, enabling the model to capture features ranging from fine details to larger-scale structures within the crowd.

By incorporating MSCA into the model architecture, we ensure that the model optimally allocates attention across different scales, thereby improving its accuracy in crowd counting tasks. This module not only enhances the model's capability to recognize objects at various scales but also complements the multi-level nature of the BiFPN, further refining the model's feature representation and attention mechanism across the entire image.

The combination of BiFPN, MSCA, and ECA creates a multi-directional, powerful, and flexible model architecture. This combination not only enhances representational capabilities but also optimizes attention and object recognition at every detail and scale in the image, making the model particularly suitable for the crucial task of crowd counting.

*2) **Density Map Estimator (DME)**:* Following the aggregation and synthesis of essential features from various VGG layers by the Bi-directional Feature Pyramid Network (BiFPN), our focus shifts to the Density Map Estimator (DME) layer—a pivotal element influencing the accuracy of the density map. Structuring the DME layer is as shown in Figure 1, not only showcases flexibility but also maximizes the utilization of information from BiFPN's output layers, rendering the crowd counting process more robust and versatile than ever.

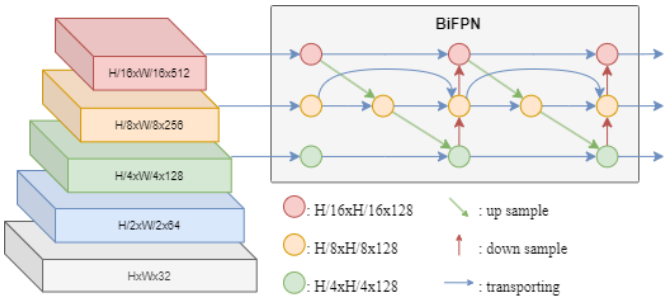

Fig. 3. Bi-directional Feature Pyramid Network

Connencing with Layer 3 of BiFPN, this layer possesses dimensions H/16 x W/16 (with HxW representing the input image's size). Within each layer, the process encompasses resizing through transposed convolution, information optimization via 3x3 convolutions, and concludes with ReLU activation. Each layer's output comprises 128 channels, concurrently concatenated with the lower BiFPN layer through concatenate to ensure tight interaction and efficiency. Specifically, transposed convolution aims to efficiently restore the size to the original input image dimensions, while the 3x3 convolutional layer adjusts each detail, creating a diverse information space with unique features. This iterative process continues through each subsequent BiFPN layer, progressing from the highest to the lowest. The meticulous tuning of each layer, coupled with astute concatenation from higher to lower layers, enhances the model's flexibility and establishes bidirectional interaction between layers, facilitating the creation of an effective density map amidst crowded objects of varying sizes.

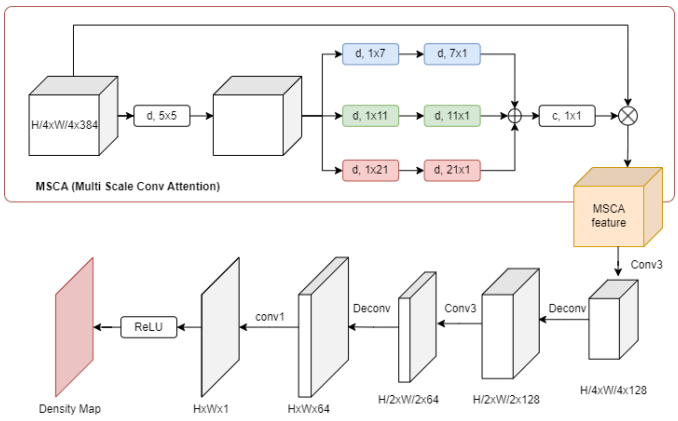

Fig. 4. Density Map Estimation Module

Following three consecutive layers, the outcome is a feature map with 256 channels, and dimensions H/4 x W/4. To match the input image size (HxW) and reduce channels to 1, we undertake two iterations of transposed convolution and convolution, as depicted in Figure 1. Ultimately, the ReLU activation function is applied after the final convolutional layer, as density map values are invariably positive. Consequently, DME generates high-resolution density maps mirroring the

input size by amalgamating essential information from the BiFPN block.

B. Density Map Generation

In a densely populated and crowded scene, each object is represented by three key annotations: c_i (indicating the position), h_i (representing the height), and w_i (indicating the width) for each entity. Each of these parameters holds specific significance in describing the spatial attributes of the objects within the crowd context. The parameter c_i assumes a pivotal role in the accurate determination of the spatial location of each object. It serves as a key factor in establishing the precise coordinates of an object within the given space, enabling a detailed understanding of its position amidst the crowded surroundings. Simultaneously, the values of w_i and h_i contribute to furnishing a holistic perspective on the dimensions of the object. These dimensions collectively provide comprehensive insights into the size and spatial extent of the entities under consideration. Analogous to widely adopted bounding box methodologies in advanced object detection approaches, the combination of w_i and h_i acts as a robust descriptor, encapsulating the object's physical characteristics and facilitating a more nuanced representation within the research framework.

However, the application of object detection and bounding box annotation methods has laid bare certain limitations. Specifically, detection-based model's efficacy encounters challenges in scenarios with overlapping objects, leading to an undesirable increase in inference times for the object detection model [33]. To address these challenges and enhance the robustness of our approach, we draw inspiration from the groundbreaking research of Zhang et al [22]. In a departure from conventional methods, we advocate for an innovative alternative: the creation of a density map derived from the precise positions of the object specimens, eschewing the direct reliance on point annotations. This strategic shift in approach aims to surmount the constraints associated with overlapping objects and optimize the efficiency of our quantity estimation task.

In a specific instantiation, this alternative approach is implemented on our dataset comprising M trained images denoted as $I_m\{m = 1, ..., M\}$. Within each image I_m, with N objects are annotated, the set of positions $c_i\{i = 1, ..., N\}$ collectively forms a spatial point map. From This point map 2D, and drawing inspiration from the seminal work of Victor Lempitsky and Andrew Zisserman [33], serves as the foundational basis for formulating a density map. Specifically, the authors employ a 2D Gaussian kernel filter expressed as follows:

$$G_{xy} = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

In this equation, G_{xy} symbolizes the normalized 2D Gaussian kernel with a variance σ and mean value at the position c_i. Then, at each position of an annotated object in the image, it is substituted with a normalized 2D Gaussian kernel as delineated in Figure 5. Due to the normalization of the Gaussian kernel, information pertaining to both the position and the aggregate number of objects in the image is conserved. Nonetheless, it becomes evident that this methodology has foregone details regarding the size of each individual object. For instance, certain studies tends to designate sigma as a constant, resulting in uniform outputs upon traversal through the Gaussian filter for objects of distinct sizes, thereby relinquishing size-specific information. This method ostensibly demonstrates efficacy only in scenarios wherein object sizes are presupposed to be uniform.

On the other hand, similar to the fixed σ approach, within the context of maintaining homogeneity in object sizes, methodologies posited by select papers [22], [41] deploy dynamic σ predicated on the average distance between nearest k-neighbors of objects. This approach presupposes that the spatial distance between an object and its contiguous counterparts dictates its size within the image. However, this solution exhibits suboptimal generalization across diverse cases. In our specific problem domain, even when the size of an object markedly surpasses that of other entities, It is often overlapped by numerous smaller objects. Consequently, utilizing inter-object distances as a size determinant becomes untenable

Confronted with the complexity due to object size heterogeneity, it is recognize that conventional methodologies employed are oftentimes deemed suboptimal and deficient in accuracy. Consequently, to attain optimal precision in density estimation, careful consideration of individual dimensions within the subject population is necessary. From this challenge, we propose a methodology to generate a density map while retaining details about the object size, under the assumption that the sigma parameter of the Gaussian kernel is proportional to the size of the object. This approach ensures a more comprehensive evaluation of size diversity density within a densely populated, thereby augmenting the overall accuracy of object population density estimation. Algorithm 1 furnishes pseudocode employed for generating a density map based on the proposed methodology.

Within Algorithm 1, for each annotation $\{c_i, h_i, w_i\}$ within an object population image dataset, we dynamically instantiate a normalized Gaussian kernel for each object, denoted as G_i, and σ_i represents the variance, and k_i denotes the kernel size. This tailored Gaussian kernel G_i is uniquely assigned to each individual object. Such a approach guarantees that the dataset label, denoted as L, preserves crucial information pertaining to the size characteristics of each object as well as information about that location of the object. On the other hands, in specific scenarios where close-up shots of object are taken, the resulting image may cause the object to appear larger compared to the actual image size. Subsequently, as these object undergo Gaussian filtering, the variance of the filter increases dueto the enlarged size of the object . This augmented variance leads to the disappearance of the density map at the this object's location, creating a challengefor generating density map efficiently. To mitigate this issue, our proposed solutioninvolves incorporating a scaling factor, denoted as α. This α factor, with a value less than 1, is multiplicatively applied to the variance, effectively controlling the image loss experienced by the object during the Gaussian filtering process. Through systematic experimentation, weobserved that

setting the α value to 0.5 yielded the most favorable outcomes, demonstrating its efficacy in addressing the challenge posed by excessive variance during image processing.

Algorithm 1 Algorithm generates a density map by applying a Scale-Adaptive Gaussian Kernel filter

Input: N annotations: $\{c_i, h_i, w_i\}$
Output: Output label $L(H \times W)$

for $\{c_i, h_i, w_i\}$ in N annotations **do**
  $\sigma_i \leftarrow \alpha\sqrt{h_i \cdot w_i}$
  $k_i \leftarrow 2\sigma_i$
  for $x \leftarrow 0$ to k_i **do**
    for $y \leftarrow 0$ to k_i **do**
      $g_{xy} \leftarrow \frac{1}{2\pi\sigma_i^2} e^{-\frac{(x-\frac{k_i}{2})^2+(y-\frac{k_i}{2})^2}{2\sigma_i^2}}$
    end for
  end for
  $G[x,y] \leftarrow \begin{bmatrix} g_{00} & g_{01} & \cdots & g_{0k} \\ g_{10} & g_{11} & \cdots & g_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ g_{k0} & g_{k1} & \cdots & g_{kk} \end{bmatrix}$
  $G[x,y] = \frac{1}{\sum_{x=0}^{k}\sum_{y=0}^{k} G[x,y]} \cdot G[x,y]$
  for $x = 0$ to k **do**
    for $y = 0$ to k **do**
      $l_i[x + c_i - \frac{k}{2}, y + c_i - \frac{k}{2}] = G[x,y]$
    end for
  end for
  $L \mathrel{+}= l_i$
end for
return $L(H \times W)$

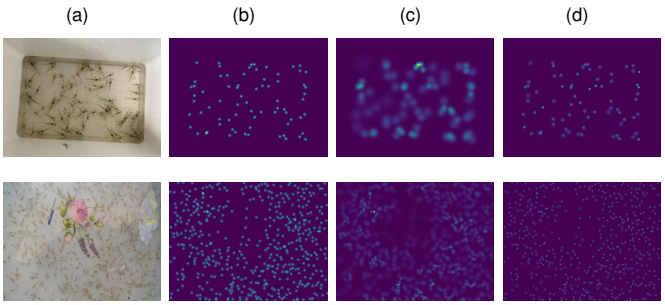

Fig. 5. Generate a depth map by applying the Gaussian kernel. (a) Original image of shrimp larvae, (b) Density map applying Gaussian kernel with constant σ, (c) Density map is generated applying Gaussian kernel with dynamic σ based on the average k-nearest neighbor distance, (d) Density map is generated applying Gaussian kernel with dynamic σ based on size

IV. Experiments Evaluation

A. Acquisition of Datasets

Our study is centered around the intricate task of estimating the quantity of objects present in images under conditions that involve a complex mixture of objects with varying dimensions. The paramount challenge we address is the inherent diversity in the sizes of these objects. Consequently, identifying and characterizing objects with heterogeneous size distributions becomes a critical aspect of our research.

TABLE I
Number of Shrimp Larvae

Number of Shrimp larvae	Number of images
under 100	2647
between 100 and 200	1443
between 200 and 300	448
between 300 and 400	194
between 400 and 500	140
between 500 and 600	32
between 600 and 700	115
Total	5019

The focus on shrimp larvae as our chosen subjects is grounded in their inherent characteristics. The rapid and disparate growth rates among individuals within a group of larvae create a scenario where the size of one sample can significantly exceed that of another within the same population as shown in Figure 6. This variability underscores the need for an innovative approach to object detection and counting, particularly in dynamic environments such as aquaculture facilities.

Our adoption of a supervised learning methodology necessitates the availability of annotated datasets, a pivotal component in training our model. To ensure the relevance and authenticity of our data, we meticulously collected segmented videos in a systematic manner during on-site filming at various aquaculture facilities. This rigorous approach is designed to replicate the complexities of real-world shrimp production environments, reinforcing the robustness and applicability of our experimental results.

In pursuit of a model that thrives in practical deployment scenarios, we executed a comprehensive video recording strategy. This involved capturing multiple videos under diverse conditions, including alterations in camera angles, variations in lighting parameters, and fluctuations in both the quantity and size of shrimp populations within the tanks. Our intention is to create a dataset that encapsulates the intricacies of the shrimp farming milieu, ensuring that our model is well-equipped to handle the challenges posed by real-world scenarios.

From the extensive library of recorded videos, we meticulously extracted images by adopting a sampling approach, selecting one frame every 60 frames from each recorded shrimp video. This process generated a substantial collection of images that served as the foundation for our subsequent analyses.

The next crucial step in our methodology involved annotating each image to facilitate effective utilization of the supervised learning approach. A key consideration was the level of annotation required. Recognizing the natural inclination of humans to count objects, we deemed point annotation as an indispensable feature in our dataset. Furthermore, the annotation of object sizes emerged as another critical aspect of our research. Each object within the frame was meticulously annotated with three parameters: xi (position), wi (width), and hi (height), providing a holistic representation of the object in question.

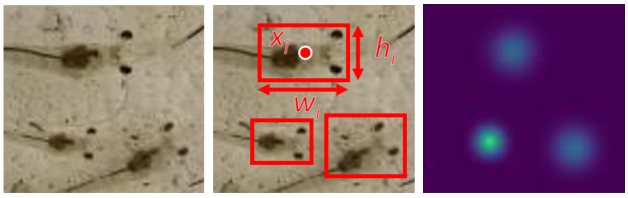

Fig. 6. (a): Original image, (b) Annotated image, (c) Density map generated by Size-Adjustable Gaussian Kernel. We annotated three kinds of keypoints on each larva from their bound box: c_i, h_i and w_i correspond to the center, height and width of the bounding box.

B. Evaluation Metric

In this study, we employ two pivotal metrics to gauge the predictive prowess of the model: Mean Absolute Error (MAE) and Mean Squared Error (MSE).

MAE is expressed by the following formula:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Here, N represents the number of samples, y_i denotes the ground truth value, and $\hat{y}_i$ signifies the predicted value. MAE quantifies the average absolute difference between predicted and actual values of the crowd size.

MSE is computed using the formula:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

This formula calculates the average squared difference between predicted and actual values. The incorporation of both MAE and MSE allows for a comprehensive assessment of the model's accuracy and sensitivity in the context of crowd counting, offering valuable insights into various types of prediction errors. We anticipate that this amalgamation will streamline the evaluation process and elevate the model's performance in real-world applications.

C. Evaluation and Comparison

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT METHODS

Method	MAE			RMSE		
	Fix	k-NN	Scale	Fix	k-NN	Size
CAN	14.44	17.88	10.37	17.52	22.18	13.05
CFFNet	19.99	27.45	13.91	23.71	32.38	17.18
CSRNet	12.06	17.67	13.23	14.58	22.33	16.76
SANet	10.77	36.3	11.91	13.6	43.05	14.59
TEDNet	11.55	19.89	11.08	14.61	24.74	13.73
SGANet	36.69	43.38	20.2	39.93	52.55	25.25
Ours Model	12.59	18.99	9.81	16.05	23.6	12.77

We compared our proposed model with several state-of-the-art methods for crowd counting and density estimation, including CAN, CFFNet, CSRNet, SANet, TEDNet, and SGANet. Each of these methods employs different strategies for handling object counting, offering a diverse set of benchmarks against which to assess our approach.

The results of our experiments, as presented in Table II, clearly demonstrate the superior performance of our model. Our method achieves a significantly lower MAE and MSE compared to the other approaches, indicating higher accuracy and robustness in object counting. This improvement can be attributed to the effectiveness of our Size-Adjustable Gaussian Kernel in capturing object size variations and generating accurate density maps, along with the powerful feature extraction and attention mechanisms incorporated within our CNN architecture.

Moreover, we also delved deeper into the performance of various models under different conditions related to object size handling. We specifically examined three scenarios: using a fixed Gaussian kernel (Fix), a kernel with dynamic sigma based on k-nearest neighbors (k-NN), and our proposed Size-Adjustable Gaussian Kernel (Scale). This analysis provided valuable insights into the strengths and weaknesses of each approach.

Fixed Kernel (Fix): Models employing a fixed Gaussian kernel with constant sigma exhibited the lowest accuracy across the board. This approach, while simple, fails to account for the diverse sizes of objects within the dataset. Consequently, the generated density maps were less accurate, leading to higher MAE and MSE values for all models.

k-Nearest Neighbors (k-NN): Utilizing a dynamic sigma based on the average distance between k-nearest neighbors showed some improvement over the fixed kernel approach. This method attempts to adapt to size variations by considering the local density around each object. However, it still falls short of capturing individual object sizes accurately, particularly in cases where objects overlap significantly or exhibit large size discrepancies. As a result, while k-NN models demonstrated better performance than those with fixed kernels, their accuracy remained inferior to our proposed method.

Size-Adjustable Gaussian Kernel (Size): Our proposed approach, employing a Size-Adjustable Gaussian Kernel, consistently outperformed the other methods across all models. By dynamically adjusting the kernel parameters based on individual object dimensions, we ensured that the generated density maps accurately reflect the size variations within the population. This led to a significant reduction in both MAE and MSE, demonstrating the superior accuracy and robustness of our method.

V. CONCLUSION

This paper presented a novel approach for object counting, specifically addressing the challenges posed by diverse object sizes and dense scenarios. Our proposed method, utilizing a Size-Adjustable Gaussian Kernel for density map generation, effectively incorporates individual object size information, leading to significant improvements in counting accuracy compared to existing methods.

Besides, the integration of the BiFPN architecture and attention mechanisms, such as ECA and MSCA, within our

CNN model further enhanced feature extraction and object recognition across different scales. Extensive experiments on a challenging shrimp larvae image dataset confirmed the efficacy of our approach, showcasing its ability to handle size variations and dense clusters effectively.

The superior performance of our model paves the way for more precise and reliable object counting in various applications, including crowd analysis, traffic monitoring, ecological surveys, and industrial quality control. Future work could explore incorporating additional object features, such as shape or texture, into the density map generation process and investigating the application of our method to other object counting tasks with similar challenges.

REFERENCES

[1] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[2] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.

[3] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1140–1156, 2022.

[4] C. C. Loy, K. Chen, S. Gong, and T. Xiang, "Crowd counting and profiling: Methodology and evaluation," in *Modeling, Simulation and Visual Analysis of Crowds: A Multidisciplinary Perspective*. Springer, 2013, pp. 347–382.

[5] S. Armalivia, Z. Zainuddin, A. Achmad, and M. A. Wicaksono, "Automatic counting shrimp larvae based you only look once (yolo)," in *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*. IEEE, 2021, pp. 1–4.

[6] L. Zhang, X. Zhou, B. Li, H. Zhang, and Q. Duan, "Automatic shrimp counting method using local images and lightweight yolov4," *Biosystems Engineering*, vol. 220, pp. 39–54, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1537511022001234

[7] L. Xu, H. Deng, Y. Cao, W. Liu, G. He, W. Fan, T. Wei, L. Cao, T. Liu, and S. Liu, "Bass detection model based on improved yolov5 in circulating water system," *PloS one*, vol. 18, no. 3, p. e0283671, 2023.

[8] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2913–2920.

[9] F. P. Nurmaida, B. S. B. Dewantara, A. I. Gunawan, A. R. Mahendra, J. W. Prayanata, Z. D. M. Fasya, and S. W. Trianto, "Comparison of cnn-based design for shrimp seed counting machine," in *2023 International Electronics Symposium (IES)*. IEEE, 2023, pp. 493–498.

[10] K.-T. Nguyen, C.-N. Nguyen, C.-Y. Wang, and J.-C. Wang, "Two-phase instance segmentation for whiteleg shrimp larvae counting," in *2020 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2020, pp. 1–3.

[11] Y. Hashisho, T. Dolereit, A. Segelken-Voigt, R. Bochert, and M. Vahl, "Ai-assisted automated pipeline for length estimation, visual assessment of the digestive tract and counting of shrimp in aquaculture production." in *VISIGRAPP (4: VISAPP)*, 2021, pp. 710–716.

[12] T. Hong Khai, S. N. H. S. Abdullah, M. K. Hasan, and A. Tarmizi, "Underwater fish detection and counting using mask regional convolutional neural network," *Water*, vol. 14, no. 2, 2022. [Online]. Available: https://www.mdpi.com/2073-4441/14/2/222

[13] R. Kesvarakul, C. Chianrabutra, and S. Chianrabutra, "Baby shrimp counting via automated image processing," in *Proceedings of the 9th International Conference on Machine Learning and Computing*, 2017, pp. 352–356.

[14] M. Solahudin, W. Slamet, and A. Dwi, "Vaname (litopenaeus vannamei) shrimp fry counting based on image processing method," in *IOP Conference Series: Earth and Environmental Science*, vol. 147, no. 1. IOP Publishing, 2018, p. 012014.

[15] J. Kaewchote, S. Janyong, and W. Limprasert, "Image recognition method using local binary pattern and the random forest classifier to count post larvae shrimp," *Agriculture and Natural Resources*, vol. 52, no. 4, pp. 371–376, 2018.

[16] C.-T. Yeh and M.-C. Chen, "A combination of iot and cloud application for automatic shrimp counting," *Microsystem Technologies*, pp. 1–8, 2019.

[17] E. A. Awalludin, M. M. Yaziz, N. A. Rahman, W. N. J. H. W. Yussof, M. S. Hitam, and T. T. Arsad, "Combination of canny edge detection and blob processing techniques for shrimp larvae counting," in *2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, 2019, pp. 308–313.

[18] C.-T. Yeh and M.-S. Ling, "Portable device for ornamental shrimp counting using unsupervised machine learning." *Sensors & Materials*, vol. 33, 2021.

[19] J. Zhang, G. Yang, L. Sun, C. Zhou, X. Zhou, Q. Li, M. Bi, and J. Guo, "Shrimp egg counting with fully convolutional regression network and generative adversarial network," *Aquacultural Engineering*, vol. 94, p. 102175, 2021.

[20] W.-C. Hu, L.-B. Chen, M.-H. Hsieh, and Y.-K. Ting, "A deep-learning-based fast counting methodology using density estimation for counting shrimp larvae," *IEEE Sensors Journal*, vol. 23, no. 1, pp. 527–535, 2022.

[21] D. Liu, B. Xu, Y. Cheng, H. Chen, Y. Dou, H. Bi, and Y. Zhao, "Shrimpseed_net: Counting of shrimp seed using deep learning on smartphones for aquaculture," *IEEE Access*, 2023.

[22] 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.

[23] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[24] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100.

[25] Y. Xia, Y. He, S. Peng, Q. Yang, and B. Yin, "Cffnet: Coordinated feature fusion network for crowd counting," *Image and Vision Computing*, vol. 112, p. 104242, 2021.

[26] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5099–5108.

[27] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, and L. Shao, "Crowd counting and density estimation by trellis encoder-decoder networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6133–6142.

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

[29] B. Wang, H. Liu, D. Samaras, and M. H. Nguyen, "Distribution matching for crowd counting," *Advances in neural information processing systems*, vol. 33, pp. 1595–1607, 2020.

[30] Q. Wang and T. P. Breckon, "Crowd counting via segmentation guided attention networks and curriculum loss," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15 233–15 243, 2022.

[31] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6142–6151.

[32] Z. Fan, H. Zhang, Z. Zhang, G. Lu, Y. Zhang, and Y. Wang, "A survey of crowd counting and density estimation based on convolutional neural network," *Neurocomputing*, vol. 472, pp. 224–251, 2022.

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

[34] D. Guo, K. Li, Z.-J. Zha, and M. Wang, "Dadnet: Dilated-attention-deformable convnet for crowd counting," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1823–1832.

[35] S. Bai, Z. He, Y. Qiao, H. Hu, W. Wu, and J. Yan, "Adaptive dilated network with self-correction supervision for counting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4594–4603.

[36] Y.-J. Ma, H.-H. Shuai, and W.-H. Cheng, "Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation," *IEEE Transactions on Multimedia*, vol. 24, pp. 261–273, 2021.

[37] X. Chen, Y. Bin, N. Sang, and C. Gao, "Scale pyramid network for crowd counting," in *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2019, pp. 1941–1950.

[38] J. Ma, Y. Dai, and Y.-P. Tan, "Atrous convolutions spatial pyramid network for crowd counting and density estimation," *Neurocomputing*, vol. 350, pp. 91–101, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219304059

[39] W. Zhai, M. Gao, Q. Li, G. Jeon, and M. Anisetti, "Fpanet: feature pyramid attention network for crowd counting," *Applied Intelligence*, pp. 1–18, 2023.
[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[41] W.-C. Hu, L.-B. Chen, M.-H. Hsieh, and Y.-K. Ting, "A deep-learning-based fast counting methodology using density estimation for counting shrimp larvae," *IEEE Sensors Journal*, vol. 23, no. 1, pp. 527–535, 2022.