Size-Adaptive Density Map Estimation and Multiscale Feature Extraction Network for Robust Prawn Postlarvae Counting

Thai-Ha Dang, Fabio Dos Santos Neto, Xinrong Li, *Senior Member, IEEE*,
Breana Smithers, Edward Mager, and Miguel Acevedo

***Abstract*—Accurate prawn postlarvae counting is critical for optimizing aquaculture production and management. Traditional manual counting methods are labor-intensive and prone to errors, while existing automated approaches face challenges such as small object sizes, overlapping instances, and variations in postlarvae dimensions. To address these limitations, we propose P2CountNet, a novel deep-learning framework that integrates size-adaptive density map estimation with a multiscale feature extraction network for precise and robust prawn postlarvae counting. The size-adaptive density map dynamically adjusts to varying postlarvae sizes, ensuring accurate counting even in dense and heterogeneous environments. Meanwhile, the multi-scale feature extraction network captures features across multiple resolutions, effectively detecting tiny and overlapping postlarvae. Experimental evaluations on a newly curated prawn postlarvae dataset demonstrate that P2CountNet significantly outperforms state-of-the-art counting methods, achieving lower mean absolute error (MAE) and mean squared error (MSE). Notably, the proposed method achieves an average counting accuracy of 94.32% for testing datasets. Key contributions include the development of the size-adaptive density estimation technique, the design of a multiscale feature extraction network tailored for small object counting, and extensive evaluations validating the approach. This work provides a robust and efficient solution for prawn postlarvae counting, with promising applications in advancing aquaculture practices.**

***Index Terms*—attention mechanism, deep learning, density estimation, prawn postlarvae counting, size-adjustable Gaussian kernel.**

I. Introduction

Shrimp farming has become one of the most significant and fastest-growing sectors in the global seafood aquaculture industry. According to the Food and Agriculture Organization (FAO), shrimp farming contributes substantially to world aquaculture production, with 8.5 million tons produced in 2022 [1]. This industry not only meets the growing demand for seafood but also plays a crucial role in supporting the economies of many developing countries by providing employment opportunities and boosting foreign exchange earnings through exports.

The success of shrimp and prawn farming is heavily based on management practices that improve production efficiency while ensuring sustainability [2]. A key aspect of shrimp and prawn aquaculture management is the accurate estimation of postlarvae populations. Accurate counting of these postlarvae is crucial to maximize growth rates and feed conversion efficiency, as overstocking can lead to resource competition, higher mortality, and disease spread, while under-stocking may result in under-utilization of pond capacity [3]. Accurate postlarvae counting also allows for better resource allocation, enabling farmers to optimize the use of feed, aeration, and water quality management, which leads to cost savings and reduced environmental impacts. Additionally, knowing the initial postlarvae population aids in yield prediction and planning, allowing farmers to schedule harvests and meet market demands efficiently. Lastly, monitoring postlarvae populations is vital for disease management, as it enables early detection of mortality events that may signal disease outbreaks or adverse environmental conditions, allowing for timely interventions to prevent losses and pathogen spread.

Conventional approaches to counting shrimp and prawn postlarvae rely on manual methods, where skilled workers employ microscopes or visual estimation to count postlarvae in sampled volumes of water samples [4]. However, random sampling often introduces significant inaccuracies because postlarvae vary in size and are not evenly dispersed. These inaccuracies can lead to inequitable transactions between hatcheries and aquaculture producers, as well as miscalculations in feed requirements and final harvest yields. Moreover, the sampling process itself risks harming the postlarvae. To address this, a non-contact electronic fry counter was developed, utilizing infrared emitters paired with a photodiode [5]. Despite this innovation, such techniques remain time-consuming, labor-intensive, and prone to human error and inconsistent results. As aquaculture operations expand and automation gains traction, the demand for precise, efficient, and automated counting solutions for shrimp and prawn postlarvae continues to rise.

Recently, various modern convolutional neural network

Manuscript received December 2024;date of current version. The associate editor coordinating the review of this article and approving it for publication:... (Corresponding author: Xinrong Li.)

This work was supported in by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) and National Science Foundation (NSF) Cyber-Physical Systems (CPS); Grant Number: 13711289

T. H. Dang, X. Li, B. Smithers, and M. Acevedo are with the Department of Electrical Engineering, University of North Texas, Denton, TX 76205 USA, (e-mail: thaihadang@my.unt.edu; xinrong.li@unt.edu; breana.smithers@unt.edu; miguel.acevedo@unt.edu)

D. S. Fabio and E. Mager are with the Department of Biological Sciences, University of North Texas, Denton, TX 76205 USA, (e-mail: fabiodossantosneto@my.unt.edu and edward.mager@unt.edu).

E. Mager and M. Acevedo are with the Advanced Environmental Research Institute, University of North Texas, Denton, TX 76205 USA.

Fig. 1. Example of high-density prawn postlarvae crowd in an aquaculture environment with more than 1000 counts.

(CNN) approaches have been proposed to extract more discriminative deep features, increasing the accuracy of shrimp counting. For example, Thai *et al*. [6] used a U-Net architecture to segment possible shrimp candidates from backgrounds and then count these possible individuals. Similarly, Zhang *et al*. [7] employed a Mask R-CNN to isolate shrimp regions for white shrimp postlarvae counting. Other CNN-based methods have used various object detectors to count prawns if their sizes in the image are larger. Armalivia *et al*. [8] used a lightweight YOLOv4 to develop an automatic shrimp counting system. Furthermore, Yu *et al*. [9] used data augmentation methods to synthesize objects representing peeled shrimp and then trained the YOLOv4 detector to count them. However, the above detection methods cannot accurately and effectively estimate counts in high-density scenes. This is especially true for counting prawn postlarvae, as shown in the example of Fig. 1. Counting prawn postlarvae is similar to crowd counting but more difficult since prawns swim in a 3-D space, but people in a crowd (observed from a bird's view) move on a 2-D plane with fewer occlusions.

Detection-based methods [8], [7] are effective for distinct and non-overlapping objects but face challenges in dense situations where overlaps occur because these methods depend on bounding boxes, which can lead to errors in counting and localization. Similarly, regression-based techniques [10], [11] provide a direct connection between image features and object counts but frequently miss important spatial details, leading to a lack of information about the arrangement and distribution of objects in the scene.

Approaches focusing on density maps have recently shown potential as a viable solution [12], utilizing spatial cues to produce maps representing how objects are scattered throughout the scene. However, these techniques typically rely on static Gaussian filters, with uniform kernel parameters, disregarding the natural differences in object dimensions [13]. As a result, performance suffers, especially in scenarios involving groups characterized by varying object scales. Although some approaches endeavor to overcome this challenge by applying adaptive Gaussian filters informed by the mean spacing between adjacent targets [14], they remain less accurate than methods that integrate size data for each object.

To address the challenges in shrimp and prawn postlarvae counting, in this paper, we propose P2CountNet, a novel deep-learning framework specifically designed to provide an accurate counting solution. This approach leverages the latest advances in CNN and incorporates innovative techniques tailored to the unique requirements of prawn postlarvae counting. One of the key components of our proposed P2CountNet is the size-adaptive density map estimation method, based on incorporating size information and adjusting the density contribution of each detected individual based on its size. Such an approach ensures that both large and small postlarvae are accurately represented, particularly in densely packed scenes, improving the overall counting accuracy. Additionally, P2CountNet includes a multiscale feature extraction network designed to capture the characteristics of small objects like prawn postlarvae across different scales. The key contributions of our work are summarized in the following:

1) We introduce a dynamic size-adaptive mechanism within the density map estimation process, allowing the model to adjust to varying postlarvae sizes and improve counting accuracy.
2) Our proposed framework effectively handles occlusions and overlapping postlarvae, significantly outperforming existing methods in high-density scenarios.
3) We validate our approach through comprehensive experiments on real-world datasets, demonstrating substantial improvements over state-of-the-art techniques in terms of accuracy and computational efficiency.

The remainder of this paper is organized as follows. Section II reviews related works covering existing methods in crowd counting and shrimp postlarvae counting. Section III introduces the proposed P2CountNet model, detailing its architecture, the size-adaptive density map estimation, and the multi-scale feature extraction network. Section IV presents the experiments and evaluation, including model accuracy comparisons, performance metrics, and an ablation study. Finally, Section V concludes the paper by summarizing key findings and discussing potential future research directions.

II. Related Works

A. Crowd Counting

Crowd counting has become an increasingly important task within computer vision, with significant applications in public safety monitoring and crowd behavior analysis. Over time, numerous convolutional neural network (CNN)-based methods have been introduced to estimate crowd density maps from input images [15]. These techniques typically feed the image into a backbone network, which generates a density map through its final layer. The overall crowd count is then derived by integrating the values within this map. For instance, the work in [16] introduced the Multi-column Convolutional Neural Network (MCNN), a straightforward yet powerful architecture that links input images to their corresponding density maps. It leverages multiple convolutional kernels of varying sizes to capture multi-scale features, enabling effective estimation of head counts at different scales.

To reduce redundant computation of convolutional features across sub-regions, Xiong *et al.* [17] proposed generating multi-resolution feature maps by segmenting a dense region into smaller sub-regions, with their counts derived from a pre-defined closed set. Jiang *et al.* [18] addressed scale variation by incorporating multiscale contextual cues directly into the regression model. In [19], a density attention network was introduced to create scale-specific attention masks, enabling the model to concentrate on relevant spatial resolutions for crowd counting. Miao *et al.* [20] utilized a densely connected architecture to preserve multiscale features, along with an attention mechanism to suppress background noise. Thanasutives *et al.* [21] developed M-SFANet, a dual-path multiscale fusion network combining SFANet and SegNet, aimed at achieving both accuracy and efficiency in crowd counting tasks. To enhance generalization, Zhang *et al.* [19] proposed a CNN model employing a switching strategy for alternating between density and count estimation. Despite these advancements, many of these approaches still struggle with efficiency, posing challenges for real-time applications such as prawn counting systems.

Recent advancements in transformer-based detectors, such as DETR [22] and Swin Transformer variants [23], have demonstrated strong performance in handling occlusions and scale variation for general-purpose object detection tasks. These models leverage global attention to capture long-range dependencies and yield competitive detection accuracy (mAP). However, in the context of dense and translucent prawn postlarvae, where individuals overlap heavily and instance boundaries are inherently ambiguous, detection-based approaches often struggle with under- or over-counting. In contrast, density-based methods aggregate local evidence without enforcing explicit separation, making them more suitable for aquaculture counting tasks. Therefore, while we acknowledge the progress in transformer-based detection, our focus remains on density estimation frameworks, which have been widely adopted in crowd counting and are better aligned with the challenges of prawn postlarvae counting.

B. Shrimp and Prawn Counting

Underwater vision presents domain-specific challenges beyond generic crowd counting. Detection refinement for benthic structures (e.g., *Nephrops norvegicus* burrows) improves precision in noisy seabed imagery [24]. Context-enriched light detectors have been adapted for counting in complex aerial scenes, and density-aware routing mechanisms modulate processing by local crowd levels [25]. Our approach is complementary: instead of box refinement or route selection, we predict a feature-driven scale map and regress densities, which proves robust to translucency and partial visibility common in hatcheries.

Shrimp and prawn postlarvae counting is more challenging than crowd counting since people in a crowd move on a plane (when observed from an aerial perspective), whereas prawns swim in a three-dimensional environment, with further complications by potential occlusions. Traditional vision-based methods simplify the background and isolate potential individual regions using a threshold technique to extract individual candidates effectively. For example, Kesvaraku *et al.* [26] utilized an adaptive thresholding method and a blob detection technique to count individuals. Solahudin *et al.* [27] used a thresholding method to extract possible individual regions and then perform morphological operations such as opening and closing to isolate their centers for counting. Kaewchote *et al.* [28] marked possible shrimp candidates to extract their local binary pattern (LBP) features and verified them based on a random forest classifier. However, traditional vision-based methods may not be accurate or stable in counting small shrimp with severe occlusions.

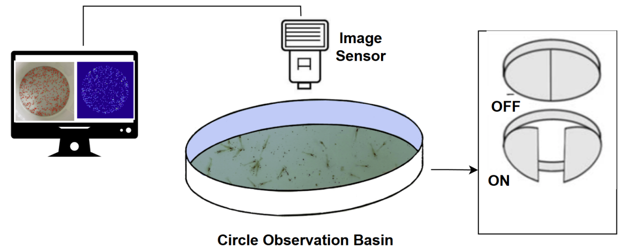


Fig. 2. Application of the prawn postlarvae counting system using the proposed P2CountNet methodology.

C. Density Map Generation

Determining object quantities in scenes that are filled with closely packed, intersecting entities is a widely explored research area, and many effective remedies have been suggested to overcome it. Loy *et al.*. [29] identified three main strategies for dealing with this challenge: detection-driven approaches [8], [30], [31], regression-oriented techniques [10], [11], and density estimation strategies [28], [32], [33], [27].

Notably, detection-driven methodologies excel at pinpointing objects within visual data [34]. Such methods rely on advanced object detection infrastructure. Their effectiveness stems from carefully refined detection models that can distinguish and capture rich object attributes. Well-known examples include YOLO (You Only Look Once) [8], [7] and R-CNN (Region-based Convolutional Neural Network) [35], [30], [31]. Such networks, thoroughly prepared to recognize targets, utilize bounding boxes to count objects in visual scenes.

However, the accuracy of detection-driven methods degrades in cases featuring highly packed or occluded elements because accurate bounding box placement is difficult in heavily crowded and overlapping object scenes [36]. This shortcoming endures even though the frameworks proficiently absorb object information. Consequently, these solutions yield the best performance when targets present well-defined boundaries.

To overcome the constraints inherent in detection-based methodologies, regression-based approaches directly map 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 variations in the spatial distribution of objects—an essential factor impacting accurate counting.

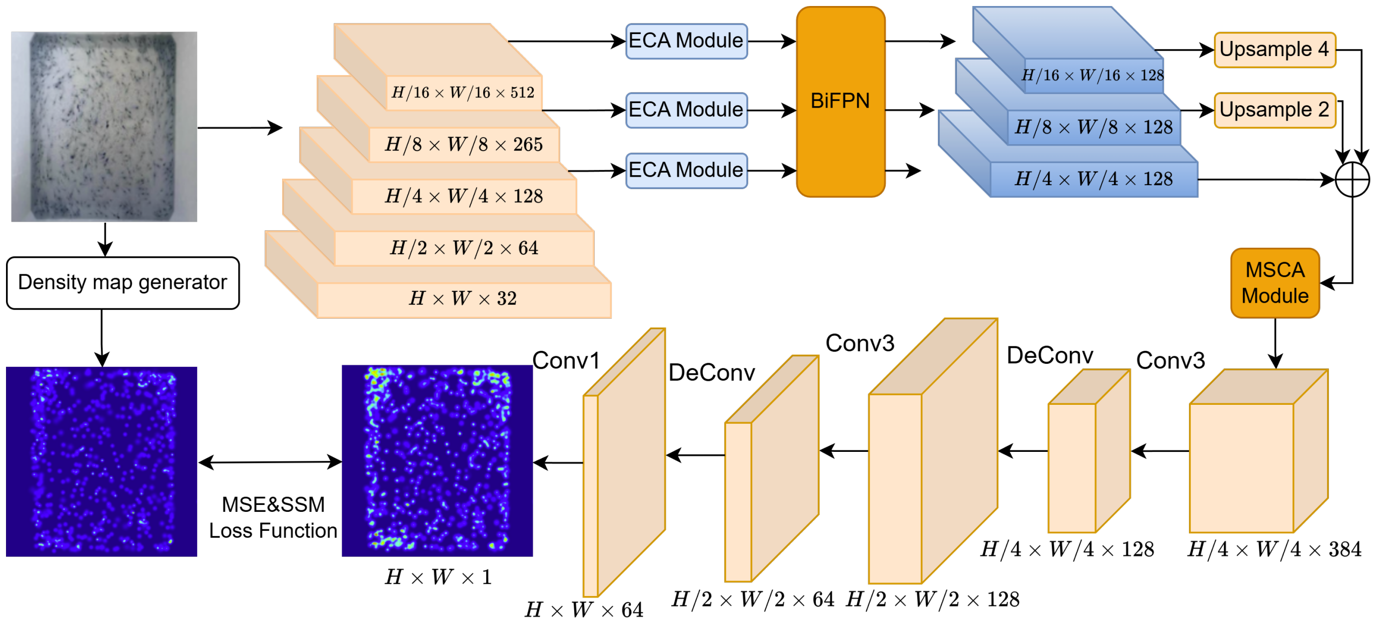


Fig. 3. Overview of the P2CountNet architecture with modified VGG16 [37] as backbone and BiFPN [38] to connect feature maps.

III. P2CountNet for Prawn Postlarvae Counting

Figure 2 illustrates the implementation of the prawn postlarvae counting system using the proposed P2CountNet approach. To begin the process, the user pours a sample of prawn postlarvae into a circular observation basin positioned 80 cm below the imaging sensor. The system then captures an image, which is processed using a deep learning model to count the postlarvae. Once the counting is completed, the interior of the observation basin is reset and prepared for the next counting cycle. Figure 3 shows the architecture of P2CountNet, which includes two primary innovations, i.e., size-adaptive density map estimation and a multiscale feature network. These components work together to provide robust counting in complex aquaculture environments.

A. Adaptive Density Map Generation

In highly crowded visual environments, such as those encountered when examining groups of prawn postlarvae, each individual entity is characterized by three essential parameters: c_i, h_i, and w_i. These attributes correspond to the entity's spatial coordinates, its height, and its width, respectively. By incorporating these elements, it becomes possible to achieve a more structured depiction of the scene, enabling researchers to understand both the precise location and the physical dimensions of each object present. In a setting where numerous entities cluster together, c_i offers an indispensable reference point for pinpointing the exact placement of each individual within a densely packed space. This parameter ensures that even within a visually complex setting, slight positional differences are captured, allowing a clearer interpretation of object distributions and arrangements.

Simultaneously, h_i and w_i work in concert to describe the geometric contours of the entity, encompassing its vertical and horizontal extents. Their combined influence provides insight into the overall shape and occupied area of the object, establishing how it fits spatially alongside neighboring entities. This integrated perspective is reminiscent of the bounding-box schemes frequently employed in sophisticated object detection methods. Here, however, these parameters are tailored to the unique challenges of densely populated conditions, where identifying individual entities is not simply a matter of detection but also requires careful analysis of location and shape. By considering c_i, h_i, and w_i together, researchers can move beyond merely counting objects, focusing instead on how their placement, size, and relative positioning reveal hidden patterns, interactions, and structural characteristics within intricate environments.

However, the application of object detection and bounding box annotation methods introduces certain limitations. Specifically, the detection-based model's efficacy encounters challenges in scenarios with overlapping objects, leading to an undesirable increase in inference time for the object detection model [12]. To address these challenges and enhance the robustness of our proposed approach, we draw inspiration from the research of Zhang *et al* [16]. 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 training images denoted as $\{I_m\}_{m=1}^{M}$. Within each image I_m, with N objects annotated, the set of positions $\{c_i\}_{i=1}^{N}$ collectively forms a spatial point map. Using concepts from the work of Victor

Lempitsky and Andrew Zisserman [12], a 2D map serves as the foundational basis for formulating a density map. Specifically, these 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 4. so that the 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 tend to designate σ as a constant, resulting in uniform outputs upon traversal through the Gaussian filter for objects of distinct sizes, thereby losing size-specific information. This method demonstrates efficacy only in scenarios wherein object sizes are assumed to be uniform.

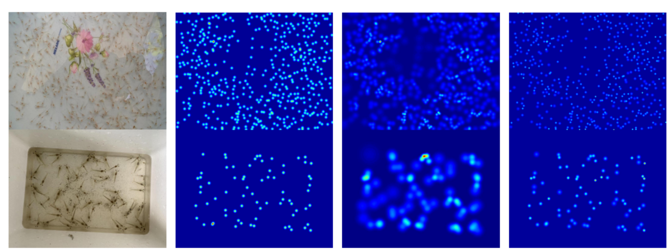
Within the context of maintaining homogeneity in object sizes, alternative methodologies [32], [16] deploy dynamic σ predicated on the average distance between nearest k-neighbors of objects. This approach assumes that the spatial distance between an object and its contiguous counterparts dictates its size within the image. However, this solution is difficult to generalize across diverse cases. In our problem domain, when the size of an object markedly surpasses that of other entities, it is often overlapped by numerous smaller objects. Consequently, it becomes untenable to utilize inter-object distances as a size determinant.

Due to the complexities caused by the heterogeneity of the size of the object, conventional methodologies are frequently deemed suboptimal and deficient in accuracy [39]. Consequently, carefully considering individual dimensions within the subject population is necessary to attain optimal accuracy in density estimation. To address this challenge, in this paper, we propose a new methodology to generate a density map while retaining details about the object size under the assumption that the parameter σ of the Gaussian kernel is proportional to the object's size. This approach ensures a more comprehensive evaluation of size diversity density within a densely populated area, thereby augmenting the overall accuracy of object population density estimation. Algorithm 1 furnishes the pseudocode that could be employed to generate 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, σ_i represents the variance, $\lfloor\cdot\rfloor$ represents floor operation, and k_i denotes the kernel size. The tailored Gaussian kernel G_i is uniquely assigned to each individual object. Such an approach guarantees that the dataset labels, denoted as D, preserve crucial information pertaining to the size characteristics of each object as well as information about the location of the object. On the other hand, in specific scenarios where close-up shots of the object are taken, the resulting image may cause the object to appear larger compared to the actual image size. Subsequently, as these objects undergo Gaussian filtering, the variance of the filter increases due to the enlarged size of the object. This augmented variance leads to the disappearance of the density map at this object's location, creating a challenge for generating density maps efficiently. To mitigate this issue, our proposed solution involves incorporating a non-negative

Algorithm 1 Size-adaptive Gaussian kernel density map generator algorithm.

Input: N annotations: $\{c_i, h_i, w_i\}_{i=1}^{N}$
Output: Output density map $D \in \mathbb{R}^{(H\times W)}$
Initialize $D \leftarrow \mathbf{0}^{H\times W}$
for each annotation $i = 1$ to N **do**
  Compute standard deviations: $\sigma_i \leftarrow \alpha\sqrt{h_i w_i}$
  Define kernel size: $k_i \leftarrow \lfloor 2\sigma_i \rfloor$
  for $x \leftarrow 0$ to k_i **do**
    for $y \leftarrow 0$ to k_i **do**
      $g_{xy} \leftarrow \frac{1}{2\pi\sigma_i^2} \exp\left(-\frac{(x-\frac{k_i}{2})^2+(y-\frac{k_i}{2})^2}{2\sigma_i^2}\right)$
    end for
  end for
  $G[x,y] \leftarrow \begin{bmatrix} g_{00} & g_{01} & \cdots & g_{0k_i} \\ g_{10} & g_{11} & \cdots & g_{1k_i} \\ \vdots & \vdots & \ddots & \vdots \\ g_{k_i0} & g_{k_i1} & \cdots & g_{k_ik_i} \end{bmatrix}$
  $G[x,y] = \frac{G[x,y]}{\sum_{x=0}^{k_i}\sum_{y=0}^{k_i} G[x,y]}$
  for $x = 0$ to k_i **do**
    for $y = 0$ to k_i **do**
      $l_i[x + x_i - \frac{k_i}{2}, y + y_i - \frac{k_i}{2}] = G[x,y]$
    end for
  end for
  $D \mathrel{+}= l_i$
end for
return density map D =0



(a) (b) (c) (c)

Fig. 4. Comparison of density map outputs with different kernel types: (a) shows the original image of prawn postlarvae, (b) presents the density map generated using a Gaussian kernel with constant σ, (c) displays the density map created using a Gaussian kernel with dynamic σ based on the average k-nearest neighbor distance, and (d) illustrates the density map generated using a Gaussian kernel with dynamic σ based on size.

scaling factor, denoted as α with a value less than 1, which is multiplied by the variance, effectively controlling the image loss experienced by the object during the Gaussian filtering process. Through systematic experimentation, we observed 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.

We model the density field $D(\mathbf{x})$ as a superposition of Gaussian kernels centered at labeled postlarvae positions $\{\mathbf{x}_i\}_{i=1}^{N}$, but *with spatially varying scale* predicted from the image:

$$D(\mathbf{x}) = \sum_{i=1}^{N} \frac{1}{2\pi\,\sigma_i^2} \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|_2^2}{2\sigma_i^2}\right) \tag{2}$$

where $\sigma_i = \alpha\, s(\mathbf{x}_i) + \beta$, $s(\cdot) \in [0,1]$ is a learned, *dimensionless* local scale map produced by a lightweight scale head operating on multi-scale features. The affine parameters (α, β) are learned jointly with the network (initialized to cover a reasonable kernel support; no manual tuning is required). Unlike fixed-σ kernels, the spatially adaptive formulation in (2) aligns the kernel support with local perspective, translucency, and overlap patterns that are characteristic of prawn postlarvae in tanks.

Concretely, the backbone-BMFPN stack yields features $\{F^l\}$ across levels $l \in \{P_3, \ldots, P_7\}$. A scale head $g(\cdot)$ with depth-wise separable 3×3 convolutions and sigmoid output produces $s = \sigma(g(\mathrm{Concat}(F^l)))$. The density head $h(\cdot)$ predicts a base density map $\tilde{D}$. At training time, the ground-truth density is generated using (2); at inference, h directly regresses D while g provides an auxiliary constraint (below). This decouples scale estimation from density regression and stabilizes learning in highly crowded regions.

Classical density-estimation methods often derive σ_i from the average distance to the k nearest annotations ("geometry-adaptive" kernels) or from manually designed perspective maps. While effective on pedestrian scenes, these strategies rely on point-pattern geometry that can be unreliable for translucent, overlapping postlarvae with strong appearance ambiguity. Our approach differs in three key aspects:

1) **Feature-driven scale prediction.** Instead of computing σ_i from annotation geometry, we learn a *dimensionless* scale map $s(\mathbf{x})$ from image features and map it affinely to $\sigma_i = \alpha s(\mathbf{x}_i) + \beta$ (Eq. 2). This leverages actual visual evidence (blur extent, edges, contrast) tied to translucency and micro-bubbles.
2) **Regularized, calibration-free adaptation.** The scale-consistency and boundedness losses constrain $s(\mathbf{x})$ to correlate with local structure without manual perspective calibration or scene-specific tuning; loss weights are fixed across datasets.
3) **Decoupled heads for stability.** We decouple density regression and scale estimation via separate heads sharing the BMFPN features, which improved stability in crowded, low-contrast regions relative to single-head designs.

Together, these design choices differentiate our method from prior kNN-/perspective-based adaptive kernels and are motivated by the unique optics of underwater hatchery imaging.

B. Bi-directional Multiscale Feature Pyramid Network Architecture (BMFPN)

To tackle the challenge of crowd counting, our approach leverages VGG16 as the foundational backbone of the model. By integrating efficient channel attention (ECA) with the bidirectional feature pyramid network (BiFPN), we can achieve improved deep feature extraction and enhanced focus on critical areas, enabling the model to manage complex and crowded scenarios effectively. To further boost its scalability across varying object sizes, the model incorporates the multiscale conv attention (MSCA) module for additional refinement. This innovative combination creates a powerful and flexible framework designed to redefine crowd analysis in practical, complex, real-world environments.

1) Feature Map Extraction (FME): In this research, we selected VGG16 [37] as the core network, utilizing its strong ability to extract image features. Due to its deep, layer-rich design, VGG16 performs consistently well across numerous computer vision challenges. Its adaptability in both mid-level and deeper layers suits the intricate nature of feature extraction essential for crowd counting. However, given VGG16's substantial complexity, which demands extensive memory and computational support, we chose to halve the channel count in each layer. Tests confirmed this change did not significantly reduce its feature detection capabilities while producing a leaner and more efficient model.

Within this framework, Layer 3 of VGG-16 captures fine-grained details, making it particularly effective for detecting small or distant objects. Its primary role is to extract intricate patterns and integrate them into a comprehensive visual representation. Layer 4 serves as a transitional tier, balancing detailed feature extraction with broader contextual understanding, enabling the recognition of medium-sized entities and their spatial relationships. Layer 5, at a lower resolution, emphasizes high-level abstractions, capturing overall structure and prominent patterns within the scene. This hierarchical feature extraction approach, inherent to the VGG architecture, makes it highly effective for crowd analysis and object detection, where both local and global contexts are crucial [40]. Consequently, this study employs a combination of Layers 3, 4, and 5 as a unified feature map to leverage their complementary strengths.

Drawing on the multi-level feature representations from these VGG16 layers, we identified the need to incorporate specialized attention mechanisms into our design, such as the ECA module. This integration fulfills several important objectives. As shown in Fig. 5, mechanisms that highlight key components of the input data strengthen the model's ability to identify and emphasize critical cues. This capability proves immensely valuable, particularly for complex tasks like counting individuals in crowded scenes. The ECA module, proposed by Zhang *et al*. [41], is specifically designed to efficiently distill essential signals at the channel level. Instead of combining spatial and channel-based elements, it computes an attention score for each channel independently, reducing computational demands. This streamlined approach employs a one-dimensional convolution applied along the channel axis, followed by a sigmoid function to produce a set of channel-wise attention weights.

Furthermore, merging distinct attentional techniques enables the system to capitalize on their combined merits while diminishing their individual drawbacks. Within this framework, embedding ECA into a BiFPN arrangement establishes a powerful interplay. Although BiFPN constructs a hierarchical pyramid that integrates multi-layered information, ECA refines the focus within every channel, ensuring that crucial details receive appropriate emphasis throughout the entire feature pyramid.

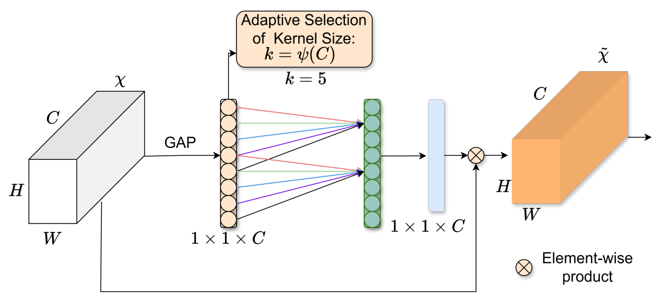


Fig. 5. Illustration of the efficient channel attention (ECA) module. After applying global average pooling (GAP) to the feature maps, the ECA module computes attention weights using a lightweight 1D convolution with an adaptive kernel size k. The value of k is dynamically determined based on the number of channels C.

This integration not only improves how effectively the model identifies and exploits features spanning various scales, but it also refines how computational resources are allocated. The enhanced interplay between these attention components strengthens the system's aptitude for distinguishing and counting entities within crowded scenes, thus making it adaptable to a wide array of challenging, real-world conditions. By embedding an ECA block into each of the third, fourth, and fifth layers, we construct a reliable, adaptable, and resource-efficient model that performs exceptionally well in multiple contexts.

After consolidating the extracted features with a BiFPN, the processed information passes through the MSCA module. This MSCA module is vital for intensifying the focus within the model, directing particular emphasis on features appearing at various scales in the image. It employs convolutional operations at multiple size levels concurrently, enabling the model to perceive both fine-grained details and larger structures among the crowd.

Integrating the MSCA module into our overall design ensures the model's attention is optimally distributed across various spatial resolutions, thereby enhancing accuracy in counting tasks. In addition to refining object recognition over a range of sizes, MSCA synergizes with the BiFPN to improve the model's feature representation and attention strategy at all layers. Bringing together BiFPN, MSCA, and ECA yields a multi-directional, potent, and adaptable architecture. This setup not only upgrades how features are represented but also heightens attentional accuracy and object detection capabilities at every scale. Consequently, the resulting model is exceptionally well-suited to the demanding requirements of crowd counting.

2) Density Map Estimator (DME): After the BiFPN assembles and merges critical features from different layers of VGG, attention turns to the DME layer, a key component for obtaining an accurate density map. Its structure, depicted in Fig. 6, not only adapts to various conditions but also fully harnesses the information available from the BiFPN output layers, making the crowd-counting pipeline notably more durable and resourceful.

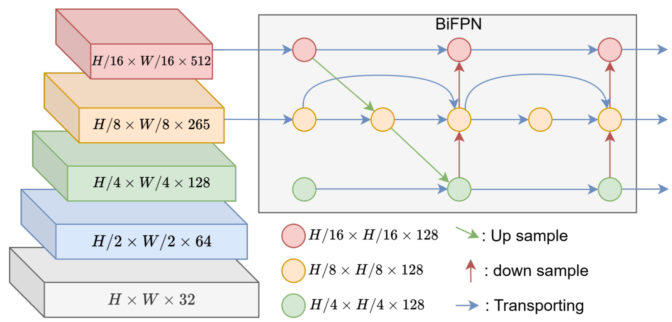


Fig. 6. Bi-directional feature pyramid network structure for multiscale feature fusion.

When connected to BiFPN's Layer 3, the DME layer initially handles data at a resolution of $H/16 \times W/16$, where H and W represent the height and width of the original input. Within this stage, the information undergoes resizing through transposed convolutions, refinement with 3x3 convolutions, and activation through ReLU. Each step generates outputs with 128 channels, which are subsequently concatenated with the next lower BiFPN layer. This ensures a tight and effective interplay between layers. The transposed convolution contributes to restoring dimensions toward the original input size, while the 3x3 convolution fine-tunes the details, yielding a rich diversity of feature representations.

This layered procedure repeats as the process advances through the subsequent BiFPN layers, moving from top to bottom. By carefully calibrating each stage and incorporating intelligent concatenation from higher to lower layers, the model retains a high degree of adaptability. This facilitates two-way interaction between layers and supports the production of a high-quality density map suitable for counting objects in crowded scenes, regardless of their varying scales.

After processing through three sequential layers, the result is a feature map comprising 256 channels with dimensions of $H/4 \times W/4$. To adjust this to the size of the original input image ($H \times W$) and reduce the channel count to a single output, we apply two rounds of transposed convolution followed by convolution, as illustrated in Fig. 7. Finally, the ReLU activation function is employed at the last convolutional layer since density map values are inherently non-negative. This process enables the DME to produce high-resolution density

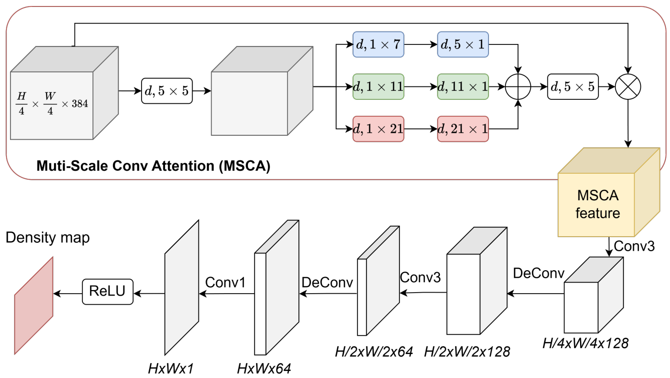


Fig. 7. Density map estimation module.

maps matching the input dimensions, effectively combining critical information extracted from the BiFPN block.

Our contribution is a task-specific integration for prawn postlarvae: (1) a size- and aspect-ratio–conditioned *elliptical* Gaussian density in which $\sigma_x = \alpha\, w$ and $\sigma_y = \beta\, h$ are tied to each larva's annotated width w and height h, preserving one-count mass per instance; and (2) a size-prior–guided multiscale fusion module that injects a pixelwise "size prior map" into the backbone to modulate attention and receptive fields. Unlike kNN geometry-adaptive kernels that set σ via neighbor distances (a proxy for density), our kernel is conditioned on object geometry; unlike appearance-only scale attention, our fusion is explicitly guided by the size prior. This integration targets transparency, rapid motion, and 3D distribution typical of prawn postlarvae.

C. Loss Function

The model optimizes its parameters based on a custom combination of loss functions that guide the training process by quantifying the difference between the predicted and ground truth density maps. For our proposed P2CountNet, a combination of the Mean Squared Error (MSE), pixel-level estimation error, and Normalized Cross-Correlation (NCC) is employed to balance local accuracy with global structural alignment in the density maps. The first error metric adopted in our counting loss is the MSE, defined by

$$\mathcal{L}_{\text{MSE}} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \left(D_{\text{pred}}(x, y) - D_{\text{gt}}(x, y) \right)^2, \quad (3)$$

where $D_{\text{pred}}(x, y)$ is the predicted density at pixel (x, y), $D_{\text{gt}}(x, y)$ is the ground truth density at pixel (x, y), and $W \times H$ represents the total number of pixels in the image. The second metric is the pixel-level estimation error defined as follows

$$\mathcal{L}_{\text{PEE}} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \left| D_{\text{pred}}(x, y) - D_{\text{gt}}(x, y) \right|. \quad (4)$$

The third loss component, Normalized Cross-Correlation (NCC), is introduced to encourage the model to focus on the structural alignment between the predicted and ground truth density maps. The NCC loss is defined as:

$$\mathcal{L}_{\text{NCC}} = 1 - \frac{\sum_{x=1}^{W} \sum_{y=1}^{H} D_{\text{pred}}(x, y) \cdot D_{\text{gt}}(x, y)}{\sqrt{\sum_{x=1}^{W} \sum_{y=1}^{H} D_{\text{pred}}^2(x, y)} \cdot \sqrt{\sum_{x=1}^{W} \sum_{y=1}^{H} D_{\text{gt}}^2(x, y)}}. \quad (5)$$

The total objective function for training P2CountNet is a weighted combination of the MSE, Density Estimation Error, and NCC losses:

$$\mathcal{L}_{\text{Total}} = \lambda(\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{PEE}}) + (1 - \lambda)\mathcal{L}_{\text{NCC}}, \quad (6)$$

where λ is a weighting coefficient for $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{DEE}}$, which is set to 0.9.

IV. Experiments and Evaluation

A. Acquisition of Datasets

In this research, a comprehensive dataset was curated to evaluate the performance of P2CountNet in counting prawn postlarvae within aquaculture environments. This dataset was obtained through systematic data collection efforts conducted at the Department of Biological Sciences, University of North Texas, Denton, TX, USA. The collection period spanned from May to July 2024, ensuring various environmental conditions representative of real-world aquaculture settings.

The dataset was created using postlarvae of the species *Macrobrachium rosenbergii*, specifically individuals 10, 20, and 30 days old. These different age groups allowed us to capture a range of sizes and developmental stages, which are essential for evaluating the model's capability to handle variability in object size and appearance. The animals were carefully captured and placed in a white-bottom bucket with approximately 1 inch of water in depth. This setup provided a uniform background, facilitating clear visibility of each individual and aiding in the accuracy of the annotations.

A 5 megapixel (MP) camera was used to capture high-resolution images, with each image having a resolution of 2560×1920 pixels. The camera was mounted directly above the tank, positioned 0.8 m away from the horizontal plane to achieve an overhead view of the postlarvae. This configuration allowed for clear visualization of the postlarvae while minimizing distortion. The camera transmitted the acquired images and videos to a computer and the cloud for storage via a network cable, ensuring secure and efficient data handling.

To enhance the model's robustness, images were taken under controlled conditions and in diverse lighting scenarios. To replicate real-world variability, images were captured under different lighting conditions, including natural sunlight and artificial lighting, and in water of varying clarity, from clear to turbid. This diversity is crucial for developing a model capable of generalizing across the range of environmental conditions encountered in prawn farming.

The dataset includes images with varying population densities of prawn postlarvae, from sparse distributions to highly crowded scenes. Figure 8 shows sample images illustrating the diversity of postlarvae distributions and densities used for model evaluation. This variation in population densities provides a valuable range for assessing the model's ability to handle overlapping instances and different postlarvae densities. Table I provides a summary of the number of images categorized by the prawn postlarvae samples used in the experiment.

To ensure that the model is trained effectively, precise and consistent annotations were necessary. The annotation process was meticulously conducted, with each larva in an image being carefully box-marked using CClaberler software as in the NWPU-Crowd dataset [42]. CClaberler annotation software was used to mark each larva's exact location and size, allowing for detailed ground truth data. Annotators, trained in identifying prawn postlarvae even in challenging conditions, provided the following annotations for each larva: (1) the centroid coordinates c_i to denote the larva's center position, and (2) the height h_i and width w_i to capture the larva's size. These dimensions were measured by drawing bounding boxes around each larva to represent its spatial extent. Each image was reviewed carefully, and any discrepancies were resolved through a consensus process or consulting with aquaculture experts. This rigorous quality assurance process contributed to creating a reliable dataset and enhancing the quality of training data available to the model. Figure 9b illustrates an example of annotation on the prawn postlarvae.

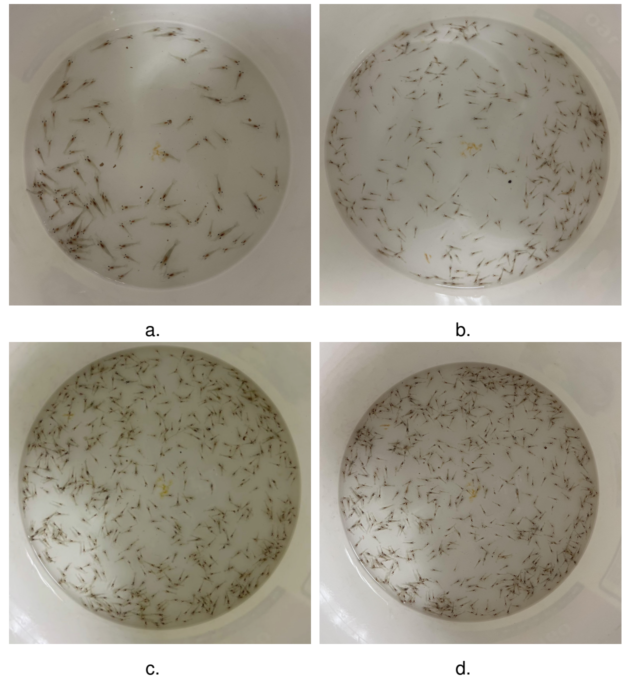


Fig. 8. Sample image of prawn postlarvae with (a) 137 prawns, (b) 308 prawns, (c) 752 prawns, and (d) 797 prawns.

TABLE I
QUANTITY OF PRAWN POSTLARVAE UTILIZED IN THE STUDY

Number of prawn postlarvae	Number of images
fewer than 100	647
100 to 200	473
200 to 300	368
300 to 400	294
400 to 500	170
500 to 600	82
more than 700	45
Total	2079

The dataset was randomly shuffled and split into training (1,455 images, 70%), validation (249 images, 15%), and test sets (249 images, 15%) to ensure objective model evaluation and prevent overfitting. The training set develops the model, the validation set tunes hyperparameters and applies early stopping, and the test set assesses the final performance. Each subset retains a representative distribution of postlarvae densities, sizes, and environmental conditions for reliable results.

Data augmentation techniques were applied to the training set to enhance data diversity and improve the model's generalization ability. Geometric transformations, including random rotations (up to ±15 degrees), horizontal and vertical flipping, and scaling adjustments (up to ±10%), were used to simulate different orientations and perspectives. When applying rotations, the bounding box coordinates were rotated correspondingly to maintain proper alignment with the transformed objects, ensuring that annotations remained accurate. This transformation was achieved by applying the same rotation matrix to both the image and its corresponding bounding boxes. Additionally, color jittering techniques were used to adjust brightness, contrast, and saturation levels, mimicking varying lighting conditions in real-world prawn farming environments. Gaussian noise was also introduced to simulate sensor noise and water turbidity effects, further increasing data variability. To ensure a balanced dataset, each original training image underwent augmentation, generating two additional transformed images per instance. As a result, the training set expanded from 2,079 original images to 6,237 images. These augmentation techniques significantly improved the robustness of the model by introducing variations in the training data, enhancing its ability to generalize to unseen images.

B. Evaluation Metric

To quantitatively assess the model's performance, we employ two standard evaluation metrics widely used in crowd counting tasks: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics provide insights into how well the predicted counts and density maps generated by the model match the ground truth values. MAE is expressed by the following formula:

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \hat{y}_i \right|, \tag{7}$$

where N is the total number of test images, y_i represents the ground truth count for the ith image, and $\hat{y}_i$ represents the predicted count for the ith image, obtained by summing the predicted density map over all pixels. MAE quantifies the average absolute difference between predicted and actual values of the crowd size.

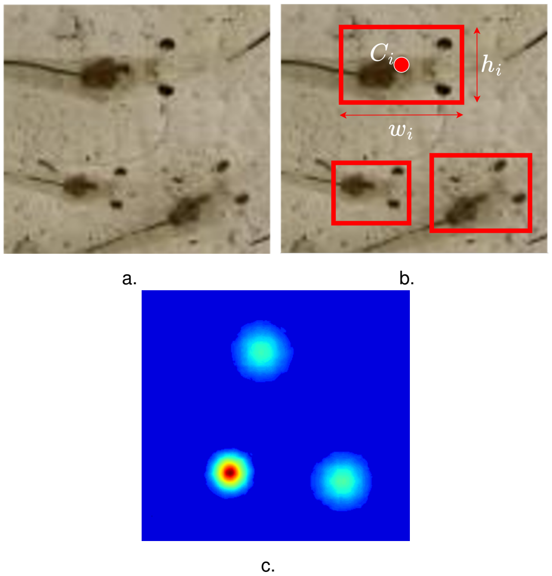


Fig. 9. Example annotations in the prawn postlarvae dataset for (a) original image, (b) labeled image, and (c) density map produced using a size-adaptable Gaussian kernel.

The RMSE is another commonly used metric in regression and counting tasks. Unlike MAE, RMSE penalizes larger errors more heavily due to the squaring operation. It emphasizes the impact of outliers or instances where the model makes particularly large mistakes, making it useful for identifying cases where the prediction is far from the ground truth. RMSE is computed using the formula in the following:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{8}$$

This formula calculates the average squared difference between predicted and actual values. The incorporation of both MAE and RMSE 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.

Each image was resized to a fixed size of 1280×720 for training and testing. Our method was pre-trained on ImageNet using the Adam optimizer. In the training stage, we used an Nvidia RTX 4090 GPU to implement our model and the stochastic gradient descent (SGD) to optimize our model parameters with a batch size of 64. We set the momentum to 0.9 and the weight decay to 0.0005. Initially, the learning rate was set to 0.05 and decreased by a factor of 0.1 every 50 epochs. We used several data augmentation techniques during training, including rescaling, cropping, flipping, and Gaussian blurring, to improve the accuracy. Finally, we evaluated our model's accuracy on the test dataset.

C. Evaluation and Comparison

The average accuracy of P2CountNet was evaluated across several test sets containing varying numbers of prawn postlarvae. As shown in Table II, the model consistently achieved high accuracy across different test conditions, with the accuracy slightly decreasing as the number of prawn postlarvae increased. For the testing image containing 233 postlarvae, P2CountNet reached an accuracy of 98.72%, reflecting its high accuracy in low-density scenarios. As the number of postlarvae increased to 352 and 473, the model maintained strong performance with accuracies of 97.08% and 96.73%, respectively.

When tested on images with 502 and 683 postlarvae, the accuracy slightly dropped to 95.94% and 95.88%, indicating the model's ability to manage moderate to high-density prawn postlarvae populations effectively. Finally, for the test set with 780 postlarvae, P2CountNet achieved an accuracy of 94.32%, which is still very competitive despite the increased complexity caused by higher densities. Overall, the model demonstrated robust and consistent performance across different prawn postlarvae counts, proving its reliability for real-world applications.

TABLE II
AVERAGE ACCURACY OF THE PROPOSED MODEL FOR EACH TEST SET

Number of prawn postlarvae	**Number of images**	**Average accuracy (%)**
23	32	100.00%
57	29	99.38%
86	34	98.51%
143	24	98.93%
233	30	98.72%
352	27	97.18%
473	25	96.73%
502	24	95.94%
683	19	95.88%
780	16	94.32%

To evaluate the performance of P2CountNet, we conducted comprehensive experiments comparing it with several state-of-the-art methods for crowd counting and density estimation. The methods selected for comparison include Context-Aware Network (CAN) [43], Multi-Column Convolutional Neural Network (MCNN) [16], Congested Scene Recognition Network (CSRNet) [44], Coordinated feature fusion network (CFFNet) [45], YOLOV8 [46], YOLOV12 and Scale Aggregation Network (SANet) [47]. All training experiments were conducted on a workstation equipped with an NVIDIA RTX 4090 GPU (24 GB GDDR6X memory), Intel i9 13900k CPU @ 4.5 GHz, and 128 GB DDR5 RAM, running Ubuntu 22.04 with CUDA 12.1 and cuDNN 8.9. This setup ensured sufficient computational capacity for large-batch training, efficient parallel data loading, and stable convergence of deep neural networks.

Figure 10 presents the accuracy comparison of the predicted density heatmap of prawn postlarvae with fewer than 1,000 prawn postlarvae for both our proposed P2CountNet and the SANet model. Table III illustrates the accuracy comparisons of prawn postlarvae counting across different models, including YoloV8, SANet, CRSNet, and our proposed model, for test cases containing 223, 502, 780, and 989 prawn postlarvae. As shown, our model consistently provides the most accurate predictions, with the closest counts to the ground truth in all cases.

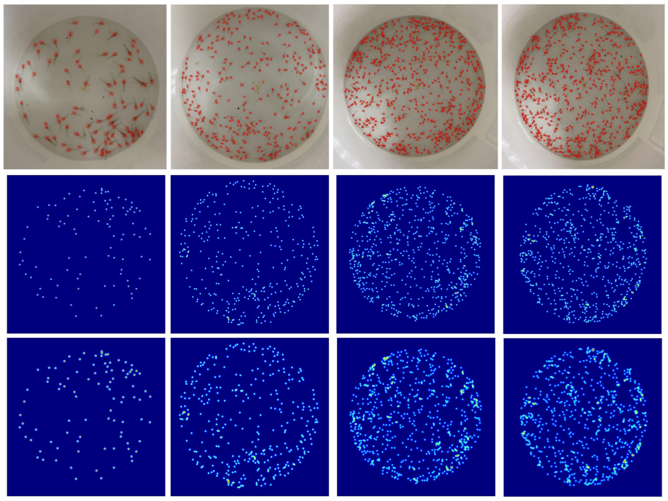


Fig. 10. Comparison of P2CountNet's and ground truth's predicted density heatmaps for the prawn postlarvae counts of 223, 502, 780, and 989 (a–d).

TABLE III
ACCURACY COMPARISONS OF PRAWN POSTLARVAE COUNTING

Number of prawn	223	502	780	989
YoloV8 [46]	192	429	620	810
YoloV12	195	437	613	817
SANet [47]	209	473	605	797
CRSNet [48]	213	486	612	961
Our Model	220	497	764	981

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT METHODS

Method	MAE			RMSE		
	Fix	k-NN	Scale	Fix	k-NN	Scale
CAN [43]	14.44	17.88	10.37	17.52	22.18	13.05
CFFNet [45]	19.99	27.45	13.91	23.71	32.38	17.18
CSRNet [44]	12.06	17.67	13.23	14.58	22.33	16.76
SANet [47]	10.77	36.30	11.91	13.60	43.05	14.59
MCNN [16]	11.55	19.89	11.08	14.61	24.74	13.73
Our Model	12.59	18.99	9.81	16.05	23.60	12.77

To evaluate the effectiveness of different density estimation techniques, we examined three approaches: a fixed Gaussian kernel (Fix), a dynamic sigma kernel based on k-nearest neighbors (k-NN), and our proposed Size-adaptive Gaussian kernel (Scale). The Fixed Kernel (Fix) approach employs a constant sigma value for all objects. While computationally simple, this approach fails to accommodate the natural variations in object sizes. The k-nearest Neighbors (k-NN) method improved upon the fixed kernel by dynamically adjusting the sigma based on the average distance between k-nearest neighbors. This approach allows for some adaptability to local density variations, making it more effective in moderately dense regions. However, it still struggles to accurately capture individual object sizes, particularly in cases of extreme object overlap or significant size discrepancies. In contrast, our Size-adaptive Gaussian kernel method consistently outperformed the other approaches. By dynamically adjusting the kernel size based on each object's dimensions, this approach accurately preserves size variations within the dataset.

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 our size-adaptive Gaussian kernel's ability to capture object size variations and generate accurate density maps, along with the powerful feature extraction and attention mechanisms incorporated within our CNN architecture. Our proposed method achieved a MAE of 9.81, compared to 12.59 for the fixed-kernel approach, representing a 34.8% reduction in MAE. Similarly, the RMSE was reduced from 16.05 to 12.27, a 27.8% improvement. These reductions in error metrics highlight the superior ability of the size-adjustable kernel to accurately estimate the number of postlarvae, particularly in high-density environments where traditional fixed-kernel methods struggle to distinguish overlapping or closely spaced postlarvae.

To assess the feasibility of P2CountNet in real-world deployment, we benchmarked the per-image inference latency of our model and compared it with established baselines. All timings correspond to *single-image inference* (batch size = 1) on 2D frames. Unless otherwise specified, inputs are resized as shown in Table V. For each configuration, we discard the first 50 warm-up iterations and average over 200 timed iterations with device synchronization enabled; gradients are disabled during timing.

TABLE V
PER-IMAGE INFERENCE LATENCY ON SINGLE FRAMES.

Model	Input image size	Latency (ms)
P2CountNet	1920×1920	141.2
P2CountNet	1280×1280	47.5
CSRNet [44]	1920×1200	97.6
SANet [47]	1920×1200	114.1
YOLOv8n	640×640	29.7

As shown in Table V, P2CountNet achieves 47.5 ms per image at 1280×1280 resolution (∼21 FPS), which is sufficient for near real-time monitoring in aquaculture applications. At full resolution (1920×1920), the latency is 141.2 ms (∼7 FPS), which is acceptable for offline or semi-online analysis. By comparison, CSRNet and SANet exhibit slower inference at similar resolutions, while YOLOv8n demonstrates the fastest latency due to its lightweight detection-oriented design. These results indicate that P2CountNet strikes a balance between accuracy and efficiency, making it suitable for practical deployments where high-resolution monitoring is required.

YOLOv8n achieves $\sim$ 33.7 FPS at 640×640 (29.7 ms), which is generally considered *real-time* on modern GPUs. However, this number is not directly comparable to density/regression models at multi-megapixel inputs: (i) the YOLO timing uses a substantially *lower* input resolution, which reduces computational load; (ii) YOLO performs object detection with non-maximum suppression, whereas P2CountNet and density-based methods estimate fine-grained counts across dense, partially transparent, and occluded populations where higher spatial resolution is often required. In practice, real-time feasibility for counting depends jointly on input resolution, scene density, and acceptable accuracy; at 1280×1280, P2CountNet operates near real-time while preserving counting fidelity in dense aquaculture scenes.

D. Ablation Study and Analysis

1) Effectiveness of Proposed Model: To evaluate the impact of key architectural components on model performance. The study compares three configurations: (1) a fixed-size density map, (2) a modified VGG16 network without BiFPN, and (3) the full proposed model. The results, shown in table VI, demonstrate a consistent improvement in counting accuracy with each architectural enhancement. Notably, the accuracy increased from 90.48% using a fixed-size density map to 95.8% with the full proposed model, highlighting the effectiveness of integrating advanced feature extraction, density map generation, and attention mechanisms.

TABLE VI
ABLATION STUDY RESULTS FOR DIFFERENT MODEL CONFIGURATIONS

Model Configuration	MAE	RMSE	Counting Accuracy (%)
Fix-size density map	12.59	16.05	90.48%
VGG16 without BiFPN	11.94	14.79	92.17%
Our Model	9.81	12.77	95.8%

2) Study of Loss Function: To better understand the contribution of each loss function in P2CountNet, we conducted an ablation study to analyze how the model performs with different combinations of loss components. The goal of this study is to evaluate the individual and combined impact of the MSE, PEE, and NCC losses on model performance. By systematically removing one or more of these components, we can measure their importance in achieving accurate and reliable prawn postlarvae counting. The quantitative results of the ablation study are summarized in table VII.

The proposed combination loss performs the best because it effectively balances pixel-level accuracy with structural consistency. MSE ensures the minimization of significant pixel-wise errors, PEE contributes to a balanced approach to small and large discrepancies, and NCC maintains the structural integrity of the density map, especially in challenging regions.

TABLE VII
ABLATION STUDY RESULTS FOR DIFFERENT LOSS CONFIGURATIONS

Loss configuration	MAE	RMSE
$\mathcal{L}_{\text{MSE}}$ only	12.80	16.92
$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{PEE}}$	10.70	14.83
$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{NCC}}$	10.55	14.65
Our Loss (Using eq. 6)	9.81	12.77

3) cross-dataset study on crowd counting: To assess generality beyond aquaculture, we conducted a cross-dataset study on ShanghaiTech Part B, retraining MCNN [16], CSRNet [44], SANet [47], and our model under identical protocols (same resizing, 512×512 random crops, horizontal flip, Adam with 1×10^{-4} LR, batch size 8, 300 epochs). Evaluation follows the standard MAE/RMSE on the test set with counts obtained by integrating predicted density maps. Our full model achieves 10.39/13.61 (MAE/RMSE), outperforming CSRNet (13.51/17.24) and SANet (11.16/13.93).

TABLE VIII
CROSS-DATASET COMPARISON ON SHANGHAITECH CROWD COUTING DATASET

Method	**MAE**	**RMSE**
MCNN [16]	**11.35**	**14.92**
CSRNet [44]	**13.51**	**17.24**
SANet [47]	**10.16**	**13.93**
Our model	**10.39**	**13.61**

V. CONCLUSION

This paper presents a novel deep-learning model for accurately counting prawn postlarvae in aquaculture environments. Leveraging a Size-Adjustable Gaussian Kernel and a Bidirectional Multiscale Feature Pyramid Network, our approach effectively addresses the challenges of counting small, overlapping objects with significant size variability. Through extensive evaluation of a diverse, self-labeled prawn postlarvae dataset, P2CountNet demonstrated superior performance across varying population densities and achieved high accuracy and robustness in complex, high-density scenarios. Our model's consistent performance across different prawn postlarvae counts and densities highlights its potential for real-world aquaculture applications. By providing a reliable and efficient solution for postlarvae counting, P2CountNet can aid in optimizing prawn farming operations, enabling better management of stocking densities, resource allocation, and yield predictions. These improvements contribute to both the economic efficiency and sustainability of aquaculture practices.

Despite the strong overall performance, the proposed method shows degradation in several challenging scenarios.

In particular, frames containing more than 1000 prawns create extreme-density conditions where heavy occlusion, kernel overlap, and local peak merging lead to undercounting or over-smoothing. Additional failures arise in low-visibility conditions such as water turbidity, bubbles, or debris, where distractors produce spurious responses in the density map. Illumination variability, glare, and motion blur further suppress fine structural cues, while scale or viewpoint mismatches push larvae sizes beyond the trained priors. These factors collectively highlight the limits of robustness in uncontrolled aquaculture environments and motivate future work on domain adaptation and uncertainty-aware counting models.

In future work, we aim to further enhance the model's adaptability by exploring additional augmentation techniques and expanding the dataset to include different aquatic species and environmental conditions. Moreover, integrating P2CountNet with real-time monitoring systems on hardware like NVIDIA Jetson platforms could enable live tracking and counting in aquaculture facilities, providing a complete solution for intelligent aquaculture management.

References

[1] FAO, "The state of world fisheries and aquaculture 2024," *Food and Agriculture Organization*, vol. 2024, pp. 1–244, 2024.

[2] J. H. Primavera, "Tropical shrimp farming and its sustainability," in *Tropical mariculture*. Elsevier, 1998, pp. 257–289.

[3] M. L. Pazmiño, L. Chico-Santamarta, A. Boero, and A. D. Ramirez, "Environmental life cycle assessment and potential improvement measures in the shrimp and prawn aquaculture sector: A literature review," *Aquaculture and Fisheries*, 2024.

[4] X. Li, R. Liu, Z. Wang, G. Zheng, J. Lv, L. Fan, Y. Guo, and Y. Gao, "Automatic penaeus monodon larvae counting via equal keypoint regression with smartphones," *Animals*, vol. 13, no. 12, p. 2036, 2023.

[5] Y. Duan, L. H. Stien, A. Thorsen, Ø. Karlsen, N. Sandlund, D. Li, Z. Fu, and S. Meier, "An automatic counting system for transparent pelagic fish eggs based on computer vision," *Aquacultural engineering*, vol. 67, pp. 8–13, 2015.

[6] T. T. N. Thai, T. S. Nguyen, and V. C. Pham, "Computer vision based estimation of shrimp population density and size," in *2021 International symposium on electrical and electronics engineering (ISEE)*. IEEE, 2021, pp. 145–148.

[7] 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.

[8] 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.

[9] H. Yu, X. Liu, H. Qin, L. Yang, and Y. Chen, "Automatic detection of peeled shrimp based on image enhancement and convolutional neural networks," in *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, 2022, pp. 439–450.

[10] 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.

[11] 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.

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

[13] 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.

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

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

[16] 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.

[17] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8362–8371.

[18] X. Jiang, L. Zhang, M. Xu, T. Zhang, P. Lv, B. Zhou, X. Yang, and Y. Pang, "Attention scaling for crowd counting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4706–4715.

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

[20] Y. Miao, Z. Lin, G. Ding, and J. Han, "Shallow feature based dense attention network for crowd counting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 765–11 772.

[21] 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 *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 2382–2389.

[22] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[23] H. Gong, T. Mu, Q. Li, H. Dai, C. Li, Z. He, W. Wang, F. Han, A. Tuniyazi, H. Li *et al.*, "Swin-transformer-enabled yolov5 with attention mechanism for small object detection on satellite images," *Remote Sensing*, vol. 14, no. 12, p. 2861, 2022.

[24] A. Naseer, E. N. Baro, S. D. Khan, and Y. Vila, "A novel detection refinement technique for accurate identification of nephrops norvegicus burrows in underwater imagery," *Sensors*, vol. 22, no. 12, p. 4441, 2022.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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, p. 222, 2022.

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[38] 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.

[39] F. Kamalov, “Kernel density estimation based sampling for imbalanced class distribution,” *Information Sciences*, vol. 512, pp. 1192–1201, 2020.

[40] A. G, N. K. S, S. S. V, and S. S, “Density map based estimation of crowd counting using vgg-16 neural network,” in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, 2023, pp. 704–708.

[41] Y. Zhang, S. Choi, and S. Hong, “Spatio-channel attention blocks for cross-modal crowd counting,” in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, December 2022, pp. 90–107.

[42] Q. Wang, J. Gao, W. Lin, and X. Li, “NWPU-crowd: A large-scale benchmark for crowd counting and localization,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 6, pp. 2141–2149, 2020.

[43] 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.

[44] 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.

[45] 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.

[46] D. Reis, J. Kupec, J. Hong, and A. Daoudi, “Real-time flying object detection with YOLOv8,” *arXiv preprint arXiv:2305.09972*, 2023.

[47] 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.

[48] 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.

Thai-Ha Dang received the B.Eng. degree from Hanoi University of Science and Technology (HUST), Hanoi, Vietnam, in 2021, and the M.S. degree from Pukyong National University (PKNU), Busan, South Korea, in 2023. He is currently pursuing the Ph.D. degree in electrical engineering at the University of North Texas (UNT), Denton, TX, USA. His research interests include signal processing, computer vision, machine learning, and embedded systems.

Fabio Dos Santos Neto received a Bachelor of Science in Marine Biology degree from Hawaii Pacific University, Honolulu, HI, USA, in 1999, a Master of Professional Science in Aquaculture degree from the University of Miami, Miami, FL, USA, in 2023 and is currently a Ph.D. candidate in Environmental Science at the University of North Texas, Denton, TX, USA. He has 25 years of professional experience in Aquaculture (R&D and production). His current research focuses on sustainable shrimp aquaculture production using Brackish water desalination concentrations that integrate sensors, controls, and ecotoxicology.

Xinrong Li (Senior Member, IEEE) received his Bachelor of Engineering degree from University of Science and Technology of China (USTC), Hefei, China, in 1995, Master of Engineering degree from National University of Singapore (NUS) in 1999, and Ph.D. degree from Worcester Polytechnic Institute (WPI), Worcester, MA, USA, in 2003, all in electrical engineering. He is currently a Professor in the Department of Electrical Engineering at University of North Texas (UNT), Denton, TX, USA. His research interests include signal processing, machine learning, embedded sensor systems, wireless communications, and wireless networks. His research has been supported by Army Research Laboratory, National Science Foundation, National Institute of Food and Agriculture, North Central Texas Council of Governments, and a number of industrial partners.

Breana Smithers received her B.S. and M.S. degrees in Kinesiology from the University of North Texas, Denton, TX in 2014 and 2016. After developing her M.S. thesis on a prototype of pulse wave velocity sensing using microcontrollers, she has been performing interdisciplinary research with Electrical Engineering (EE). Since 2014, she has been a Research Assistant, Research Coordinator, and now Research Scientist with the EE department. She works on research grants involving desalination systems, renewable power, and environmental monitoring. Her current projects relate to improving crop yield and soil salinity by integrating renewable power, sensors, controls, and ecotoxicology with decoupled aquaponics using brackish groundwater and desalination concentrate for sustainable food production.

Edward Mager received a Bachelor of Science in Biology degree from Saint Louis University, St. Louis, MO, USA, in 1995, a Master of Science in Biotechnology degree from Northwestern University, Evanston, IL, USA, in 1996, and Ph.D. degree from the University of Miami, Miami, FL, USA, in 2010. From 2010 to 2016, he was a Postdoctoral Associate and Assistant Scientist at the University of Miami. He has been working in the Department of Biological Sciences, University of North Texas, Denton, TX, USA, as an Assistant Professor since 2016 and an Associate Professor since 2023. His research has been focused on the aquatic toxicology of environmental pollutants, the ecophysiology of aquatic organisms and aquaculture.

Miguel Acevedo received the B.S. and M.S. degrees in Electrical Engineering (EE) from the University of Texas at Austin in 1971 and 1972, the MEng in EE and Computer Science (1977), as well as the Ph.D. in Biophysics from the University of California, Berkeley. He has been with the University of North Texas (UNT) for over 30 years where he is currently a Regents Professor in the EE department. Before joining UNT, he was at the Universidad de Los Andes, Merida, Venezuela, where he served in the School of Systems Engineering. He has received numerous research grants as well as written many journal articles, book chapters, and proceeding articles. He has written four textbooks for CRC Press, Data Analysis and Statistics for Geography and Environmental Science and Engineering, Simulation of Ecological and Environmental Models, Real-time environmental monitoring: sensors and systems (Second Edition), and Introduction to renewable power systems and the environment with R.