Received: November 21, 2024. Revised: December 29, 2024.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Compact *Litopenaeus Vannamei* Post-Larvae Fry Counting Network with Optimized Scale Aggregation Network

Zulfikar Davbi Mahendra Fasya[1]* Agus Indra Gunawan[1] Bima Sena Bayu Dewantara[2]

[1]*Department of Electrical Engineering, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia*
[2] *Department of Informatics and Computer Engineering,*
Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia
* Corresponding author's Email: fikardavbi12@gmail.com

Abstract: Knowing the number of *Litopenaeus vannamei* post-larvae fry (PL) is vital in the proliferation process. Traditional methods employ a small spoon for counting which is labor-intensive and has poor accuracy resulting in an uncertain number of biomasses. Slow counting process also contributes to fry hypoxia due to excessive fry contact. This underscores the needs of a compact, high-speed and high-accuracy PL counting network. This paper introduces an Optimized Scale Aggregation Network (OSA-Net), a compact fry counting network based on density map regression, designed for edge devices with a small parameter size of 660 KB. Squeeze-and-Excitation Network embeds the channels pruned network backbone to compress less contributed channels. The model trained with Local Pattern Consistency Loss combined with Euclidean Loss to enhance predicted density map quality. Trained on the *Politeknik Elektronika Negeri Surabaya Litopenaeus Vannamei one* (PENSLV-1) dataset, proposed model obtained high accuracy with Mean Absolute Error (MAE) of 1.99 and Mean Squared Error (MSE) of 2.69 which indicate superior counting effectiveness in under different density levels and PL sizes.

Keywords: Density maps estimator, Automatic fry counting, Channel attention mechanism, Scale aggregation network, *Vannamei* post-larvae counting.

1. Introduction

In recent years, the demand for precise and rapid fry counting equipment has significantly increased among various aquaculture operations, particularly in white leg shrimp PL hatcheries [1, 2]. This is primarily driven by the small size of PL, which typically ranged from 0.45 to 0.55 cm for PL aged 5 days [3]. Traditionally in Indonesia, small hatcheries have relied on manual methods, such as using small ladles to count fry in batches, often measured in ethnomathematics unit called "rean," where one rean is roughly equivalent to 5,000 fries [4-6]. However, such methods are prone to inaccuracies and potential fraud, as the actual number of fries per rean can vary significantly. This variability can lead to inaccurate biomass estimations, resulting in overfeeding, deteriorating water quality, and increased operational inefficiencies [7]. Moreover, manual counting is labour-intensive, time-consuming, and can cause significant mortality due to hypoxia from excessive handling of PL [8]. Consequently, there is an urgent need for high-accuracy, non-invasive, and high-speed counting methods to address these challenges.

The computer vision approach strategically solved the challenges, providing a middle range pricing, easy to apply, and applicable to different platforms [9, 10]. However, the approaches may struggle to handle the small size of PL and occlusion scenarios in dense settings. Density maps regression methods have proven to solve counting small object challenges in fry counting applications by representing features in density distribution offering simpler deep learning networks and enabling overlap tolerance capabilities [11-14]. Several fry counting methods incorporating deep learning and density maps have been employed prior to this research. Li

et al [8], introduced MSENet that utilizes multi-column convolutional neural networks (MCNN) [14] incorporated with squeeze and excitation networks (SE-Net). Multiple SE-Nets are embedded in every branch of MCNN to enhance extraction features while keeping the model lightweight. Multi-column architecture addressed a variety of sizes of fries in real applications achieving 3.33 in MAE and 0.1 MB of parameters. Qu et al [15], proposed shrimp larvae counting network with overlapping splitting and Bayesian-dm-count loss (SLCOBNet). Two vision transformer architecture (Twin-SVT) embeds with feature pyramid aggregation for richer context of information are employed. The feature inputs are through overlapping splits operation into fixed-shape size blocks. Bayesian-DM-Count-Loss is used for training the models and achieved 3.27, 3.61, and 1.28 in MAE for slight noise, turbid, and dark lighting counting conditions respectively. Liu et al [16], modify Congested Scene Recognition (CSRNet) to achieve highly accurate PL counting called Shrimpseed_Net. Six large 512×512 convolutional layers in the CSRNet are removed, 128 dilated convolutional layers are added at the back end to tolerate layer removal. Layers modifications resulting 133% inference speed compared to original CSRNet while maintaining counting performance. Hsieh et al [17], utilize modified Scale Aggregation Networks architecture to count PL called ShrimpCountNet. Since scale aggregation network (SANet) [18] applications are focused on crowd counting, modification of the network to adapt with PL features is employed. Tensor decomposition networks (TedNet) are applied prior to each scale aggregation module (SA-Module) operations in ShrimpCountNet feature map encoder. The addition enables the network to extract deeper features in PL counting scenarios achieving performance of 6.54 points MAE improvement compared to SANet.

The research surveys explore deep learning techniques for fry counting scenarios, demonstrating their applicability and drawbacks. SLCOBNet [15], being not explicitly designed as a compact model, may face challenges in running efficiently on portable devices due to its inherently complex architecture. The Shrimpseed_Net [16] also has a long architecture. Moreover, they did not account for variations in PL sizes and density levels. MSENet [8] is not specifically designed for vannamei PL counting and the MCNN backbone architecture may limit the ability of multi scale leads to reduce in counting performance. The ShrimpCountNet [17] does not account for different sizes of post-larvae (PL). Additionally, the dataset is limited to 159 fries, which does not adequately represent dense counting scenarios.

Density maps regression method is widely used in crowd counting scenarios. CMTL [11], (CNN-based Cascade Multi-task Learning) integrates two tasks within a unified deep learning framework. The high-level prior task performs classification that enables the network to group features based on crowd density levels. Therefore, global features can be distinct based on coarse counting generated by the high-level prior. The high-level prior task is integrated with the density map estimator through a cascaded CNN architecture, enabling the network to effectively address a wide range of crowd density levels. Li et al [12], proposed CSRNet which consists of a front-end network to extract 2D features using architecture of Visual Geometry Group (VGG16), and dilated CNN layers for the back-end network. Dilated CNN enables the network to generate high quality density maps and expand the receptive field without losing resolution caused by the front-end network. SFCN [13], is built on top of ResNet101 and VGG as backbone before the regression layer. The approach improves crowd counting performance in challenging, real-world scenarios by incorporating supervised learning into the backbone architecture. Leveraging a mainstream backbone allows the network to be fine-tuned effectively for unseen data, enhancing its adaptability and generalization capabilities. MCNN [14], utilizes multi-column convolutional network architecture to extract multi-scale features on one single Image, so a lightweight model can easily be achieved.

Counting small objects such as PL and human heads in crowds may face many challenges, a density maps method found to be successful in crowd counting settings. However, in PL counting scenarios MCNN [14] will struggle with overlapping of PL and inaccurate counting in a variety of PL sizes due to its lack's aggregation architecture. CMTL [11] and SFCN [13], require a sufficient quality of datasets to enable effective learning, which is not applicable to fine-grained objects such as PL where underwater environments may vary significantly. CSRNet [12] and SFCN [13] architectures are utilizing complex backbones which increase computational complexity.

This paper introduces a PL counting network called optimized scale aggregation network (OSA-Net) based on density maps regression techniques. Designed for edge devices, OSA-Net addresses real application challenges such as fry adhesion and variety of density levels while keeping the compactness of the model. OSA-Net also considered

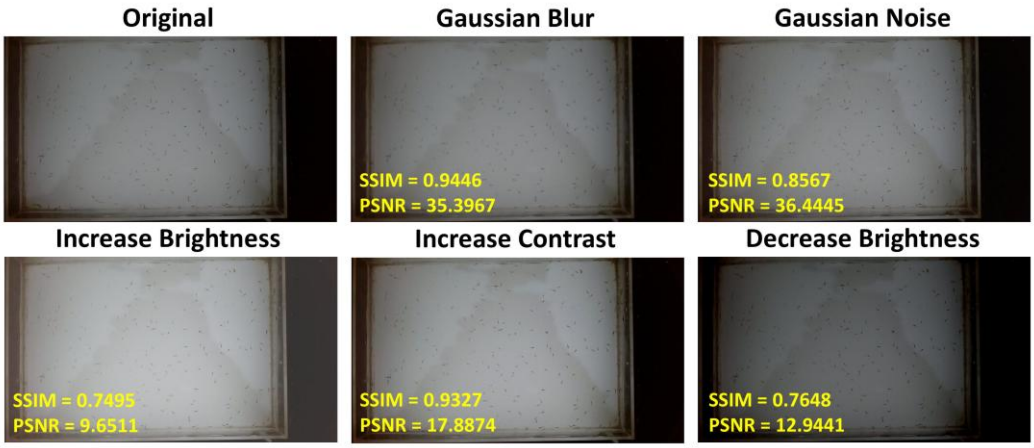

Figure. 1 PENSLV-1 dataset samples with PSNR and SSIM similarity value

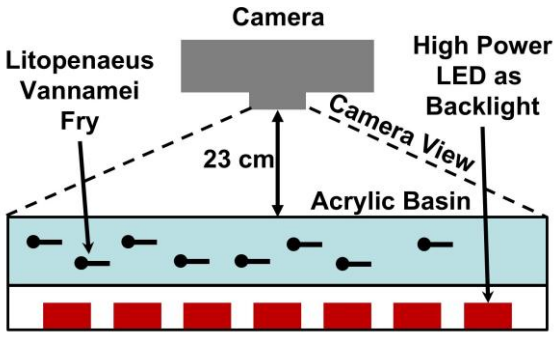

Figure. 2 Dataset capture scenario in custom made basin.

different fry sizes for training which are beneficial in counting different ages of PL. The main contributions can be shortened as:

1. A compact network called OSA-Net, designed for edge devices to count *Litopenaeus Vannamei* post-larvae (size of parameter: 0.66 MB).
2. Optimized scale aggregation module (OSA-Module) with few numbers of channels and squeeze-and-excitation networks embedding to generate more overlapping tolerance density maps (MAE: 1.99 and MSE: 2.69).
3. PENSLV-1 dataset consists of varied age and density of Litopenaeus *Vannamei* post-larvae to train model, and additional PENSLV-2 dataset to test model on unseen data.

The structure of this paper is as follows: Section 2 outlines the network architecture and describes the experimental configuration; Section 3 reports the results and provides a detailed discussion; and Section 4 concludes with a summary of the study.

2. Methodology

2.1 Dataset and device settings

This paper introduces *Politeknik Elektronika Negeri Surabaya Litopenaeus vannamei* one (PENSLV-1) dataset, consisting of 306 images of real living PL from small scale local hatcheries in Tuban, Indonesia. A custom basin was created using acrylic material to capture from a single top-down angle and additional backlight to address PL transparency, as shown in Fig. 2. The dataset includes real living PL fries of varying sizes and densities, ranging from 123 to 217 fries, aged 5 and 8 days, to simulate different counting scenarios under fixed camera conditions as shown in Fig. 1. Images are captured at a resolution of 960 × 540 pixels, down sampled from the original 1920 × 1080 pixels captured by a Logitech c920e camera. Variations in sunlight penetration and water turbidity can adversely affect the model's performance. To address the issue, data augmentation techniques such as Gaussian blur, Gaussian noise, contrast adjustment, and brightness modification are applied as illustrated in Fig. 1. The Structural Similarity Index Measure (SSIM) [19] and noise levels are assessed using the Peak Signal-to-Noise Ratio (PSNR) are employed to evaluate augmentation. The samples of PENSLV-1 provided at github.com/FIKARDAVBI/OSA-NET.

2.2 Proposed model

Number of factors are considered to design the model, first the speed of counting which requires a compact or lightweight model, second is generalization under different density levels and PL sizes. Lastly, fry adhesion that requires high quality predicted density maps.

Deep learning backbones such as AlexNet [20], VGG [21], ResNet [22], MCNN [14] and Scale Aggregation Networks (SANet) [18] combine various architectural improvements for image feature extraction and analysis. MCNN [14] utilizes multi-column CNNs architecture to scale features, so it covers a wide range of feature resolutions. AlexNet [20] introduced ReLU activation, GPU-accelerated

training, and massive convolutional filters, marking a breakthrough in deep learning applications. VGG [21], uses a stack of 16 layers including small 3×3 convolutional filters and max-pooling layers, enabling hierarchical and fine-grained feature extraction while keeping simplicity in design. ResNet [22] exploits residual connections to ease gradient flow and enable the training of very deep networks, ensuring efficient feature representation through residual learning. SANet [18] combines scale aggregation modules to capture and combine multi-scale features, effectively solving the issues of various object scales by combining local and global spatial information. Mentioned backbone architectures collectively provide powerful tools for applications needing robust image representation, such as density map regression.

Shrimp fry counting using density map regression requires a backbone capable of handling multi-scale and fine-grained features due to the varying size, orientation, and distribution of fry in images. Traditional architectures like AlexNet, VGG, and ResNet, while foundational in computer vision, face limitations in this context. AlexNet's limited depth and large filters hinder precise feature extraction, while VGG improves detail capture with smaller filters but suffers from high computational costs and lack of scale adaptability. ResNet addresses training challenges in deep networks with residual connections but struggles to effectively capture multi-scale features needed for accurate density maps. MCNN and SANet overcome these limitations by utilizing multi-scale inception architecture However, MCNN scaling is limited by the number of convolutional columns and makes the model less perform in unsupervised data. SANet integrates scale aggregation modules (SA-Module) that adaptively combine local and global features across multiple spatial scales, ensuring precise and robust density estimation. This makes SANet a superior choice for shrimp fry counting, as it excels in addressing the specific challenges of size variability and density distribution compared to traditional backbones with parameters only 0.91MB.

This research proposed optimized scale aggregation networks (OSA-Net) with SANet as the backbone as shown in Fig. 4. OSA-Net consists of two parts: Feature Map Encoder (FME) consists of Optimized SA Module (OSA-Module) and Density Map Estimator (DME). FME module is designed to capture a diverse range of feature resolutions, enabling it to effectively address variations of PL sizes and density levels. The DME module generates high-dimensional and high-resolution density maps to mitigate feature loss, particularly in scenarios where fry adhesion occurs. First, second, and third SA-Modules after input in SANet have 64, 128, and 128 channels respectively. However, In OSA-Net structure, OSA-Module channels are 32, 64, and 64 for first, second, and third respectively to reduce model complexity. Therefore, squeeze-and-excitation network (SE-Net) [23] introduced to OSA-Module to compensate channels trimmed which strengthen channels weight that have more feature information and compress channels weight with less contributions as shown in Fig. 4. Instance Normalization (IN) is used to enhance gradient descent of the model as SANet suggests. IN layers applied in each convolution and transposed convolutional layers.

2.2.1. Feature map encoder

FME employed stacking OSA-Modules separated by 2×2 max pooling layer as shown in Fig. 4. This design reduces the spatial dimensions by half at each pooling stage, which helps in downsampling the feature maps while preserving key information. OSA-Module is based on the inception network introduced by Szegedy et al [24], where branched convolutional columns as shown in Fig. 4 are stacked into one kernel with the number of channels that are equal between input and output. OSA-Modules utilize 1×1, 3×3, 5×5, and 7×7 convolutional layer branches to add more future dimension diversity as shown in Fig. 4. The inclusion of a 1×1 convolutional layer serves a dual purpose. First, it ensures that the information from the previous SA module is efficiently compressed and passed on, which helps in maintaining the feature contribution from earlier lay-

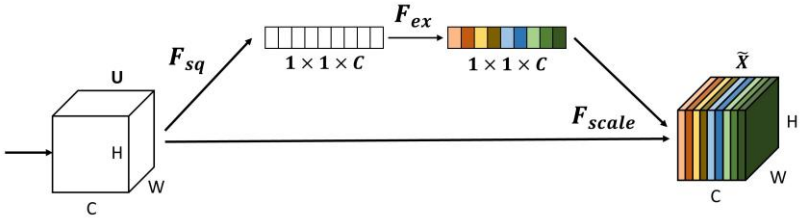

Figure. 3 Architecture of Squeeze-and-Excitation Network (SE-Net)

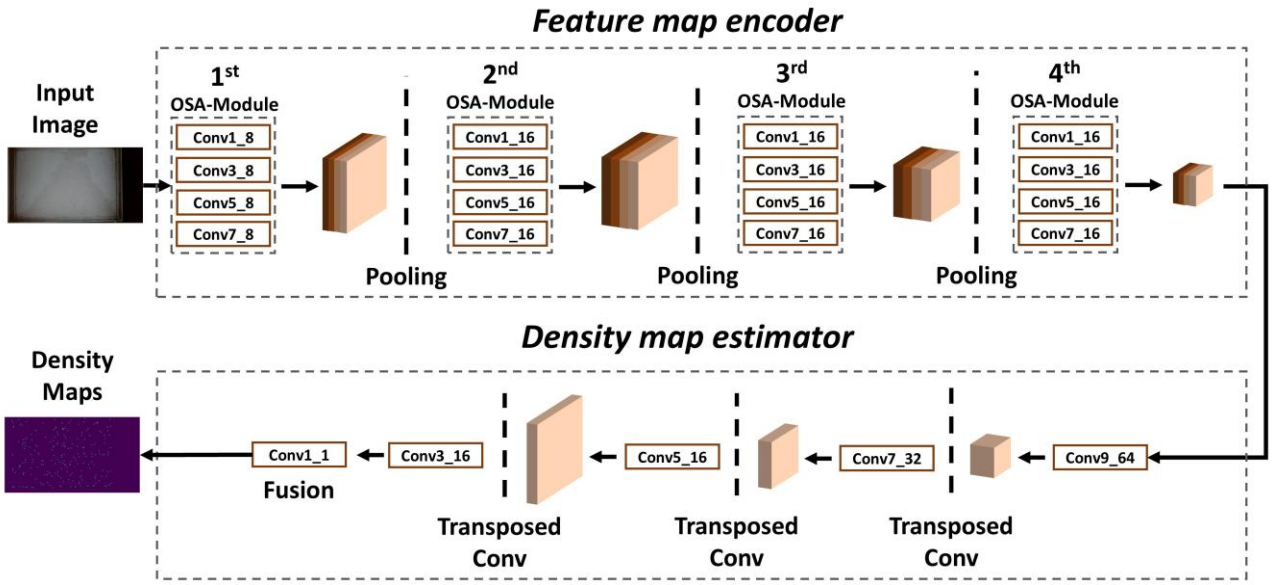

Figure. 4 OSA-Net architecture, convolutional layers dictate as “Conv(kernel size)_(channel size)

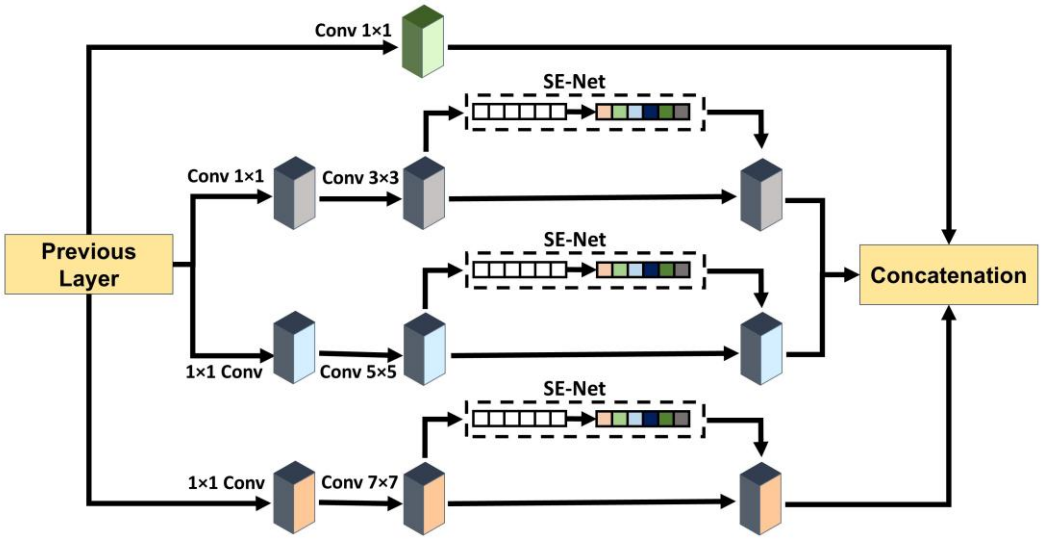

Figure. 5 Network architecture of optimized scale aggregation module (OSA-Module)

-ers. Second, it allows the network to maintain a consistent number of channels between the input and output of each OSA-Module. This design strategy is highly effective in increasing the receptive field under different PL sizes and density levels while keeping the computation efficient

$$z_c = F_{sq}(u_c) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \quad (1)$$

$$s = F_{ex}(z, \boldsymbol{W}) = \sigma\big(\boldsymbol{W_2}\gamma(\boldsymbol{W_1}z)\big) \quad (2)$$

$$\tilde{X}_C = F_{scale}(u_c, s_c) = u_c s_c \quad (3)$$

2.2.2. Attention mechanism

OSA-Net is applied with Squeeze-and-Excitation Network (SE-Net) depicted by Fig. 5 to reinforce channel wise information specifically to enhance model robustness under different density levels in PENSLV-1 dataset. It can increase generalization and feature learning without substantially increasing computational costs.

The "squeeze" operation (F_{sq}) captures global spatial information $z \in \mathbb{R}^C$ by applying global average pooling in each channel as explained by Eq.

(1). The "excitation" (F_{ex}) step learns the relationships and interdependencies between channels as explained by Eq. (2). First, transform each feature in z to $\boldsymbol{W_1} \in \mathbb{R}^{(C/r)\times C}$ dimensional. Second, apply nonlinearity γ and linearly transform it again to $\boldsymbol{W_2} \in \mathbb{R}^{C\times(C/r)}$ dimensional, which is going to reduce the dimension by the factor of r. Lastly, apply sigmoid function σ to get values between 0 and 1. Channel wise reweighting takes place by convolve u_c and s_c as described in Eq. (3). The architecture of SE-Net illustrated by Fig. 3.

2.2.3. Density Map Estimator

Generated feature maps from FME have low resolution and lost detail features information, leading to low counting accuracy. OSA-Net utilizes original DME from SANet inspired by Sindagi and Patel [25] but adding more layers to increase its performance as shown in Fig. 4. Multiple transposed convolutions and convolutions with filter size: 9×9, 7×7, 5×5, and 3×3 introduced to increase spatial resolutions by the factor of two in every layer. ReLU activation functions embed after every convolution, transpose convolution and at the end network layers since the value of density maps are always positive. Eventually, 1×1 convolutional layer take place as features fusion to predict density value at every point. High quality density maps generated by DME benefit dense PL counting which extracts futures more faithfully avoiding fail feature extractions.

Ground truth density maps require converting dots annotation in the dataset into density maps using adaptive Gaussian kernels. Whenever there are dots in pixels x_i, it is represented by Dirac's delta function $\delta(x\text{-}x_i)$ as described by Eq. (4). Then, each dot is convolved with adaptive Gaussian kernel $G_{\beta i}$ to generate ground truth Y. $G_{\beta i}$ can be determined by multiplying Gaussian kernels with the average distance $\bar{d}$ of the closest j fries relative to fry i as described by Eq. (5). Number the closest fries is limited by K. Empirically, α is equal to 0.1. The Integral of density maps result is the number of fries. This technique excels in situations where fry occlusion occurs by representing the fries as a continuous-valued density distribution, where the value of each pixel indicates clear likelihood of fries being present in that region as illustrated in Fig. 6.

$$Y = \sum_{i=1}^{H} \delta(x - x_i) * G_{\beta i}(x), with\ \beta_i = \alpha \bar{d}^i \quad (4)$$

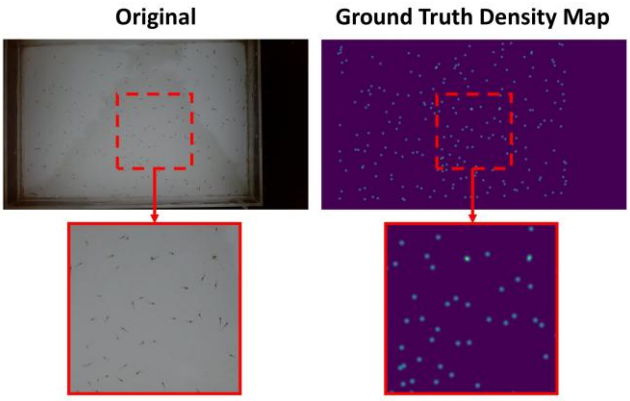

Figure. 6 Ground truth geometrically adaptive Gaussian kernel density maps

$$\bar{d}^i = \frac{1}{K}\sum_{j=1}^{K} d_j^i \quad (5)$$

2.3 Loss function

Precise density map dot's location representation enables the network to count better under different density levels avoiding fail features extraction where fry adhesion occurs. OSA-Net utilizes Local Pattern Consistency Loss (L_c) combined with Euclidean Loss (L_E), facilitating the network to evaluate based on statistics (mean, variance, and covariance). Those effective to evaluate the precision of generated density maps feature pattern structure. The total loss can be defined using Eq. (6), summing L_E and tuned L_c. In OSA-Net case, the value of λ_C (Lambda) is 0.0005 that will be discussed later.

$$L = L_E + \lambda_C L_C \quad (6)$$

2.3.1. Euclidean distance

For pixel wise error estimation, Euclidean Loss (L_E) is used to compare predicted density maps pixels with ground truth pixels which are described by Eq. (7). Where N is the number of pixels, F is the predicted density map and Y is the associated ground truth density map. The calculated inaccuracy in each pixel is summed up and normalized by the number of pixels in the image since the dimension can be different.

$$L_E = \frac{1}{N}\sum_{i=1}^{N}(F_i - Y_i)^2 \quad (7)$$

2.3.2. Local pattern consistency loss

Local Pattern Consistency Loss (L_c) utilizes SSIM, which compares similarity based on local

mean (μ_F), local variance (σ_F^2), and covariance (σFY) calculated by Eq. (8), Eq. (9), and Eq. (10) respectively in each x on predicted density maps F during training and Ground truth Y. Where, $L(p)$ is representing weight in p offset to the center, and $\boldsymbol{P}$ contains every kernel position. The result of 1 meaning the images are identical and -1 otherwise.

$$\mu_F(x) = \sum_{p \in \boldsymbol{P}} L(p) \cdot F(x+p) \tag{8}$$

$$\sigma_F^2(x) = \sum_{p \in \boldsymbol{P}} L(p) \cdot [F(x+p)^2 - \mu F(x)]^2 \tag{9}$$

SSIM then can be obtained point to point in density maps by using Eq. (11) where C1 and C2 being small numbers with the values following [19] to mitigate division by zero. L_C can be defined by averaging the SSIM value of each location x with the number of pixels N as described by Eq. (12).

$$SSIM = \frac{(2\mu_F + C_1)(2\sigma_{FY} + C_2)}{(\mu_F^2 + \mu_Y^2 + C_1)(\sigma_F^2 + \sigma_Y^2 + C_2)} \tag{11}$$

$$L_C = 1 - \frac{1}{N}\sum_x SSIM(x) \tag{12}$$

3. Experiment and discussion

3.1 Training settings

Training conducted in NVIDIA RTX 2060 SUPER with hyperparameters described by Table 1. Weight initialization is random with seed fixed for the same model generation. Normalization takes place for preprocessing the dataset before feed to the network.

3.2 Metrics and Evaluation

Mean Absolute Error (*MAE*) calculated using Eq. (13) and Mean Squared Error (*MSE*) calculated using Eq. (14) are applied to evaluate model fry counting performance. Where, YC is the ground truth numbers, FC is counting prediction and M is the number of samples. Determining model complexity is evaluated using size of parameters and Floating-Point Operations per Second (FLOPs).

The result of fry counting using proposed model shows remarkable results as shown in Fig. 8. Models can maintain counting consistency on different fry sizes and density, with less than 2% counting difference between ground truth and estimated count. Predicted density maps are evaluated using SSIM with reference to the ground truth density maps. The model predicted density maps similarity remains higher than 98% compared to ground truth.

OSA-Net tested in PENSLV-2 Dataset in Fig. 7, which consists of 100 images of 12 days old PL with 1920×1080 pixels, captured from different basins, angles, and lightning down sampled to 540×960. The results show that OSA-Net performs remarkably well in unseen data with different PL sizes and density levels with error rate under 11%.

$$MAE = \frac{1}{M}\sum_{i=1}^{N}|YC_i - FC_i| \tag{13}$$

$$MSE = \sqrt{\frac{1}{M}\sum_{i=1}^{N}|YC_i - FC_i|} \tag{14}$$

3.3 Comparative studies

3.3.1. Fry counting networks

Fry counting using a single capture angle is prone to fry adhesion leading to poor model performance. The role of backlight in the custom basin is crucial in this case. The slightly transparent body of PL is beneficial in a collision scenario which changes color to be darker due to dimmed more backlight compared to single PL.

OSA-Net is compared with state-of-the-art fry counting models: SLCOBNet [15], Shrimpseed_Net [16], MSENet [8], and ShrimpCountNet [17]. All models are trained with the same hyperparameter and hardware setup as OSA-Net.

Table 1. Hyperparameters value for OSA-Net training

Initialization	Random
Epoch	300
Batch Size	4
Learning rate	0.0001
Lambda	0.0005
Optimizer	Adam

Table 2. Comparison with crowd counting models in PENSLV-1 dataset.

Model	**Params↓**	**FLOPs↓**	**MAE**	**MSE**
CMTL	2.46 MB	6.02 G	4.30	5.61
SFCN	38.6 MB	31.62 G	3.70	4.41
CSRNet	16.2 MB	20.41 G	3.65	4.19
MCNN	**0.13 MB**	**1.36 G**	7.32	9.44
OSA-Net	0.66 MB	1.99 G	**1.99**	**2.69**

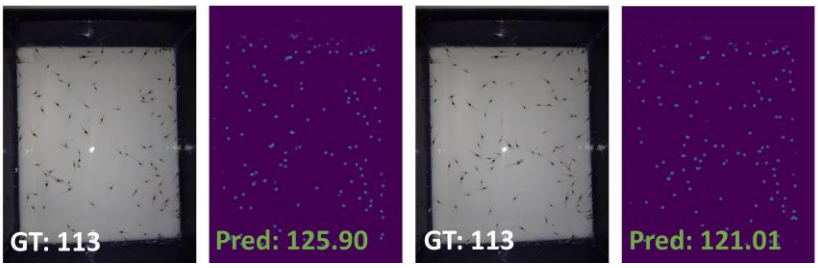

Figure. 7 OSA-Net counting prediction in unseen dataset (PENSLV-2).

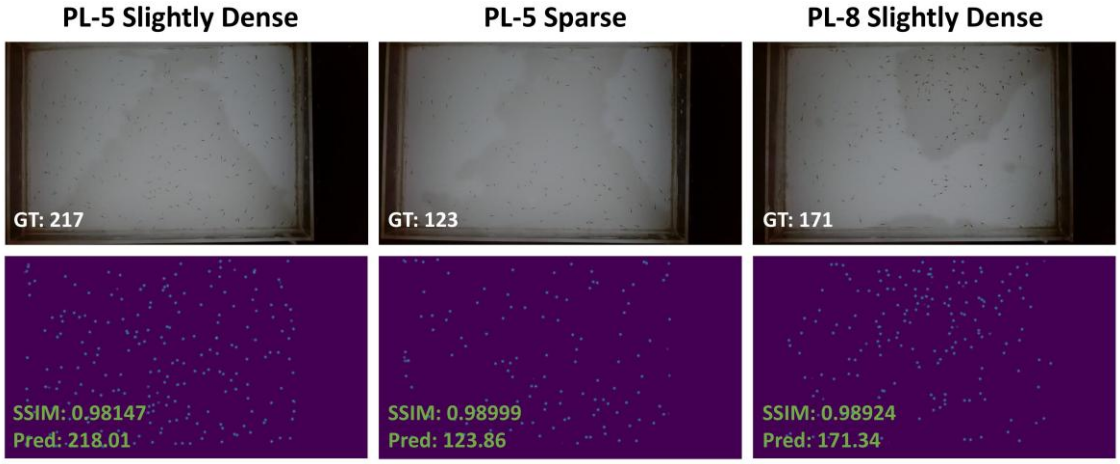

Figure. 8 OSA-Net fry counting results in PENSLV-1 dataset case, PL age dictate as "PL-days old".

$$\sigma FY(x) = \sum_{p \in \boldsymbol{P}} W(p) \cdot [F(x+p) \cdot Y(x+p)] - [\mu_F(x) \cdot \mu_Y(x)] \tag{10}$$

As shown in Table 3 and Fig. 9, SLCOBNet has 108.06 MB of parameters and makes it categorized as heavy models due to its complex architecture which require splitting images resulting in 4.09 Tera FLOPs of model complexity. Shrimpseed_Net with modified CSRNet architecture achieved descent accuracy with MAE (3.64), yet the model complexity is the second largest. ShrimpCountNet with SANet and TedNet combination architecture resulting in MAE (2.24) but has 0.92 MB of parameters which is slightly higher than original SANet. MSENet is the most compact model, yet OSA-Net counting performance remains wins with 33.7% MAE difference. OSA-Net captures complex patterns more effectively due to its aggregation architecture and SENet embedding. However, OSA-Net has slightly higher FLOPs (1.99 G vs 1.38 G) and longer training and validation times compared to MSENet. These computational costs are offset by its improved performance. In real-world applications of PL counting where counting accuracy is paramount, validation time of less than one second difference than MSENet is acceptable. These metrics highlight OSA-Net superior accuracy, particularly in handling challenging fry occlusions in a short time.

Table 3. Comparison with state-of-the-art fry counting models in PENSLV-1 dataset scenario.

Model	**Params**	**FLOPs**	**Train Time**	**Val Time**	**MAE↓**	**MSE↓**
SLCOBNet	108.06 MB	4.26 T	78.9 s	10.2 s	11.45	15.8
Shrimpseed_Net	2.84 MB	10.21 G	23.97 s	1.93 s	3.64	4.38
MSENet	**0.13 MB**	**1.38 G**	**10.73 s**	**1.04 s**	3.00	3.70
ShrimpCountNet	0.92	4.92 G	69.44	2.7 s	2.24	2.81
OSA-Net	0.66 MB	1.99 G	23.52 s	1.94 s	**1.99**	**2.69**

Table 4. OSA-Net performance comparison with SANet and OSA-Net without SE-Net in PENSLV-1 dataset.

Model	**Params↓**	**FLOPs↓**	**Train Time↓**	**Val Time↓**	**MAE↓**	**MSE↓**
SANet	0.91 MB	4.09 G	34.32 s	2.68 s	2.80	3.60
OSA-Net-W	**0.61 MB**	**1.76 G**	**17.91 s**	**1.38 s**	3.21	4.25
OSA-Net	0.66 MB	1.99 G	23.52 s	1.94 s	**1.99**	**2.69**

Table 5. Backbone models comparison trained in PENSLV-1 dataset.

Model	Params↓	FLOPs↓	Train Time↓	Val Time↓	MAE↓	MSE↓
ResNet101	27.67 MB	22.82 G	74.39 s	4.28 s	3.24	4.40
ResNet50	8.7 MB	7.64 G	26.50 s	2.00 s	3.01	3.72
AlexNet	61.1 MB	**0.75 G**	**10.70 s**	**1.01 s**	3.43	4.20
VGG	7.7 MB	13.90 G	36.65 s	2.34 s	3.65	4.68
VGG Decoder	8.4 MB	14.60 G	39.09 s	2.50 s	6.57	8.39
SANet	**0.91 MB**	4.09 G	34.32 s	2.68 s	**2.80**	**3.60**

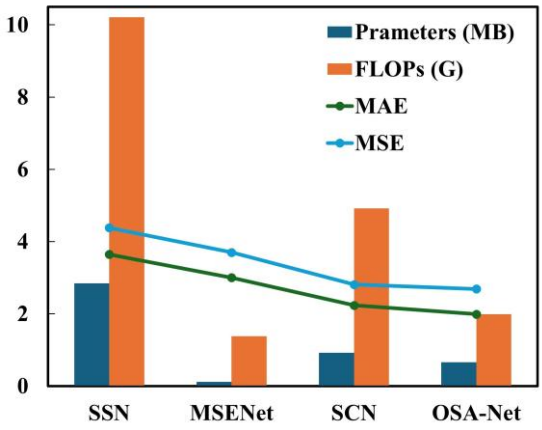

Figure. 9 Stats comparison with other fry counting models in PENSLV1 dataset exclude SLCOBNet, network names SSN refer to Shrimpseed_Net, SCN refer to ShrimpCountNet.

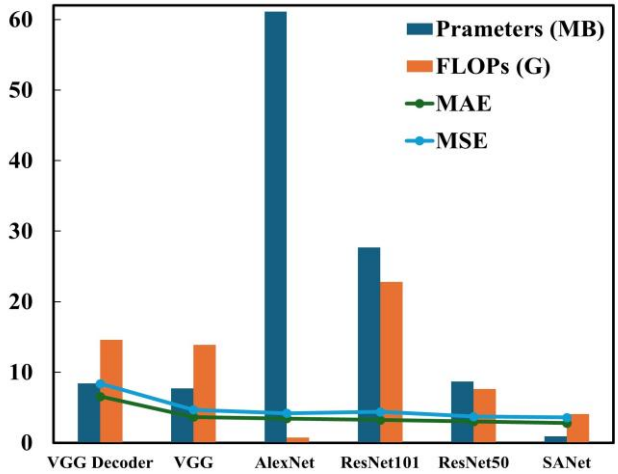

Figure. 11 Stats comparison of backbone models in PENSLV-1 dataset.

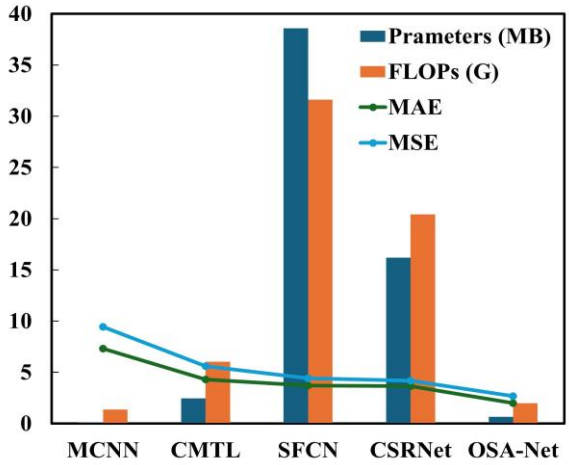

Figure. 10 Stats comparison with other crowd counting models in PENSLV1 dataset.

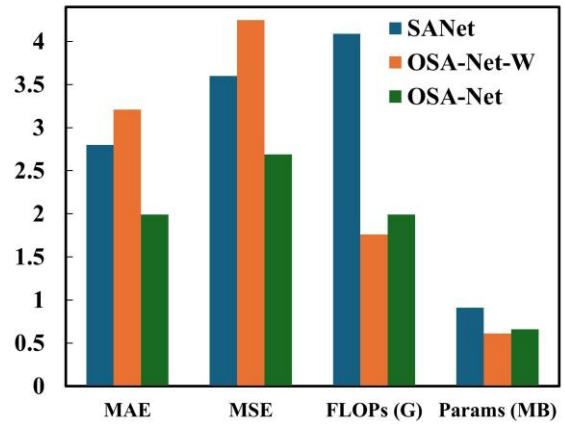

Figure. 12 Stats comparison of OSA-Net with SANet and OSA-Net-W in PENSLV-1 dataset.

3.3.2. Crowd counting networks

OSA-Net evaluated and compared with publicly available density map crowd counting models: CMTL [11], CSRNet [12], MCNN [14] and SFCN [13]. All models trained on the PENSLV-1 dataset.

The comparison aims to validate OSA-Net's effectiveness in addressing the complex challenges of fry counting, including varying densities, occlusions, and visual noise which are similar with crowd counting networks addressed. By training all models under the same hyperparameter settings and hardware setup as OSA-Net, the evaluation ensures a fair assessment of its performance relative to established crowd counting networks.

As shown in Table 2 and Fig. 10, MCNN is the lightest model, but its counting performance is suboptimal. CMTL also provides a compact architecture, but the performance is still behind OSA-Net. CSRNet shows a great result but comes with high model complexity (20.41 G). Another great performance achieved with SFCN but the model complexity is also high (31.62 G). OSA-Net remains wins in the counting performance with MAE (1.99) -

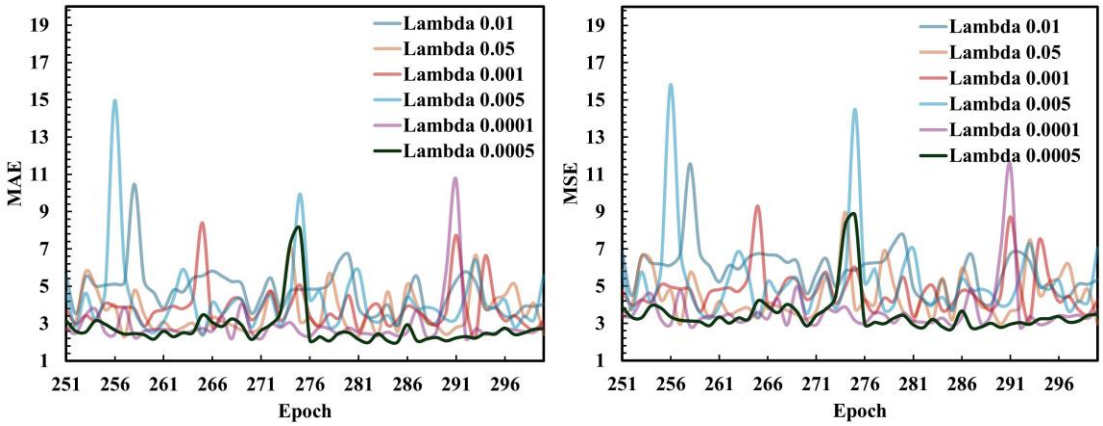

Figure. 13 OSA-Net training MAE and MSE curves corresponding to different lambda values.

Table 6. Best MAE and MSE correspond to different lambda values.

Lambda	**0.01**	**0.05**	**0.001**	**0.005**	**0.0001**	**0.0005**
MAE↓	2.30	2.01	2.17	2.35	2.04	**1.99**
MSE↓	2.92	2.65	2.83	2.99	2.60	**2.69**

-and MSE (2.69) among other density map regression models. This underscores the model's superior counting under fry adhesion and varied density levels while maintaining compacity of the model with (0.66MB) model parameter and (1.99 G) FLOPs model complexity which strongly indicates OSA-Net compatibility deployment to edge devices. (0.63 G) FLOPs difference with MCNN is acceptable since the model accuracy difference is more than 300% or 3 times better.

3.4 Ablation studies

3.4.1. Backbone

Model ability to run on edge devices is dependent on the network complexity. Varied backbones ResNet [22], AlexNet [20], and VGG [21] are compared with the OSA-Net backbone which is SANet [18] to validate model simplicity. All models trained on the PENSLV-1 dataset.

The results shown in Table 5 and Fig. 11, ResNet101 and AlexNet have over (25 MB) parameters followed by ResNet50, VGG, and VGG decoder that have over (7.5 MB) parameters which indicate high model complexities. However, The FLOPs of AlexNet are the lowest compared to ResNet101, ResNet50, VGG, and VGG decoder with (0.75 G). SANet demonstrates the most acceptable backbone to be implemented in edge devices with only (0.91 MB) parameters. In terms of counting performance, SANet performs remarkably well with MAE (2.80) and MSE (3.60) compared to AlexNet (the lowest FLOPs). This concludes SANet validation as a backbone model, which demonstrates both high accuracy counting and compactness.

3.4.2. Attention mechanism

The role of SE-Net is investigated closely. The backbone model of SANet and OSA-Net without SE-Net (OSA-Net-W) are compared with OSA-Net. All models are trained using PENSLV-1 dataset to show SE-Net addition effectiveness to compensate for channels pruned in counting PL under different PL sizes and density levels. The results shown in Table 4 and Fig. 12, SANet exhibit high model complexity with a parameter size of (0.91 MB) and (4.09 G) FLOPs, demonstrating remarkable counting performance compared to OSA-Net-W. This suggests that performance is influenced by the number of channels. However, the integration of SE-Net into OSA-Net-W results in significant improvements, achieving a 38% reduction in MAE (1.99) and a 36.71% reduction in MSE (2.69), outperforming the original SANet while utilizing fewer parameters and FLOPs. This underscores the critical role of SE-Net in compressing less informative channels while amplifying those with greater feature significance. SE-Net compensates for pruned channels in OSA-Net-W, enabling the creation of a compact model that maintains high performance while preserving efficiency called OSA-Net.

3.5 Optimal lambda search

Since OSA-Net model training is affected by the value of L_C that tune by lambda (λ_C), searching for the best value of lambda is employed. This search is beneficial for model versatility under similar or different scenarios such as different types of fish.

Initial parameters except lambda are the same as Table 1. The testing involves different lambda values: 0.01, 0.05, 0.001, 0.005, 0.0001, 0.0005. As the results shown in Fig. 13 and Table 6, Lambdas 0.0005, 0.05, and 0.0001 show the best result with 0.0005 as the best model performance. Lambdas 0.005 and 0.01 have the poorest model performance and many oscillations in MAE and MSE plot meaning the model performs less in some data samples and fails to converge. From the statistics it can be concluded that keeping lambda values around 0.05, 0.0001 or 0.0005 will help the OSA-Net model to converge.

4. Conclusion

Counting *Litopenaeus vannamei* post-larvae fry manually can take hours and has poor accuracy. Computer-vision counting methods strategically solved the problem by its non-invasiveness, high accuracy, and speed. In this work, we introduced OSA-Net, a compact *Litopenaeus vannamei* fry counting network based on density map regression techniques. By utilizing scale aggregation networks that have been channels pruned, the model managed to be compact or lightweight with parameters of 660 KB. OSA-Module with squeeze-and-excitation network addition is built to achieve high accuracy counting performance while still maintaining model compactness. PENSLV-1 dataset is built with several density levels and fry sizes to train and validate the model. Optimal lambda search is employed for model versatility information under implementation of similar or different scenarios. To validate performance, several comparative studies with publicly available models along with ablation studies to verify the role of each component in the network are employed. The results show the model can perform remarkably under different fry density levels and sizes with 1.99 and 2.69 of MAE and MSE respectively. Generalization test with PENSLV-2 datasets shows the model robustness under unseen data with under 11% of error counting rate. The minimum parameter and complexity of OSA-Net is beneficial for implementation in edge devices for portable counting such as jetson-nano or raspberry-pi. Despite the success, challenges remain, particularly with complex fry adhesion, where fries stick together in highly dense counting. Future work could address this by incorporating additional viewpoints to capture more spatial details of the fry in different angles.

Notations:

Variable	Notation
U	Input feature map with dimensions C×H×W in SE-Net.
C, H, W	Number of channels, height, and width of features respectively.
u_c	input feature map in c-th channel U$[u_1, u_2, \dots, u_c]$.
z, z_c	channel-wise descriptor obtained by global average pooling.
s, s_c	Final channel-wise attention weight.
$\boldsymbol{W_1}, \boldsymbol{W_2}$	Learnable weight matrix, used for dimensionality reduction and expansion.
σ	Sigmoid activation function.
γ	ReLU activation function.
r	Reduction dimension ratio, equal 16.
$\tilde{X}_C$	The recalibrated output feature map.
Y	Ground truth density maps.
F	Predicted density maps.
x	Spatial coordinate of fry dots in density maps.
$G_{\beta i}(x)$	Gaussian kernel centered at x_i with adaptive bandwidth β_i.
β_i	Adaptive bandwidth for the kernel at x_i.
α	Scaling factor for dots bandwidth in Gaussian filter kernels, equal to 0.1.
$\bar{d}$	Average distance to the nearest fries relative to x_i.
K, N, M	Number of nearest fries considered, number of pixels and, number of test samples respectively.
L, L_E, L_C	Total loss, Euclidean loss, and Local Pattern Consistency Loss respectively.
λ_C	Lambda for L_C tuning, equal to 0.0005.
μ_F, μ_Y	Local means of predicted and ground truth density maps respectively.
σ_F^2, σ_Y^2	Local variance of predicted and ground truth density maps respectively.
σFY	Covariance of predicted and ground truth density maps.
$\boldsymbol{P}$	All positions of the kernels, $\boldsymbol{P} = \{(-1,-1),..,(1,1)\}$.
p	Offset from the center of kernels, $p \in \boldsymbol{P}$.
$L(p)$	Weight in p locations, (weight is not updated in back propagation).
$SSIM(x)$	Structural Similarity Index Metrics of x.
YC, FC	Ground truth and predicted number of fries respectively.
C_1, C_2	Constant to avoid division by 0, the values are following [11].
MAE	Mean Absolute Error
MSE	Mean Squared Error

Conflicts of Interest

The authors declare no conflict of interest. PENSLV-1 dataset will be made available only upon request.

Author Contributions

Conceptualization, 1st and 2nd Authors; methodology, 1st and 3rd Authors; hardware and software, 1st Author; data collection, 1st Author; writing, 1st Author; editing and review, 2nd and 3rd Authors; computation resource, 2nd and 3rd Authors; *Litopenaeus vannamei* fries, 1st Author.

Acknowledgements

Authors would like to express highest gratitude to the aquaculture technology laboratory at Politeknik Elektronika Negeri Surabaya for their invaluable support to this research.

References

[1] D. Li *et al.*, "Automatic counting methods in aquaculture: A review", *Journal of the World Aquaculture Society*, Vol. 52, No. 2, pp. 269–283, 2020.

[2] F. P. Nurmaida *et al.*, "Comparison of CNN-Based Design for Shrimp Seed Counting Machine", In: *Proc. of 2023 International Electronics Symposium (IES)*, Denpasar, Indonesia, pp. 493-498, 2023.

[3] Z. Zhao et al., "A study on the effect of temperature training on compensatory growth and pathogen resistance of post-larval Litopenaeus vannamei", *Aquaculture International*, Vol. 32, No. 6, pp. 7387–7411, 2024.

[4] S. Armalivia, Z. Zainuddin, A. Achmad and M. A. Wicaksono, "Automatic Counting Shrimp Larvae Based You Only Look Once (YOLO)", In: *Proc. of 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, Bandung, Indonesia, pp. 1-4, 2021.

[5] S. V Siar, W. L. Johnston, and S. Y. Sim, *Study on economics and socio-economics of small-scale marine fish hatcheries and nurseries, with special reference to grouper systems in Bali, Indonesia*, Vol. 2, Network of Aquaculture Centres in Asia-Pacific, Bangkok, p. 36, 2004.

[6] M. T. Budiarto, R. Artiono, and R. Setianingsih, "Ethnomathematics: Formal Mathematics Milestones for Primary Education", *Journal of Physics: Conference Series*, Vol. 1387, No. 1, p. 012139, 2019.

[7] L. Rossi, C. Bibbiani, B. Fronte, E. Damiano, and A. Di Lieto, "Validation campaign of a smart dynamic scale for measuring live-fish biomass in aquaculture", In: *Proc. of 2022 IEEE Workshop on Metrology for Agriculture and Forestry*, Perugia, Italy, pp. 111–115, 2022.

[8] W. Li, Q. Zhu, H. Zhang, Z. Xu, and Z. Li, "A lightweight network for portable fry counting devices", *Applied Soft Computing*, Vol. 136, p. 110140, 2023.

[9] J. Li, J. Sun, X. Cui, B. Jiang, S. Li, and J. Liu, "Automatic Counting Method of Fry Based on Computer Vision", *IEEJ Transactions on Electrical and Electronic Engineering*, Vol. 18, No. 7, pp. 1151–1159, 2023.

[10] M. Wang, H. Cai, Y. Dai, and M. Gong, "Dynamic Mixture of Counter Network for Location-Agnostic Crowd Counting", In: *Proc. of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, pp. 167-177, 2023.

[11] V. A. Sindagi and V. M. Patel, "CNN-Based cascaded multi-task learning of high-level prior and density estimation for crowd counting", In: *Proc. of 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Lecce, Italy, pp. 1-6, 2017.

[12] Y. Li, X. Zhang and D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes", In: *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 1091-1100, 2018.

[13] Q. Wang, J. Gao, W. Lin and Y. Yuan, "Learning From Synthetic Data for Crowd Counting in the Wild", In: *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 8190-8199, 2019.

[14] Y. Zhang, D. Zhou, S. Chen, S. Gao and Y. Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network", In: *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 589-597, 2016.

[15] Y. Qu, S. Jiang, D. Li, P. Zhong, and Z. Shen, "SLCOBNet: Shrimp larvae counting network with overlapping splitting and Bayesian-DM-count loss", *Biosystems Engineering*, Vol. 244, pp. 200–210, 2024.

[16] D. Liu *et al.*, "Shrimpseed_Net: Counting of Shrimp Seed Using Deep Learning on Smartphones for Aquaculture", *IEEE Access*, Vol. 11, pp. 85441–85450, 2023.

[17] 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.

[18] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale Aggregation Network for Accurate and Efficient

Crowd Counting", In: *Proc. of Lecture notes in computer science*, Springer, pp. 757–773, 2018.
[19] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity", In: *Proc. of IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp. 600-612, 2004.
[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", *Communications of the ACM*, Vol. 60, No. 6, pp. 84–90, 2017.
[21] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size", In: *Proc. of 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Kuala Lumpur, Malaysia, pp. 730-734, 2015.
[22] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", In: *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778, 2016.
[23] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks", In: *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7132-7141, 2018.
[24] C. Szegedy *et al*., "Going deeper with convolutions", In: *Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 1-9, 2015.
[25] V. A. Sindagi and V. M. Patel, "Generating High-Quality Crowd Density Maps Using Contextual Pyramid CNNs", In: *Proc. of 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 1879-1888, 2017.