IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 April 2024, accepted 9 May 2024, date of publication 13 May 2024, date of current version 21 May 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3400310*

METHODS

# Crowd Counting and Individual Localization Using Pseudo Square Label

**JIHYE RYU AND KWANGHO SONG**
INFINIQ Corporation, Gangnam-gu, Seoul 06232, Republic of Korea
Corresponding author: Kwangho Song (khsong@infiniq.co.kr)

This work was supported by the Institute of Civil Military Technology Cooperation funded by the Defense Acquisition Program Administration and Ministry of Trade, Industry and Energy of Korean Government under Grant 23-CM-Al-15.

**ABSTRACT** Recent work in crowd counting focuses on counting over detected individuals rather than estimating the number of people in the image. However, existing crowd localization methods directly detect the head point or region of individuals, which may entail non-responsibility of the outputs that fall outside the grid. Our proposed Pseudo Square Label Network (PSL-Net) presents a novel method for crowd counting and localization, which takes advantage of the anchor-free detection in which PSL-Net predicts the probability of the center point that fall into the responsible grid, while indirectly detecting an individual outside of the responsible grid through box regression and centerness estimation. This study proposes to supervise with pseudo square label(PSL), which is generated around point annotation with fixed size. Furthermore, we design a partial many-to-one matching algorithm to assign precise labels by only matching within PSL during the training phase, and associate the predicted points with their responsible grids through centerness during the inference phase. As a result, not only PSL-Net achieves state-of-the-art on ShanghaiTech Part A and B, which are the most popular datasets in crowd counting, but also achieves state-of-the-art among the point detection-based methods in crowd localization.

**INDEX TERMS** Crowd counting, crowd localization, anchor-free object detection, point estimation, video surveillance.

## I. INTRODUCTION

After the recent crowd incident [1], the necessity of crowd counting, which estimates the number of people in crowd images, has been raised in the field of crowd analysis. In general, conventional crowd counting estimates the number of individuals in the crowd using a neural network that utilizes image patches or density maps. However, users of surveillance systems in the real world expect to monitor crowds and capture abnormal situations, which cannot be achieved through conventional crowd counting only. Therefore, the current crowd counting research has been focused on crowd localization, predicting the exact location of each individual, rather than solely predicting the number or density map as in a general approach.

The crowd localization is distinguished into two representatives as shown in Fig. 1. The first is conventional anchor-based detection with bounding box(bbox) [2], [3], [4], [5], [6], displayed at the top, and anchor-based detection with points [10], [11], [12], [13], displayed at the middle. Additionally, there are segmentation-based approaches [7], [8], [9]. The bbox detection method involves directly predicting the left-top position of the head bbox (x, y), length of the bbox side (w, h) and probability within a responsible grid (i.e., a grid cell). Since there is one anchor point per grid cell detecting bbox, there may be inevitable false negatives when dense crowds occupy the same grid cell, as shown at the top of Fig. 1. The point detection method involves directly predicting the center coordinate (x,y) of the human head and probability. However, there is no constraint on the reachable distance from each anchor point. Consequently, several anchor points exceed the responsible grid to match with the ground truth(GT) points located outside of the grid, leading to an increase in non-responsible predictions.

In order to compensate for the shortcomings of previous methods, we propose a novel method, Pseudo-Square-Label

The associate editor coordinating the review of this manuscript and approving it for publication was Alessia Saggese.

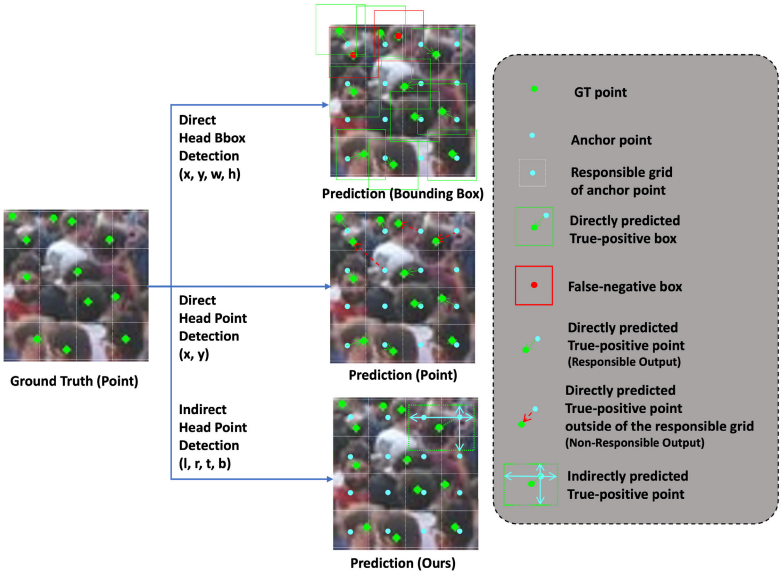



**FIGURE 1.** Comparative illustration of head detection on crowd by ours between the other method.

Network (PSL-Net), to indirectly estimate the center of the human head region as the (x,y) coordinate form by predicted Bbox as shown in Fig. 1. Specifically, for the responsible prediction of the center point on the human head region, PSL-Net uses the way of anchor-free object detection that separates the anchor point on output and the center point of the object. Therefore, unlike the previous anchor-based works, which directly estimates the surrounding Bbox area(x, y, w, h) [2], [3], [4], [5], [6] or the center coordinates (x, y) [10], [11], [12], [13] of the human head object for each anchor point, PSL-Net directly estimates the distances (l, r, t, b) from the anchor point to the four sides of bbox containing the anchor point and indirectly estimates the coordinate of the human head center point. Thus, the responsibility of prediction on PSL-Net is always preserved, even if the center point of the human head falls the outside of the responsible grid, because every output is generated from an anchor point fixed at the center of the grid cell. PSL-Net also employs specific strategies to ensure accurate predictions. During training, it utilizes a partial many-to-one label assignment to create positive pairs based on local adjacency. Meanwhile, PSL-Net estimates the centerness of predictions in inference phase, which represents the likelihood of a human head being present at each anchor point. This approach minimizes the inclusion of irrelevant positive samples during training and selects reliable outputs with high probability and centerness in the inference phase.

Additionally, because the ultimate objective of PSL-Net is to estimate head center coordinate precisely, the PSL-Net does not need the costly man-made Bbox label which has the same size with each head on training phase. Instead, the PSL-Net make and use predetermined pseudo square labels(PSL), which do not require additional optimization process fitting bbox to the real head size, to avoid expensive annotation costs. The size of the predetermined PSL is designated through grid search about the head object sizes, which varies with distance, angle and resolution of the camera, in crowd images. And as a result, PSL-Net utilizes 3 types of pseudo-square labels that are large enough to recognize the existence of the human head with the naked eye.

In summary, The main contributions of this work are:

1) We propose the PSL-Net using the way of anchor-free object detection to make responsible estimation of head center points even if it is placed outside of the responsible grid.
2) We propose a partial many-to-one matching algorithm used in the training phase and centerness estimation used in inference phase to accurately assign positive labels to proximate ground truth-anchor pairs.
3) We propose a label assignment step using three types of predetermined pseudo square labels without any additional annotation costs. Additionally, PSL-Net can be applied to crowd analysis tasks such as individual recognition, tracking, or flow analytics at the crowd level.
4) Our proposed PSL-Net achieves state-of-the-art performance in crowd counting across multiple benchmark datasets, surpassing existing point/box detection frameworks in crowd localization benchmarks.

Following Sec.II briefly reviews existing studies in the field of crowd counting and elaborates the head point detection methods. Sec.III explains the details of the PSL-Net such as responsible prediction on the inference phase, label assignment algorithm on the training phase, and the network design. Sec.VI presents experimental results of PSL-Net and the performance comparison with other works.

Ablation studies and the consideration also explained in this section. Finally, Sec.V provides a summary of this study.

## II. RELATED WORKS

This section reviews the related works for crowd counting and crowd localization, a branch of crowd counting. The works are based on bbox detection, segmentation, and point detection, which are similar to the proposed methods.

Most of the recent works for crowd counting make the density map on pixel level [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24] or image patch level [25], [26], [27] to precisely estimate the number of people in a crowd. They use effective spatial information extracted by various local features and its fusion, context information extracted through attention mechanism, and perspective information. Those works have been nominated as a good candidate for crowd surveillance systems. However, individual locations on a crowd, which is required from recent crowd analysis, cannot be identified by density map-based methods without an additional post-processing [28] or an independent localization module [29] to find the peak point on the density map. Also, even if additional processes to get individual locations are used, it is still difficult to distinguish individuals in congested crowds [15].

Crowd localization based on bbox detection has evolved from redirecting general purpose object detection models [2], [5] to developing the optimized novel architecture [30]. Furthermore, recent studies [3], [4], [7] devise pseudo bbox label optimization that adjusts the size of each person's head, which uses additional modules or optimization processes, to avoid the problems related with annotation cost on crowd images. Most of these utilize the way of anchor-based bbox detection that predicts the bbox coordinates and the probability of the object at each anchor point. In other words, visually identifiable objects can be predicted on each anchor point. As shown in Fig. 1, if crowd density increases, the number of people in each grid cell also increases. Then it makes it challenging to predict bounding boxes for each specific object due to object occlusion and can significantly reduce performance by increasing the number of false negatives.

Recently, the crowd localization based on the point detection has been proposed to supplement the shortcomings of bbox detection [10], [11], [12], [13]. These methods also detect heads like the conventional anchor-based methods, but the prediction objective is the center point instead of the object region of the human head. Therefore, the label takes the form of a coordinate (x,y) which helps to minimize the limitations related to occlusion between heads and avoid the shortcomings about costs of bbox format annotation. P2P-Net [10] demonstrated exceptional performance in high-density crowd images and has achieved state-of-the-art results in several benchmarks in the field of crowd counting.

To be more specific, P2P-Net [10] uses a convolutional neural network (CNN) with an encoder-decoder structure to form an output that is reduced to $1/s$ compared to the input, as shown in Fig. 2. The output includes: ① $(\Delta x, \Delta y)$, which is the mahalanobis distance between the anchor point $(A_x, A_y)$ and the predicted center point $(A_x + \Delta x, A_y + \Delta y)$ on the target image, and ② probability, which indicates whether a human exists in the predicted location ①. As long as P2P-Net [10] uses an anchor-based detection that predicts the coordinates and probability of the object at each anchor point, the predicted probability ② could be responsible only when the predicted point $(A_x + \Delta x, A_y + \Delta y)$ exist in responsible grid cell region of its anchor point $s \times s$ in target image.

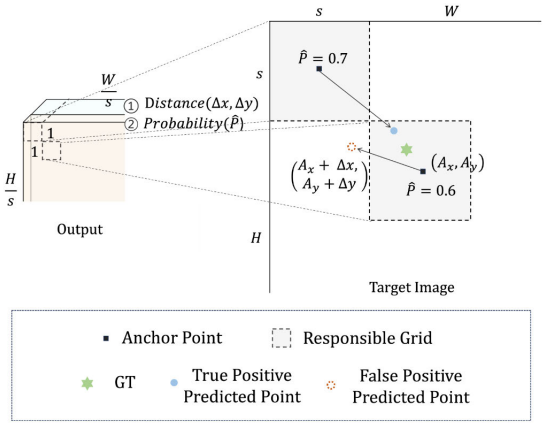



**FIGURE 2.** Non-responsible output example on P2P-Net [10].

However, the anchor point on P2P-Net [10] can predict coordinates outside of its responsible grid, as shown in Fig. 2 and it cannot be responsible for the predicted probability. In high-density crowd images, the number of people in a responsible grid are increased and the number of anchor points around the GT points are decreased. Therefore non-responsible output is increased naturally and the reliability of output would be decreased. Furthermore, it is hard to determine the most confident prediction by only probability among the candidate anchor points, and it can cause multiple false positives.

## III. METHODOLOGY

This section describes the details on the proposed method, PSL-Net. Firstly, we explain the inference process of PSL-Net for responsible output, which is emphasized as the main contribution. Next, the label assignment algorithm using PSL is proposed. Finally, details about the network design and loss functions are provided.

### A. OUTPUT CONFIGURATION

As previously stated, PSL-Net indirectly detects the head point based on the anchor-free detection inspired by FCOS [41] to ensure responsibility of prediction. Therefore, PSL-Net outputs the three predictions for each anchor point of the output, as illustrated in Fig. 3.

① Bbox distance $\hat{B}$ : The bbox regression denoted as $(l, r, t, b)$ is prediction of the distance from the anchor

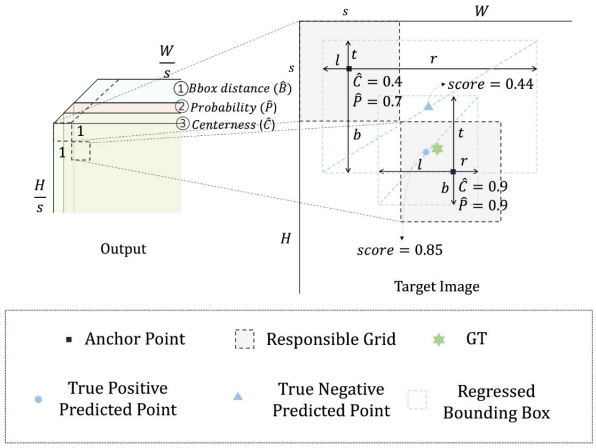



**FIGURE 3.** Responsible output example on our PSL-Net.

point to the outermost edge of the target object box, as in anchor-free object detection. It is not necessary for the center point of the predicted box to be on the responsible grid. It is important that the anchor point must exist inside the predicted box, and the center point of the predicted box may exist outside the responsible grid of the anchor point. The proposed method can preserve the responsibility for the predicted point (i.e., the person's head), regardless of whether the head is within or outside of the responsible grid.

② Probability $\hat{P}$ : The probability in object detection refers to the confidence that the predicted box area belongs to a specific class, regardless of whether an anchor-based or anchor-free method is used. However, the proposed method only requires the point of the head existing at the center of the detected bbox rather than the entire region of the head area. The proposed method predicts the probability ② of the head existing at the center of the detected box area in ①. Since each anchor point predicts based on its responsible grid, the probability may have different values when multiple anchor points simultaneously predict the same head point.

③ Centerness $\hat{C}$ : The estimation of centerness associates the responsible grid with the predicted point in the anchor-free method. It also indicates the normalized distance between the center point of the head and the anchor point, in the proposed method. The anchor-free method does not consider the association between the responsible grid and the predicted point. However, it is natural that the closer they are to each other, the more reliable the prediction is. Therefore, the proposed centerness is used to estimate the reliability of the predicted point in addition to ②.

Therefore, the proposed method evaluates the reliability of the prediction using *score* in (1), which is calculated by multiplying the probability by the weight obtained by the square root of the centerness, unlike the existing methods that use only the probability to distinguish positive and

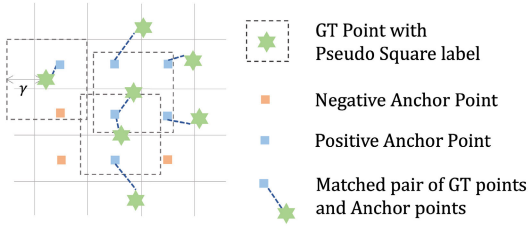



**FIGURE 4.** Illustration of the partial many-to-one matching process.

negative results. In particular, $\sqrt{\hat{C}}$ helps to further refine the localization of the PSL-Net by assigning a higher weight to the anchor point near the predicted point. Accordingly, experiments related to $\hat{C}$ have been conducted (see Table 4).

$$score = \hat{P} \times \sqrt{\hat{C}} \tag{1}$$

In summary, the proposed method preserves responsibility for predicted points obtained through bbox regression, whether they are located inside or outside the responsible grid. Fig. 3 shows that the most reliable prediction can be selected for accurate localization based on centerness, when multiple anchor points predict the same GT point with high probability.

### B. LABEL ASSIGNMENT

The PSL-based label assignment algorithm is proposed to distinguish the positive or negative predicted points for each batch in the training process, which is the method of matching the anchor point with GT based on their adjacency.

Our method is partial many-to-one matching, as shown in Fig. 4. For each GT, The pair of the anchor point and GT is matched based on the hungarian algorithm [31], which minimizes the cost matrix of the adjacency and probability. Based on the assumption that reachable distance from anchor point to GT is limited to the radius $\gamma$ of inscribed circle of the PSL. Since the hyperparameter $\gamma$ defines the maximum margin between the anchor points and GT points, the optimal value varies depending on the distribution of the distances between anchor points and GT points in the input image. The cost matrix $M$ with respect to the adjacency $\mathcal{L}_{DIoU}$ and probability $\hat{P}$ is defined as in (2):

$$\begin{aligned} M(\mathbb{A}, \mathbb{G}) &= \mathcal{L}_{DIoU}(\mathbb{A}, \mathbb{G}) - \hat{P}_j \\ &= 1 - IoU(B_j^{\mathbb{A}}, B_i^{\mathbb{G}}) + \frac{||A_j - G_i||_2}{d^2} - \hat{P}_j \end{aligned} \tag{2}$$

where $\mathbb{G}$ denotes to the set of GT points $G_i$ ($i \in [0, I)$), and the $\mathbb{A}$ denotes to the set of the anchor points $A_j$ ($j \in [0, J)$) of the input image. As PSL for each $G_i$ is assigned with a length of $2\gamma$, the each of elements in cartesian product of $\mathbb{A}$ and $\mathbb{G}$ can be used to calculating the DIoU loss [32] ($\mathcal{L}_{DIoU}(\mathbb{G}, \mathbb{A})$, $\mathcal{L} \in [0, 2]$) of bboxes $B_i^{\mathbb{G}}$, $B_j^{\mathbb{A}}$, and the probability matrix also can be comprised from the each pair. Therefore, the shape of the cost matrix $M$ is $(I, J)$.

By using the DIoU loss which is defined with the ratio of the diagonal distance ($d$) to the L2-distance between

two boxes in the cost matrix, GT could be matched to the nearest anchor points among the adjacent anchor points. Additionally, the DIoU loss encourages the predicted points to be close to the center of the head(i.e., GT), by suppressing the large regressed bbox, so that the predicted point could be near the responsible grid. During the one-to-one matching, if the unmatched anchor point in PSL region is not remained, the matched positive point with minimum DIoU could be matched repeatedly with GT. It is also intended to be suppressing the distant pairs as mentioned. As a result, The proposed label assignment algorithm prepares the asymmetrical pairs for training PSL-Net. In other words, GT can be matched only with the single predicted point, while the predicted point can be matched with multiple GT. The aforementioned process is described in Algorithm 1.

**Algorithm 1** Algorithm for Partial Many-to-One Matching

**Require:** $N$ is the number of samples in an batch; $I$ is the number of the GT points in an image; $J$ is the number of the predictions in an image; $\mathbb{G}_i$ is the set of GT, $\mathbb{G}_i \in \mathbb{R}^{Ix2}$; $B_i^{\mathbb{G}}$ is the set of PSL of the $\mathbb{G}_i$; $\mathbb{A}_j$ is Anchor points of Responsible Grid, $\mathbb{A}_j \in \mathbb{R}^{Jx2}$; $B_j^{\mathbb{A}}$ is the set of PSL of the $\mathbb{A}_j$; $\hat{P}_j$ is probability map of an image, $\hat{P}_j \in \mathbb{R}^{Jx1}$; $D$ is a function that calculates the DIoU loss between two input boxes; $H$ is a function that associates the two input matrices

**Ensure:** *$X$ is a set of matched index of predictions; $Y$ is a set of matched index of GT*

$X \leftarrow \emptyset$
$Y \leftarrow \emptyset$
**for** $0 \leq n \leq$ N **do**
  let $m_d$ be the pair-wise $D$ of $B_j^{\mathbb{A}}$ and $B_i^{\mathbb{G}}$, $m_d \in \mathbb{R}^{JxI}$
  let $m_p$ be the pair-wise matrix of $\hat{P}_j$ by $\mathbb{G}_i$, $m_p \in \mathbb{R}^{JxI}$
  $M \leftarrow m_d - m_p$
  $x, y \leftarrow H(M)$
  $ind\_x, ind\_y \leftarrow where(D(\mathbb{A}_x, \mathbb{G}_y) \geq 2)$ ▷ The maximum of DIoU is 2
  **for** $0 \leq ix, iy \leq ind\_x, ind\_y$ **do**
    $y_{iy} \leftarrow argmin(M_{ix})$
  **end for**
  $X = X \cup x$
  $Y = Y \cup y$
**end for**
**return** $X, Y$

In order to supervise the proposed method, the three types of labels for each output are assigned to the positive predicted points through the matching process: ① PSL with the length of $2\gamma$ surrounding GT, ② The one-hot encoding of probability to be a head, ③ centerness $\hat{C}$ between GT and their positive points via (3). On the other hand, the negative predicted points are only assigned the labels of ②.

$$C^* = 1 - \frac{||A_j - G_i||_2}{d^2} \tag{3}$$

To ensure identification of individuals in densely populated crowd images, we set ① PSL large enough to identify for the head rather than the fit bbox for each head. Therefore, PSL-Net forwards the features of the contextually identifiable bbox of the head with the background.

After assigning the label, the loss function is calculated by taking the weighted sum of each output loss, as shown in equation (4).

$$L = \lambda_1 L_P + \lambda_2 L_B + \lambda_3 L_C \tag{4}$$

For the positive predicted samples, $L_B$ is the loss function for the output ①, and $L_C$ is the loss function for the output ③, which is formulated as respectively (5), and (6). $L_B$ is DIoU loss of PSL $B_i^{\mathbb{G}}$ and regressed bbox $\hat{B}_j$ of its matched positive sample, and $L_C$ is the Cross Entropy loss [33] of estimated centerness $\hat{C}_j$ and its label $C_i^*$.

$$L_B = \frac{1}{I}\sum_0^I \sum_0^J \mathbb{1}_{i,j}^{obj} \mathcal{L}_{DIoU}(B_i^{\mathbb{G}}, \hat{B}_j) \tag{5}$$

$$L_C = \frac{1}{I}\sum_0^I \sum_0^J \mathbb{1}_{i,j}^{obj} L_{CE}(C_i^*, \hat{C}_j) \tag{6}$$

For training the classification, the loss function $L_P$ uses all predicted samples, including both positive and negative samples from output ②. However, the probability of the positive samples could be underestimated, as the proportion of the number of positive samples $I$ in the entire number of samples $J$ is significantly small, which result in the class imbalance. Thus, we add Cross Entropy of positive samples to the Weighted Cross Entropy [34] of the entire samples, for intensifying the positive samples in the training process, as shown in (7).

$$\begin{aligned} L_P = &-\frac{\beta}{I}\sum_0^I \sum_0^J \mathbb{1}_{ij}^{obj} L_{CE}(P_i, \hat{P}_j) \\ &- \frac{1}{J}\Big\{\alpha \sum_0^I \sum_0^J \mathbb{1}_{ij}^{obj} L_{CE}(P_i, \hat{P}_j) \\ &+ (1-\alpha)\sum_0^I \sum_0^J \mathbb{1}_{ij}^{noobj} L_{CE}(P_i, \hat{P}_j)\Big\} \end{aligned} \tag{7}$$

In addition, we use hyperparameter $\alpha$ to adjust the scale of Weighted Cross Entropy, because the negative samples are relatively large in number. Then we use the hyperparameter $\beta$ for Cross Entropy of the positive samples to avoid overestimating the positive samples.

### C. NETWORK DESIGN

The network architecture of PSL-Net consists of the three steps: encoder, decoder and detector, as shown in Fig.5. The VGG16BN [35]-based encoder extracts the low level feature

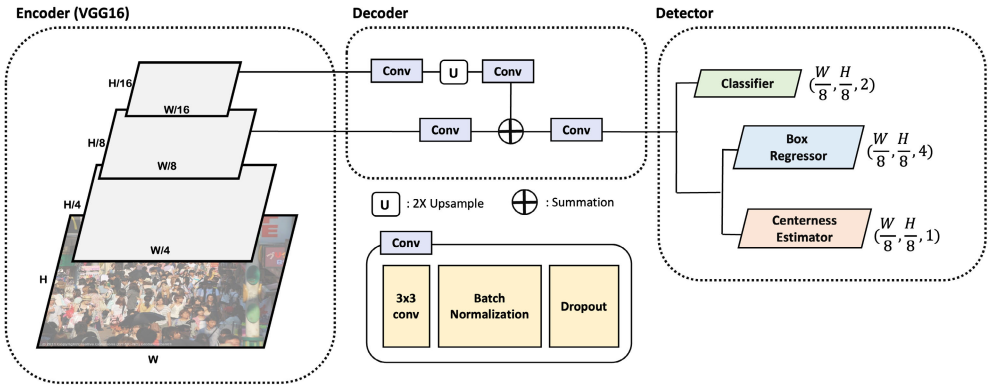



**FIGURE 5.** Network architecture of our proposed PSL-Net.

map of 1/8 size and high level feature map of 1/16 size from the input image, and then feed-forward to the decoder. This hierarchical structure using multi-level feature maps is commonly used in object detection methods [36], [37] to extract a multi-scale feature. The decoder element-wise summates the high-level feature map and the low-level feature map using a convolution module that consists of a $3 \times 3$ convolution, batch normalization and dropout. More specifically, the high-level feature map is upsampled to be the same size as the low-level feature map for element-wise summation. From this added feature map, the coarse feature map is finally extracted by the convolutional module, then forwarded to the detector. The detector is composed of the three heads as follows : the classifier for probability $\hat{P}$, the box regressor for bbox $\hat{B}$ and the centerness estimator for centerness $\hat{C}$, where the activation function for each outputs are softmax, relu, and sigmoid, respectively. In the case of a box regressor, we restrict each side of the box to $e^8$ to avoid exploding gradients. There is no constraint for the input size of the image, likewise in Fully Convolutional Network [38]. Formally, if the input size of the image is $(H, W)$, the shape of the anchor points of each output is $H/s \times W/s$. In other words, the $H/s \times W/s$ of grid cells are in the target image, where each cell has the size of $s \times s$. In terms of the computational cost, PSL-Net uses an anchor point in each grid cell, whereas [10] uses four anchor points in each grid. The number of parameters of PSL-Net is 19,203,400 while [10] has 21,579,344 parameters. Despite using only one anchor point in each grid cell, PSL-Net has shown improved performance.

## IV. EXPERIMENTS

### A. IMPLEMENTATION DETAILS

#### 1) DATASETS

For an experiment and evaluation of the proposed method, the most popular benchmark datasets such as ShanghaiTech [21] and UCF-QNRF [39] for crowd counting, and NWPU [40] for crowd localization are used.

**ShanghaiTech** [21] is a representative benchmark for crowd counting. It devided into type A(SHTA) and type B(SHTB). While SHTA mostly consists of images of extremely congested crowds, SHTB consists of images of a relatively sparse crowd. Each type has 300, 400 training data and 182, 316 test data. The average resolution of the images in SHTA is $589 \times 868$, which is smaller than other benchmarks, but, on average, 501 head annotations are spread into each image. Also, the resolution of all images in SHTB is $768 \times 1024$.

**UCF-QNRF** [39] is also a representative benchmark for crowd counting. It contains 1201 training data and 334 testing data with diverse information, such as a wide range of camera angles, light variation, and crowd density distribution, which can be used to make the crowd counting method. Also this is a huge and well-generalized benchmark, which has a diverse size of heads on multi-environment images compared to the other benchmarks. So it is used to pre-train the proposed method before its fine-tuning on evaluation phase.

**NWPU-crowd** [40] is the largest crowd localization benchmark, consisting of 5,109 images with 2,133,375 annotations. This is a well-generalized higher resolution benchmark, which has an average resolution of $2191 \times 3209$ and contains 351 negative samples. It also represents a large appearance variation of the head and supports not only point-wise annotation but also box-level annotation.

#### 2) HYPERPARAMETERS, DATA AUGMENTATIONS AND ENVIRONMENT

In the training phase, we use an Adam optimizer with a learning rate of 1e-4 and a batch size of 16, and the resolution of the input data on the batch is fixed to $128 \times 128$, which is randomly cropped from the original input image. Also, the hyperparameters for the loss function were experimentally determined to be $\alpha = 0.45$, $\beta = 0.01$, $\lambda_1 = 0.1$, $\lambda_2 = 0.01$, $\lambda_3 = 0.01$. In addition, the extremely high-resolution image samples were downsized to $1792 \times 2304$ to avoid excessive processing costs and too many negative samples from the remaining anchor points, which would degrade the overall performance.

To augment the input data, we adopted random scaling and flipping and the other details of augmentation are the same as

P2P-Net [10]. Additionally, the training and evaluation of the proposed method is conducted on the server with the NVIDIA RTX 3080Ti and Ubuntu LTS 20.04. and it is implemented by PyTorch 2.0.1 and Python 3.9.16.

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|\hat{I}_n - I_n|, \tag{8}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{I}_n - I_n)^2}$$
$$(N : Num.\ of\ test\ imgs,$$
$$I_n, \hat{I}_n : Num.\ of\ GT\ and\ positive\ output\ on\ n^{th}\ img) \tag{9}$$

To evaluate our PSL-Net on the benchmark datasets, we measure the Mean Absolute Error(MAE) as in (8) and the Root Mean Squared Error(RMSE) as in (9), which are common metrics for evaluation in crowd counting. Considering that most of the works using the formulation of RMSE presented as Mean Squared Error(MSE), we refer to RMSE as MSE when comparing with others.

In addition, we measure the precision, recall, and F1-score on NWPU, which are commonly used metrics for crowd localization evaluation.

### B. EVALUATION

We evaluate the performance of our proposed method for crowd counting based on SHTA, SHTB, and UCF-QNRF. Also, we evaluate the performance of crowd localization based on NWPU as well. Due to the extremely variable attributes on the head size in the crowd image by its resolution, we distinguish the proposed PSL-Net into 3 types based on the hyperparameter $\gamma$ that used for training as follows: PSL-Net($\gamma = 18$), PSL-Net($\gamma = 24$), PSL-Net($\gamma = 44$). Accordingly, $\gamma$ of PSL-Net($\gamma = 18$) was set to 18 considering the relatively low image resolution and mostly small head size, $\gamma$ of PSL-Net($\gamma = 44$) was set to 44 considering the high image resolution and mostly large head size, and $\gamma$ of PSL-Net($\gamma = 24$) was set to 24 in between. $\gamma$ was determined by data analysis and hyperparameter grid search of each benchmark dataset.

We compare the performance of crowd counting with crowd counting methods and crowd localization methods separately, as shown in Table 1, 2, respectively.

Compared to related works in Table 1, which only estimate the number of crowds with a density map or image patch, the three types of PSL-Net are superior to LovitCrowd [25], which is state-of-the-art on SHTA. In particular, we focus on the best-performing PSL-Net($\gamma = 18$), which achieved the MAE of 49.9 and the MSE of 77.6, reducing the MAE by 4.9 and the MSE by 3.3 compared with the existing state-of-the-art method. In addition, also on SHTB, all of the PSL-Net outperformed the existing state-of-the-art GauNet [14]. The PSL-Net($\gamma = 24$) reduces the MAE by 0.4 and MSE by 0.7. PSL-Net($\gamma = 44$) achieved the second best performance on MAE and MSE with 85.5 and 144.4, respectively on QNRF. Nevertheless, the proposed method not only reduces the MSE by 9.3 with respect to the existing best MAE [14], but also reduces the MAE by 1.5 with respect to the existing best MSE [25]. These results imply that PSL-Net is comparable to the state-of-the-arts on QNRF. Moreover, PSL-Net can localize individuals in a crowd, which other works in Table 1 cannot do.

Table 2 shows the results of the comparison with related works, which can localize individuals using bbox or point detection or segmentation. The three types of PSL-Net are superior to FGE-Net [11], which is the state-of-the-art on SHTA. The best-performing PSL-Net($\gamma = 18$), reduces the MAE by 1.7 and MSE by 7.4 compared with the existing state-of-the-art method, as on SHTB, all the PSL-Net outperform the existing state-of-the-art P2P-Net [10]. The PSL-Net($\gamma = 24$) reduces the MAE by 0.4 and the MSE by 0.7. Howerer, on QNRF, PSL-Net($\gamma = 44$) achieved the best performance with respect to the MSE, and the second best performance in terms of MAE. Since the difference with the best MAE is only 0.3, while the MSE is improved by 10.1, it could be considered comparable to the exisiting state-of-the-art FGE-Net [11]. As a result, PSL-Net fully achieved state-of-the-art based on MAE and MSE in SHTA, SHTB, and comparable performance to state-of-the-arts in crowd counting. As a result, for crowd counting, PSL-Net achieved state-of-the-art based on MAE and MSE in SHTA, SHTB, and comparable performance to the state-of-the-art on QNRF.

As mentioned above, we evaluate performance based on the NWPU test dataset. Since NWPU consists of particularly high-resolution images, the performance of our PSL-Net($\gamma = 44$), which shows outstanding performance on the high-resolution benchmark QNRF, is compared with other state-of-the-art methods based on the point detection. As shown in Table 3, the proposed method achieved the best F1-score and recall compared to the existing state-of-the-art point- based method [10], which improved by 1.5% and 4.5% respectively. Even though the precision of ours is 1% lower than that of the existing state-of-the-art P2P-Net, our PSL-Net demonstrates its superior performance in both crowd counting and localization, considering that PSL-Net outperformed P2P-Net in terms of MSE on QNRF.

### C. ABLATION STUDY

#### 1) EFFECT OF CENTERNESS AS A SCORE WEIGHT

Firstly, we examine the effect of the centerness $\hat{C}$ as the weight for the inference score by the ablation study, as shown in Table 4, where the scale of $\hat{C}$ is adjusted by the square or the square root. As mentioned before, the scale of $\hat{C}$ is amplified by its square root, while its scale is reduced by the square of $\hat{C}$, since c hat ranges from 0 to 1. In conclusion, the proposed method outperformed when increasing the scale of $\hat{C}$, which implies that the centerness of the predicted points is also important as the probability $\hat{P}$. However, the more amplified the scale of $\sqrt{\hat{C}}$ (e.g.,$\sqrt[3]{\hat{C}}$)

**TABLE 1.** Comparison of the counting performance with state-of-the-art works for crowd counting only.

| Method | Strategy | SHTA | | SHTB | | QNRF | |
|---|---|---|---|---|---|---|---|
| | | MAE | MSE | MAE | MSE | MAE | MSE |
| VGG+GPR [40] | density map | 112.4 | 176.9 | 13.1 | 19.4 | 203.5 | 343.3 |
| MCNN [21] | density map | 110.2 | 173.2 | 26.4 | 41.3 | - | - |
| DM-Count [18] | density map | 59.7 | 95.7 | 7.4 | 11.8 | 85.6 | 148.3 |
| M-SFANet+M-SegNet [15] | density map | 57.5 | 94.4 | 6.3 | 10.0 | 87.6 | 147.7 |
| GauNet [14] | density map | 54.8 | 89.1 | 6.2 | 9.9 | **81.6** | 153.7 |
| SFSL [27] | Image patch | 82.7 | 122.8 | 14.9 | 25.5 | 145.8 | 249.0 |
| TransCrowd [26] | Image patch | 66.1 | 105.1 | 9.3 | 16.1 | 97.2 | 168.5 |
| LoViTCrowd [25] | Image patch | 54.8 | 80.9 | 8.6 | 13.8 | 87.0 | **141.9** |
| PSL-Net($\gamma = 18$) | point detection | **49.9(8.8%)** | **77.6(4.0%)** | 6.0 | 9.9 | 92.9 | 156.4 |
| PSL-Net($\gamma = 24$) | point detection | 50.6 | 79.0 | **5.8(5.3%)** | **9.2(6.9%)** | 87.9 | 148.7 |
| PSL-Net($\gamma = 44$) | point detection | 50.4 | 77.9 | 6.1 | 10.0 | 85.5 | 144.4 |

**TABLE 2.** Comparison of the counting performance with state-of-the-art works on crowd counting with localization.

| Method | Strategy | SHTA | | SHTB | | QNRF | |
|---|---|---|---|---|---|---|---|
| | | MAE | MSE | MAE | MSE | MAE | MSE |
| Tiny Faces [2] | bbox detection | 237.8 | 422.8 | - | - | - | - |
| LSC-CNN [4] | bbox detection | 66.4 | 117.0 | 8.1 | 12.7 | 120.5 | 218.2 |
| PSDDN+ [3] | bbox detection | 65.9 | 112.3 | 9.1 | 14.2 | - | - |
| Topocount [8] | segmentation | 61.2 | 104.6 | 7.8 | 13.7 | 89 | 159 |
| Crowd-SDNet [12] | segmentation | 65.1 | 104.4 | 7.8 | 12.6 | - | - |
| RAZ [29] | point detection | 65.1 | 106.7 | 8.4 | 14.1 | 116 | 195 |
| P2PNet [10] | point detection | 52.7 | 85.0 | 6.2 | 9.9 | 85.3 | 154.5 |
| FGENet [11] | point detection | 51.6 | 85.0 | 6.3 | 10.5 | **85.2** | 158.7 |
| PSL-Net($\gamma = 18$) | point detection | **49.9(3.2%)** | **77.6(8.6%)** | 6.0 | 9.9 | 92.9 | 156.4 |
| PSL-Net($\gamma = 24$) | point detection | 50.6 | 79.0 | **5.8(6.1%)** | **9.2(6.9%)** | 87.9 | 148.7 |
| PSL-Net($\gamma = 44$) | point detection | 50.4 | 77.9 | 6.1 | 10.0 | 85.5 | **144.4(6.5%)** |

**TABLE 3.** Comparison of the crowd localization performance of point detection methods on NWPU test dataset.

| Methods | F1-Score | Precision | Recall |
|---|---|---|---|
| RAZ [29] | 0.599 | 0.666 | 0.543 |
| CLTR [13] | 0.694 | 0.676 | 0.685 |
| P2P-net [10] | 0.712 | **0.729** | 0.695 |
| PSL-Net($\gamma = 44$) | **0.727** | 0.719 | **0.735** |

**TABLE 4.** Experimental result with respect to centerness.

| Score (Th >0.5) | MAE | MSE |
|---|---|---|
| $\hat{P} \times \hat{C}^2$ | 50.96 | 79.58 |
| $\hat{P} \times \hat{C}$ | 50.86 | 79.79 |
| $\hat{P} \times \sqrt{\hat{C}}$ | **49.97** | **77.67** |
| $\hat{P} \times \sqrt[3]{\hat{C}}$ | 50.14 | 78.01 |

could degrade the performance, implying that the essential factor for classification is $\hat{P}$ in the end. Thus, the key effect of $\hat{C}$ could be interpreted as the auxiliary weight to produce the more reliable predictions among the candidates close to the GT.

### 2) EFFECT OF MATCHING PROCESS CONFIGURATION

Then, as shown in Table 5, we evaluate the effect of the distance metric and the cardinality of matching used between GT and anchor point in the label assignment during the training process of the proposed method. In terms of the distance metrics, we observe the difference between L2-Distance used for matching by P2P-Net [10], and DIoU of the

**TABLE 5.** Experimental result with respect to label assignment algorithm.

| Matching Cardinality | Distance metric | MAE | MSE |
|---|---|---|---|
| 1:1 | L2 distance | 52.66 | 80.36 |
| Partial N:1 | L2 distance | 55.87 | 86.54 |
| 1:1 | DIoU with PSL | 52.69 | 81.03 |
| Partial N:1 | DIoU with PSL | **49.97** | **77.67** |

proposed method, and in terms of the matching cardinality, we observe the difference between one-to-one matching and partial many-to-one matching. The experiments with other matching cardinalities are not implemented because, in the case of one(many)-to-many matching, the accuracy decreases due to the indiscriminately predicted positive sample, while in the case of many-to-one matching, the negative sample is excessively increased by sparse prediction, as indicated in P2P-Net [10]. In terms of one-to-one matching, our experimental results show similar performance regardless of the distance metrics. We guess that this is due to the fact that anchor points are assigned GT based on their distance in both metrics. We also analyze that this might decrease the reliability with respect to crowd localization, as it might contain the pairs that are far from each other. In contrast, in terms of our proposed partial many-to-one, the results show that there is a huge gap between using DIoU and using L2-distance, as the performance with DIoU is increased while the performance with L2-distance is decreased. We interpret this result that the relative distance by DIoU is more effective than the absolute pixel-level distance by L2-distance, when using the proposed method that allows the repeated anchor

points of single GT. Also, we suggest that it is effective that DIoU is based on the central distance.

#### 3) EFFECT OF PSEUDO SQUARE LABEL

We explore the effect of the proposed PSL on the bbox regression, as shown in Table 6. We compare the three types of bbox labels. The ablation study on the bbox label used the following settings: PSL generated by randomly selecting $\gamma$ from the natural numbers in the range of 18 to 44, and the man-made label fitting individual's head annotated by human, which provided by NWPU benchmark dataset, and PSL generated by setting $\gamma$ to 44. As a result, our proposed method with static PSL achieved the best performance based on the overall metrics. In terms of F1-score, the proposed method improves by 3% compared to random PSL and by 11% compared to the man-made labels. The man-made labels have the strict matchable boundary for small scale features in comparison to PSL, while they have the relaxed matchable boundary for large scale features. Thus, frequent false positives are caused by the strict man-made labels lacking background or foreground information, since the majority of crowd images have small heads, which makes it difficult to accurately identify individuals.

**TABLE 6. Experimental result with respect to Pseudo square label.**

| label | F1-Score | Precision | Recall |
|---|---|---|---|
| Man-made Bbox Label | 0.615 | 0.568 | 0.671 |
| Pseudo Square Label(random) | 0.691 | 0.717 | 0.667 |
| Pseudo Square Label(static) | **0.727** | **0.719** | **0.735** |

### D. ANALYSIS

The purpose of our proposed PSL-Net is to accurately detect individuals in a crowd with responsibility, which is based on anchor-free object detection where the region of the detection is auxiliary. Therefore, PSL based on $\gamma$ intends to provide the appropriate region for PSL-Net to accurately predict the center point of the head, instead of providing the exact region of the individual head.

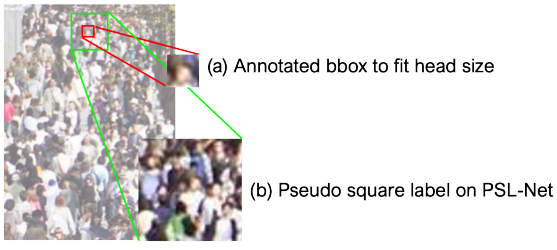



**FIGURE 6. Visual state of a person within the fit bbox(a) and within our PSL(b).**

In Fig. 6(a), It is challenging the supervision by the label (a) without the contextual information of the background, as it is not clearly identified that there is a person when annotated to fit the head size. Thus, we analyze that PSL-Net does not require optimization or annotation for the

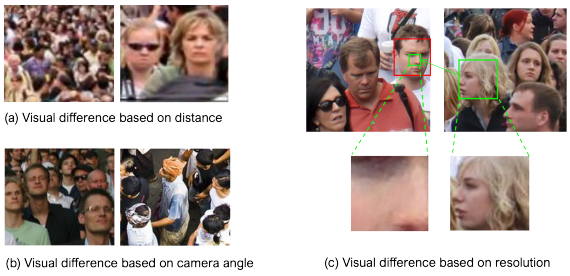



**FIGURE 7. Visual state in crowd images based on attributes.**

adaptation of PSL to fit the head size, which is demonstrated by the ablation study. However, this doesn't mean that $\gamma$ always has to be large. The large $\gamma$ makes the supervision of the proposed method challenging, as there are multiple instances that do not require the background information in the same PSL. Therefore, it is essential for PSL-Net to assign the appropriate value to $\gamma$.

We review the effect of $\gamma$ with respect to the benchmark datasets, as shown in Fig. 7, 8. Each of the benchmark datasets has differences in their attributes of distance, angle and resolution, which affect the distribution of head size in the crowd images. In accordance with the distance from camera to the individuals, we observe the various scales of the same image, as shown in Fig. 7(a). Even with the same resolution, the prediction error in the left image could significantly reduce the performance because the image is dense and lacks human-like visual features. On the contrary, the right image is sparse and intuitively identifiable as human, thus the prediction error has a negligible impact on the overall performance. In Fig 7.(b), according to the camera angle, we observe that the visual features of the heads are clearly different in spite of the similar scale. Fig. 7(c) shows the two images that differ in resolution. The scale of the heads in both images seems to be similar. Nevertheless, we notice that it is more difficult to extract the visual feature of the head from the left high-resolution image ($560 \times 560$) than from the right low-resolution image($160 \times 160$) when it is cropped as a patch of the same size ($60 \times 60$). Thus, we assign different $\gamma$ to each benchmark dataset due to the imbalance of information caused by the attributes mentioned above.

As shown above, we present the result of experiments that PSL-Net is supervised under the three different $\gamma$ of PSL. Each $\gamma$ of the PSL-Net is determined through the grid search from 16 to 48, considering the attributes of each benchmark dataset. The result shows that PSL-Net($\gamma = 18$) outperforms on SHTA, which has high crowd density and low resolution, PSL-Net($\gamma = 24$) outperforms on SHTB, which has small resolution but relatively large head size due to low density of crowd, and PSL-Net($\gamma = 44$) outperforms on UCF-QNRF, which has relatively large head size and high resolution. In Fig. 8, it can be seen that the visualization of the three types of PSL which achieve optimal performance on each benchmark dataset. Notably, the results show that PSL on the

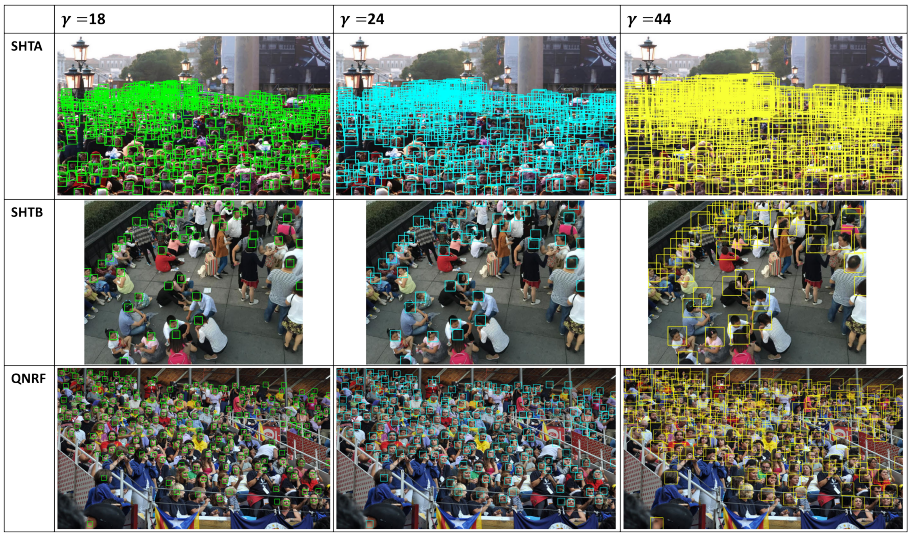



**FIGURE 8.** Visualization result of the three types of PSL on the representative image of each benchmarks.

representative image of each benchmarks contains generally visible features, which means the crowds in the image of SHTA, SHTB, and QNRF are mostly covered but not much overlapped by PSL with $\gamma = 18, 24, 44$, respectively. Since the image of QNRF has very high resolution, it can be observed that PSL is about twice as large compared with SHTA, while it is visually appropriate.

However, we found that the grid search or hyperparameter tuning based on heuristics are inevitable, when it applied to new datasets. Furthermore, in terms of label assignment process, we found that the most of the target value of the centerness is close to upper limit, following chi-square distribution. Thus, In future research, we will devise to determine $\gamma$ of PSL adaptively and to make the target value of centerness follows the normal distribution for the purpose of robustness on the various resolution.

## V. CONCLUSION

We proposed a new crowd counting and localization method PSL-Net based on pseudo square bbox label, which is responsible for prediction through anchor-free detection and precisely localizing individuals in crowd by the centerness estimation. PSL-Net preserves responsibility of the predictions by indirect detection of the outside responsible grid, unlike the existing point-based detection methods that directly detect the coordinates of the heads. In addition, PSL-Net estimates the centerness and bbox for the precise localization, and includes a partial many-to-one algorithm to match GT with the closest anchor point as possible. As a result, PSL-Net achieved state-of-the-art on ShanghaiTech Part A and Part B datasets, which are the most popular datasets in crowd counting, and partially superior than state-of-the-art on QNRF dataset. Furthermore, PSL-Net outperformed other point-detection based methods on the NWPU datasets, which is also a widely used dataset for crowd localization. In future research, we will expand our method to enhance its robustness on multi-scale features, including adaptive determination of $\gamma$ and improving the representativeness of the centerness.

## REFERENCES

[1] H.-R. Lee. *Seoul to Refurbish Pavements, Set Up Cameras in Itaewon to Help Crowd Flow*. Koreatimes. Accessed: Apr. 16, 2024. [Online]. Available: https://www.koreatimes.co.kr/www/nation/2024/04/281_349582.html

[2] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1522–1530.

[3] Y. Liu, M. Shi, Q. Zhao, and X. Wang, "Point in, box out: Beyond counting persons in crowds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6462–6471.

[4] D. B. Sam, S. V. Peri, M. N. Sundararaman, A. Kamath, and R. V. Babu, "Locate, size, and count: Accurately resolving people in dense crowds via detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2739–2751, Aug. 2021.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[6] R. Stewart, M. Andriluka, and A. Y. Ng, "End-to-end people detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2325–2333.

[7] J. Gao, T. Han, Q. Wang, Y. Yuan, and X. Li, "Learning independent instance maps for crowd localization," 2020, *arXiv:2012.04164*.

[8] S. Abousamra, M. Hoai, D. Samaras, and C. Chen, "Localization in the crowd with topological constraints," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 2, pp. 81–872.

[9] J. Gao, M. Gong, and X. Li, "Congested crowd instance localization with dilated convolutional Swin transformer," *Neurocomputing*, vol. 513, pp. 94–103, Nov. 2022.

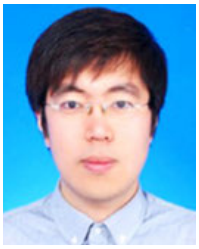
[10] Q. Song, C. Wang, Z. Jiang, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Wu, "Rethinking counting and localization in crowds: A purely point-based framework," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3345–3354.

[11] H.-Y. Ma, L. Zhang, and X.-Y. Wei, "FGENet: Fine-grained extraction network for congested crowd counting," in *MultiMedia Modeling*. Cham, Switzerland: Springer, 2024, pp. 43–56.

[12] Y. Wang, J. Hou, X. Hou, and L.-P. Chau, "A self-training approach for point-supervised object detection and counting in crowds," *IEEE Trans. Image Process.*, vol. 30, pp. 2876–2887, 2021.

[13] D. Liang, W. Xu, and X. Bai, "An end-to-end transformer model for crowd localization," in *Proc. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer Nature, 2022, pp. 38–54.
[14] Z.-Q. Cheng, "Rethinking spatial invariance of convolutional networks for object counting," 2024, *arXiv:2206.05253*.
[15] P. Thanasutives, K.-I. Fukui, M. Numao, and B. Kijsirikul, "Encoder–decoder based convolutional neural networks with multi-scale-aware modules for crowd counting," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2382–2389.
[16] Y. Ma, V. Sanchez, and T. Guha, "Fusioncount: Efficient crowd counting via multiscale feature fusion," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 3256–3260.
[17] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100.
[18] B. Wang, H. Liu, D. Samaras, and M. Hoai, "Distribution matching for crowd counting," in *Proc. Adv. Neural Inf. Process. Syst.*, Sep. 2020, pp. 1–13.
[19] Y. Wang, X. Hou, and L.-P. Chau, "Dense point prediction: A simple baseline for crowd counting and localization," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2021, pp. 1–6.
[20] Y. Wang, G. Li, Q. Zhang, J. Kim, and H. Li, "Perspective-aware density regression for crowd counting," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1214–1218.
[21] Y. Zhang, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 589–597.
[22] L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang, "Multi-scale convolutional neural networks for crowd counting," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 465–469.
[23] M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting perspective information for efficient crowd counting," 2024, *arXiv:1807.01989*.
[24] J. Gao, T. Han, Y. Yuan, and Q. Wang, "Domain-adaptive crowd counting via high-quality image translation and density reconstruction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 4803–4815, Aug. 2023.
[25] N. H. Tran, "Improving local features with relevant spatial information by vision transformer for crowd counting," in *Proc. Brit. Mach. Vis. Conf.*, vol. 729, 2022, pp. 1–14.
[26] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, "TransCrowd: Weakly-supervised crowd counting with transformers," *Sci. China Inf. Sci.*, vol. 65, no. 6, Jun. 2022, Art. no. 160104.
[27] X. Chen and H. Lu, "Reinforcing local feature representation for weakly-supervised dense crowd counting," 2022, *arXiv:2202.10681*.
[28] J. Cheng, H. Xiong, Z. Cao, and H. Lu, "Decoupled two-stage crowd counting and beyond," *IEEE Trans. Image Process.*, vol. 30, pp. 2862–2875, 2021.
[29] C. Liu, X. Weng, and Y. Mu, "Recurrent attentive zooming for joint crowd counting and precise localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1217–1226.
[30] R. Sundararaman, C. De Almeida Braga, E. Marchand, and J. Pettré, "Tracking pedestrian heads in dense crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3864–3874.
[31] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, 1955.
[32] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, Art. no. 12993.
[33] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, May 2018, pp. 1–11.
[34] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020.
[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[36] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[37] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.
[38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[39] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2018, pp. 544–559, doi: 10.1007/978-3-030-01216-8_33.
[40] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-crowd: A large-scale benchmark for crowd counting and localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2141–2149, Jun. 2021.
[41] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," 2024, *arXiv:1904.01355*.

**JIHYE RYU** received the B.S. degree in human-centered artificial intelligence and the M.S. degree in artificial intelligence and informatics from Sangmyung University, Seoul, South Korea, in 2020 and 2023, respectively. Since 2023, she has been a Researcher with INFINIQ Corporation, Seoul. She wrote some papers and articles related to network scheduling using deep reinforcement learning and face recognition using CNN. Her current research interests include abnormal detection in video surveillance, and oriented object detection using CNN and visual transformer.



**KWANGHO SONG** received the B.S. and M.S. degrees in information and communication engineering and the Ph.D. degree in electric and computer science engineering from Inha University, Incheon, Republic of Korea, in 2015, 2017, and 2022, respectively. Since 2022, he has been a Senior Researcher with INFINIQ Corporation, Seoul, Republic of Korea. He wrote some papers and articles related to video surveillance, such as person re-identification, violence detection, and kidnapping detection, using conventional machine learning and deep learning. His current research interests include the 3D semantic segmentation and abnormal contents detection using multi-modal features.