Densely-packed Object Detection via Hard Negative-Aware Anchor Attention

Sungmin Cho[1] Jinwook Paeng[1] Junseok Kwon[2]
School of Computer Science and Engineering, Chung-Ang University, Seoul, Korea
[1]{csm8167, nuggy875}@naver.com, [2]jskwon@cau.ac.kr

Abstract

In this paper, we propose a novel densely-packed object detection method based on advanced weighted Hausdorff distance (AWHD) and hard negative-aware anchor (HNAA) attention. Densely-packed object detection is more challenging than conventional object detection due to the high object density and small-size objects. To overcome these challenges, the proposed AWHD improves the conventional weighted Hausdorff distance and obtains an accurate center area map. Using the precise center area map, the proposed HNAA attention determines the relative importance of each anchor and imposes a penalty on hard negative anchors. Experimental results demonstrate that our proposed method based on the AWHD and HNAA attention produces accurate densely-packed object detection results and comparably outperforms other state-of-the-art detection methods. The code is available at here.

1. Introduction

Object detection is one of the fundamental computer vision tasks, which solves localization and classification problems in a supervised manner. With the advances of deep neural networks (DNNs), recent object detection methods have achieved large improvements in terms of speed and accuracy. These methods have also obtained high performance metrics at various scales, image resolutions, and object densities [2, 19]. Accordingly, object detection has been widely applied to various practical applications such as autonomous driving [17] and surveillance [24] under real-world environments. In particular, this object detection method has been used to develop self-checkout [3] and inventory management [33] systems in marts, convenience stores, wholesale, and retail stores [23]. As the need for product detection, recognition, and inventory management systems in a retail store has been growing, dense object detection datasets (*e.g.* WebMarket [36], CAPG-GP [9], Holoselecta [8] and SKU-110K [12]) as well as conventional object detection datasets [5, 7, 20] have been published, which contain a large number of densely-packed products in retail stores. With these datasets, a densely-packed object detection problem has emerged recently.

(*a*) (*b*)

Figure 1. **Densely-packed object detection results (a) without/ (b) with the proposed HNAA attention.** Our detection method can accurately detect a large number of densely-packed small-size objects using the proposed HNAA attention.

The densely-packed object detection problem is more challenging than conventional object detection problems, because we should deal with a large number of objects and these objects are very small in size. As shown in Fig.2, conventional object detection datasets (*e.g.* VOC [7] and MSCOCO [20]) contain a small number of objects (*e.g.* less than 10 objects in average) in an image. In contrast, densely-object object detection datasets [9, 12, 36] have a relatively large number of objects (*e.g.* more than 10, 50 and 150 objects according to datasets). In particular, the SKU-110K dataset has more than 500 objects in a single image. In addition, these datasets inevitably contain very small-size objects, as a large number of objects exist simultaneously in a single image. In this case, we cannot extract sufficient fea-

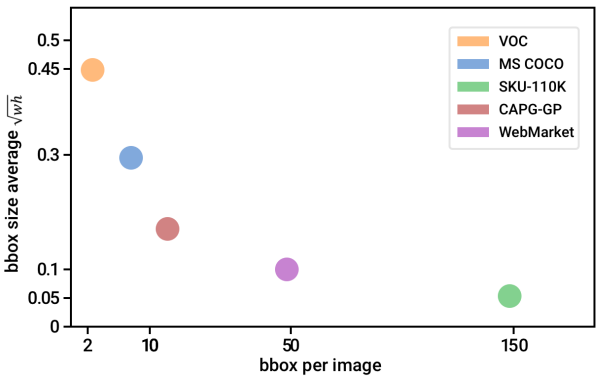


Figure 2. **Densely-packed object detection datasets have a large number of small-size objects.** The x-axis represents the average number of objects per image, and y-axis is the average size of object bounding boxes. Conventional object detection datasets are located around the left-top area, whereas densely-packed object detection datasets are located around the right-bottom area.

tures from small-size objects to identify the objects. Moreover, because objects are closely located in each other, there exist many overlaps between them, and object detection performance is easily affected by other objects. The aforementioned challenges significantly degrade the performance of conventional object detection methods [12], when densely-packed object detection datasets are used.

To detect densely-packed objects, several methods [12, 15, 19] have utilized dense anchors, which can cover a large portion of areas in an image, thus can also find object areas. For example, RetinaNet [19] used more than 120k anchors for the 800 by 800 resolution image. However, these dense anchors can be redundant and can make the detection methods computationally heavy. To solve this problem, the detection methods need to consider the relative importance of each anchor, which helps to find accurate object areas and improve the detection accuracy. Because anchors are typically built based on the center points of the objects, it is crucial to estimate the distribution of the center points accurately and use it to determine the relative importance of each anchor. However, conventional anchor assignment strategies have two limitations. First, they do not take into account the relative importance between positive anchors. Second, they do not consider positive anchors with small IOU values as hard negative anchors during the training.

To overcome the aforementioned limitations, in this paper, we propose a novel densely-packed object detection method based on advanced weighted Hausdorff distance (AWHD) and hard negative-aware anchor (HNAA) attention. The AWHD solves the problems of conventional weighted Hausdorff distance (*i.e.* sensitive to outliers and approximation to original distance) and accurately estimates center area maps of the objects. The HNAA attention determines the importance of each anchor accurately using the normalized intersection of union (normalized IoU) values and considers hard negative anchors more importantly. Using the proposed AWHD and HNAA attention, our method can accurately detect densely-packed objects. Fig.1 shows the effectiveness of the proposed HNAA attention in densely-packed object detection problems.

The main contributions of the proposed method can be summarized as follows.

- We propose the AWHD, which improves the performance of conventional WHD. The proposed AWHD accurately estimates center area maps of the objects.
- We present the HNAA attention, which determines the importance of each anchor and consider hard negative anchors importantly.
- Our AWHD and HNAA attention can be easily plugged into existing detection networks and comparably outperforms state-of-the-art methods in densely-packed object detection benchmark datasets.

2. Related Work

Object Detection. Owing to the emergence of DNNs, we have recently witnessed considerable improvement in object detection performance. DNN-based object detectors can be divided into two categories, namely two-stage detectors and one-stage detectors.

In two-stage detectors, Girshick *et al*. [11] proposed regions with convolutional neural networks (R-CNN) that firstly applied DNNs to object detection problems. After the success of R-CNN, it was extended to several variants [4, 10, 13, 28]. For object detection, these methods adopted two states, in which one was to propose the set of candidate object regions and another was to classify the candidates. In one-stage detectors, Redmon *et al*. presented YOLO and its variants [25, 26, 27] using multiple anchors in the regular grid. Recently, Bochkovskiy *et al*. introduced YOLOv4 and Scaled-YOLOv4 [2, 34] and exhibited state-of-the-art performance. Liu *et al*. [22] developed a rapid and accurate one-stage detector called single shot multi-box detector (SSD), which applied anchors with different scales into multiple layers. Then, SSD became the baseline architecture for several detectors [18, 21, 35, 32] based on multi-scale feature fusion. Lin *et al*. [19] solved class imbalance problems in object detection by applying focal loss to feature pyramid networks [18]. However, Goldman *et al*. [12] confirmed that existing detectors have difficulty in handling dense objects and verified that conventional methods showed poor performance in densely-packed scenes.

In contrast, our method exhibits state-of-the-art performance even in these densely-packed scenes.

Densely-packed object detection. The SKU-110K dataset [12] has been released recently, which contains an enor-

mous number of objects that are densely packed in supermarket shelves. Unlike existing object detection datasets [5, 7, 20], this dataset has a dense line of products from various brands, which are close to each other. The dataset contains $8,233$ training images, 584 validation, $2,941$ testing images, and $1,733,678$ instances under various conditions. Goldman *et al*. [12] evaluated conventional detectors (*e.g*. Faster-RCNN [28], YOLOv2 [26], and RetinaNet [19]) using this dataset and enhanced RetinaNet by adding the soft IoU layer and applying the EM merger unit. Kant *et al*. [15] integrated the aforementioned two components [12] by learning Gaussian distributions for all objects. It learned the object centers and minimized the overlay between the objects to improve the accuracy in dense object environments.

In contrast, our method uses the objectiveness map based on HNAA attention for accurate object detection.

Hausdorff distance. The Hausdorff distance is the longest distance between two sets of points in a metric space [1]. This distance has been widely used for image matching or object matching problems [6, 37]. Recently, it has been used as a loss function of DNNs and applied to solve various computer vision problems such as medical image segmentation [16], object localization [30], and vehicle Re-Identification [38]. Karim *et al*. [16] estimated the segmentation probability maps based on the Hausdorff distance for image segmentation tasks. Ribera *et al*. [30] found the object locations using the weighted Hausdorff distance by modifying the average Hausdorff distance in object localization tasks. Zhao *et al*. [38] introduced a new video-based vehicle re-ID benchmark and presented the Pompeiu-hausdorff distance for video-to-video matching tasks.

Our method is inspired by [30] to estimate the objectiveness maps for anchor attention. However, in contrast to [30], we introduce an advanced distance function, which can deliver accurate detection results.

Anchor assignment. Dense object detection methods [2, 19, 22, 28] based on multiple anchor boxes determined the anchor boxes that contribute to the objective function and assigned positive or negative anchors. In particular, they computed the intersection of union (IoU) values between ground truth and anchor boxes. If the values exceeded a predefined threshold, 1 was assigned, otherwise, 0 was assigned, which induced a binary mask. Ren *et al*. [28] assigned positive (negative) anchors if the IoU value is greater (less) than 0.7. Liu *et al*. [22] utilized a different threshold of 0.5 to assign positive and negative anchors, while a certain portion of negative anchors was further assigned as hard negative anchors. YOLO-based methods [26, 27] assigned one positive anchor to an object, whereas YOLO v4 [2] assigned multiple positive anchors to an object. Lin *et al*. [19] used almost anchors with IoU values of more than 0.5 or less than 0.4 to avoid the situation of being overwhelmed by easy negative samples during the classification training process. However, they used only positive anchors with IoU of more than 0.5 for the localization training.

(a) (b) (c)

Figure 3. **Comparison of WHD and AWHD.** (a), (b) and (c) show input images, estimated center area maps using the WHD and the AWHD, respectively. The proposed AWHD accurately and compactly finds center areas for densely located objects.

In contrast, our method adopts the IoU assignment that considers the relative importance of each anchor, thus enables hard negative-aware attention.

3. Proposed Method

The proposed method can detect densely-packed objects accurately using a novel attention method for the anchor assignment. In particular, our method extracts the object center area from an image and focuses on the anchor corresponding to the center area. For this, the method accurately extracts a precise and distinguishable center area of each object. Subsequently, it focuses on an important anchor based on the center area. In this section, to achieve the aforementioned goal, we propose the AWHD and HNAA attention in Sections 3.1 and 3.2, respectively.

3.1. Advanced Weighted Hausdorff Distance

The proposed AWHD, which complements the conventional weighted Hausdorff distance [30], accurately estimates the center area map for each small-size object. Then, the estimated center area map is used for anchor attention.

Weighted Hausdorff distance (WHD). The Hausdorff distance is the longest distance between two non-empty sets. Let X and Y be two non-empty subsets of the metric space (M, d), then the Hausdorff distance can be computed as

$$d_H(X,Y) = \max\left\{\sup_{x\in X} d(X,y), \sup_{y\in Y} d(x,Y)\right\}, \quad (1)$$

where $d(a,B) = \inf_{b\in B} d(a,b)$ and d can be any metric. If X and Y are finite sets, the average Hausdorff distance

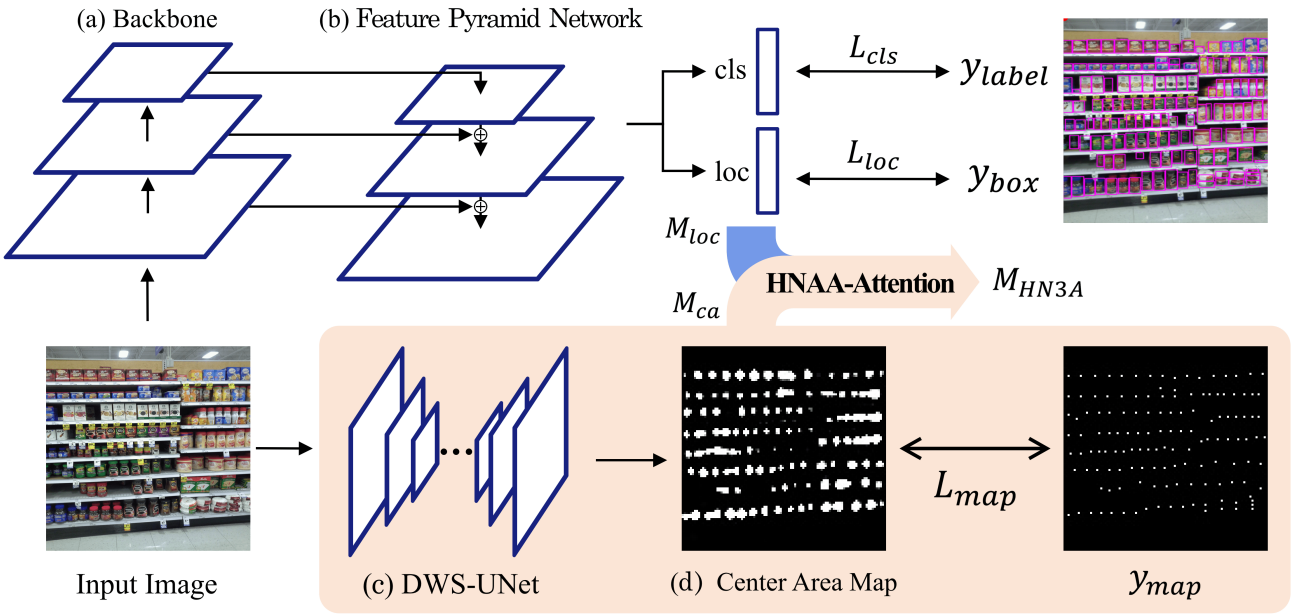


Figure 4. **Pipelines of the proposed method.** For object detection, we extract features using the backbone network in (a), and expand them into multi-scale features through FPN in (b). Simultaneously, our DWS-UNet in (c) estimates the center area map in (d), using the proposed AWHD and HNAA attention.

$d_{AH}(X,Y)$ is defined as follows:

$$d_{AH}(X,Y) = \frac{1}{|X|}\sum_{x\in X}\min_{y\in Y} d(x,y) + \frac{1}{|Y|}\sum_{y\in Y}\min_{x\in X} d(x,y), \tag{2}$$

where $d_{AH}(X,Y)$ alleviates the problem of $d_H(X,Y)$ which is sensitive to outliers [30]. The aforementioned average Hausdorff distance can be extended as the loss for DNNs in the object location task [30], which induces the WHD. Let $x \in \Omega$ be a possible pixel coordinate and Y be the set of ground truth pixel coordinates. Given Y, DNNs estimate the probability map $p = \{p_x\}$ based on the WHD, which describes the position of the object (*i.e.* $p_x = 1$ if $x \in Y$ and 0 otherwise). Then, the WHD is defined as follows:

$$d_{WHD}(p,Y) = \frac{1}{S+\epsilon}\sum_{x\in\Omega} p_x \min_{y\in Y} d(x,y) + \frac{1}{|Y|}\sum_{y\in Y} M_{\alpha\,x\in\Omega}\left[p_x d(x,y) + (1-p_x)d_{\max}\right], \tag{3}$$

where $M_\alpha(x_1,...,x_n) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^\alpha\right)^{\frac{1}{\alpha}}$ is the generalized mean function that can approximate the min function in (2) and avoid non-differentiability of the min function, and $\epsilon \approx 10^{-6}$ induces numerical stability during the training time. In (3), $S = \sum_{x\in\Omega} p_x$ and d_{max} is the largest distance in the image coordinate.

Advanced weighted Hausdorff distance (AWHD). The WHD introduces the generalized mean function M_α in (3), of which alpha should be -infinity or its approximation values. However, in practical implementation, M_α is fixed to a small negative number (*e.g.* -1) because it makes DNNs train stably. In this paper, we argue that the gap between mathematical formulations and practical implementation can induce inevitable performance drops. Moreover, because M_α is originally the approximation of the min function, M_α can result in poor performance. We also argue that $p_x \min d(x,y)$ in (3) does not exactly use the average distance, which can make the distance function sensitive to outliers like the original Hausdorff distance in (1). To complete these arguments, we conducted experiments, as shown in Fig.3, which demonstrates that we can obtain more accurate object center map when using the min function instead of M_α (*i.e.* $\min \leftarrow M_\alpha$) and using $\min p_x d(x,y)$ instead of $p_x \min d(x,y)$ (*i.e.* $\min p_x d(x,y) \leftarrow p_x \min d(x,y)$).

To solve the problems of the WHD, we present a novel AWHD while answering the following questions.

- *Question 1:* How to make the min function differentiable even without using M_α in practice?
- *Question 2:* How to modify $p_x \min d(x,y)$ to be an average distance?

$\rightarrow$ *Answer 1:* Because the min function selects only the smallest value and ignores the others, the min function can be considered as the identity operation for that one element. Thus, we can make the min function differentiable by making the gradient flow backward through the function for just that one element.

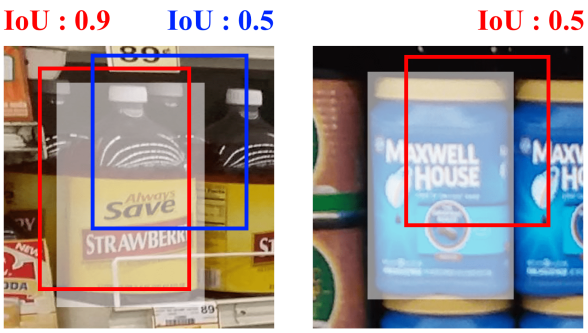


Figure 5. **Examples of anchors with various IoU values.**

$\rightarrow$ *Answer 2:* By substituting $p_x \min d(x,y)$ with $\min p_x d(x,y)$, we can induce the average distance $p_x d(x,y)$. If $d(x,y)$ is the Euclidean distance, $p_x \min d(x,y) = \min p_x d(x,y)$. However, if $d(x,y)$ is defined in arbitrary manifolds, $p_x \min d(x,y) \neq \min p_x d(x,y)$. Thus, $\min p_x d(x,y)$ exploits more accurate geometric characteristics from the average distance.

Then, the proposed AWHD can be formulated as

$$d_{AWHD}(p,Y) = \frac{1}{S+\epsilon} \sum_{x \in \Omega} \min_{y \in Y} p_x d(x,y) + \frac{1}{|Y|} \sum_{y \in Y} \min_{x \in \Omega} \left[p_x d(x,y) + (1-p_x) d_{max} \right], \tag{4}$$

where $p_x \min d(x,y)$ and M_α in (3) are changed into $\min p_x d(x,y)$ and $\min$, respectively.

Training of the AWHD. The proposed AWHD can be plugged into any existing detection network (backbone in Fig.4) and can be trained in parallel (the red box in Fig.4). To estimate the center area map, we adopt U-net [31] using light-weight depth-wise-separable convolutions [14] (denoted as DWS-UNet in Fig.4) so that it can be lightly added to existing detection networks. Subsequently, the center area map, which is estimated by DWS-UNet based on the proposed AWHD, is used to obtain the hard negative-aware anchor attention mask and accurately compute the detection loss, which is explained in the next section.

3.2. Hard Negative-Aware Anchor Attention

Using the center area map (estimated by the AWHD) and the normalized IoU mask (that is obtained by extending existing assignment methods), our method estimates a novel HNAA attention mask, which can overcome the following limitations of conventional anchor assignment strategies.

Limitations of conventional anchor assignment strategies. The loss functions of conventional anchor-based detection networks [19, 22, 2] can be typically formulated as

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{loc} + \alpha, \tag{5}$$

where $\mathcal{L}_{cls}$ is the classification loss for object classes (*e.g.* focal loss [19], cross entropy [22], and binary cross entropy [27]), $\mathcal{L}_{loc}$ is the location loss for object bounding boxes (*e.g.* smooth L1 [28, 22, 19] and IoU loss [2]), and α is an additional loss for each algorithm.

To compute the loss values in (5), conventional anchor-based detection networks utilize multiple candidates of bounding boxes and then compare them with ground-truth bounding boxes. In particular, the aforementioned candidates are assigned as positive or negative anchors by evaluating the contribution to the loss function during the training process. The evaluation process is typically performed based on the IoU values between anchors and ground-truth bounding boxes. Then, the evaluation score is implemented as a mask with binary values of $\{0,1\}$ and incorporated into the loss function by multiplying the mask with the loss. Thus, the detection loss function can be extended as

$$\mathcal{L}_{det} = M_{cls} \times \mathcal{L}_{cls} + M_{loc} \times \mathcal{L}_{loc} + \alpha, \tag{6}$$

where M_{cls} and M_{loc} denote the masks for classification and location losses, respectively.

However, conventional anchor assignment strategies have two main limitations. 1) conventional strategies do not consider the relative importance between positive anchors. For example, an anchor with the IoU value of 0.9 needs to be handled more important than an anchor with the IoU of 0.5. 2) negative anchors with large IoU values need to be considered as hard negative anchors. Moreover, because objects are densely located, these hard negative anchors should be carefully handled to improve accuracy.

Normalized IoU assignment. To overcome the first limitation above, we present a novel anchor assignment method, in which the importance of each anchor is considered to make the mask based on the IoU value. In Fig.5 (a), the white box is the ground-truth box, and blue and red boxes are anchors with IoU of 0.5 and 0.9, respectively. In this case, conventional strategies assign the mask value of 1 to both the blue and red anchors. Alternatively, if the IoU value is directly used as the mask value, the red anchor has the mask value of 0.9, which is higher than that of the blue anchor. Thus, the relative importance can be considered, if each assigned mask uses its own IoU value $\in [0,1]$. However, if there exists only a single anchor with the IoU value of 0.5 in Fig.5 (b), it does not make sense to assign the mask value of 0.5 to the anchor. To solve this problem, we present a normalized IoU assignment method, in which the IoU value is assigned to the mask and then is divided by the maximum IoU value in an image. Fig.6 shows python-type pseudo-code of our anchor assignment method.

Hard negative-aware anchor attention. To overcome the second limitation of conventional strategies, we present a novel anchor attention method, in which hard negative samples are intensively considered. Fig.7 shows the average

```
def normalized_iou_assign(G, A):
    # G : gt_boxes : (N, 4)
    # A : anchor_boxes : (B, 4)

    # compute IoU between gt and anchor
    iou = IoU(G, A)                 # (N, B)
    # ious per anchors
    max_iou_a = iou.max(dim=1)      # (B)
    # indices per anchors
    max_ids_a = iou.argmax(dim=1)   # (B)
    # ious per objects
    max_iou_o = iou.max(dim=0)      # (N)
    # normalized ious per anchors
    normalized_iou = max_ious_a
    / max_iou_o[max_idxs_a]         # (B)
    return normalized_iou
```

Figure 6. **Pseudo-code of the normalized IoU assignment.**

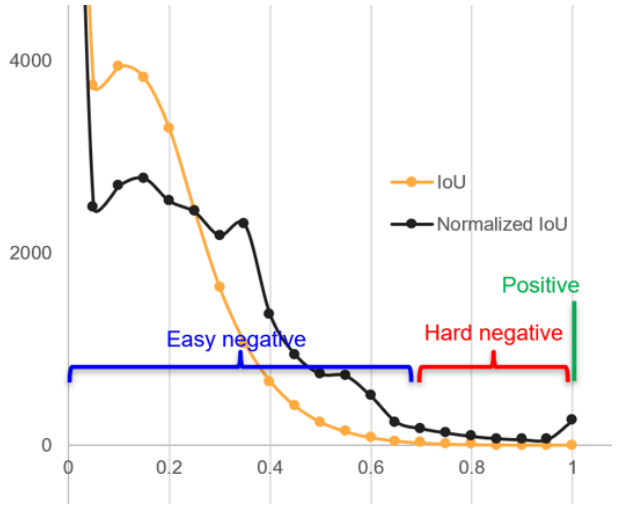


Figure 7. **Number of anchors according to different IoU values** before and after applying normalized IoU assignment method.

number of anchors according to different IoU values before and after applying the normalized IoU assignment method using the SKU-110K dataset. With the normalized IoU assignment method, anchors can have relative importance, which increases the number of anchors with large IoU values and decreases the number of anchors with small IoU values. Thus, it makes a large number of hard negative anchors with the IoU values $\in [0.7, 1)$ and enables to use them during the training process. By doing this, we can consider negative anchors with large IoU values more importantly and anchors with small IoU values less importantly.

Let the object center area map estimated by the proposed AWHD be M_{ca} and the normalized IoU mask that is obtained by extending existing assignment methods be M_{nia}. Using M_{ca} and M_{nia}, we estimate a novel HNAA attention mask (*i.e.* M_{HN3A}), as follows.

$$M_{HN3A} = \exp\left(M_{nia}^{n}\right) M_{nia} \times \exp\left(M_{ca}\right), \quad (7)$$

where M_{nia}^{n} denotes M_{nia} to the power of n, which is empirically determined in experiments. In (7), M_{HN3A} has two terms. The first term $\exp(M_{nia}^{n}) \times M_{nia} \in [0, e]$ gives a penalty to hard negative anchors. If their IoU values are closer to 1, this term amplifies the importance of hard negative anchors. Otherwise, the term exponentially lowers the importance of hard negative anchors. The second term $\exp(M_{ca})$ is used so that the center areas of the anchors are more emphasized, in which exp forces M_{ca} to have values in the increased range of $[1, e]$ compared to $[0, 1]$. Thus, our method can more accurately discriminate between hard negative and positive anchors by carefully examining the increased values of M_{ca}. Overall, M_{HN3A} estimates attention values in the range of $[0, e^2]$.

3.3. Total Loss

Our total loss extends the conventional anchor-based loss in (6), as follows.

$$\mathcal{L}_{total} = \alpha \times M_{cls}\mathcal{L}_{cls} + \beta \times M_{HN3A}\mathcal{L}_{loc} + \gamma \times \mathcal{L}_{map}, \quad (8)$$

where we use the focal loss for $\mathcal{L}_{cls}$ and conventional focal loss classification masks [19] for M_{cls}. For the localization loss $\mathcal{L}_{loc}$, we adopt the GIoU loss proposed by [29], where the importance of each anchor is considered using the proposed M_{HN3A} in (7). In (8), $\mathcal{L}_{map}$ is computed based on the proposed AWHD in (4). In experiments, α, β, and γ were empirically fixed to $1, 1$, and 0.01, respectively.

4. Experimental Results

We experimentally demonstrate the effectiveness of the proposed method. By conducting an ablation study on each component of our method, we show that each component is useful for densely-packed object detection. In addition, we empirically determine the value of n in (7). Then, we show that our proposed anchor attention based on the AWHD is superior to conventional WHD in densely-packed object detection problems and demonstrate that our detection method outperforms other state-of-the-art methods using various densely-packed object detection benchmarks.

4.1. Settings

Implementation details. We used RetinaNet [19] as the baseline and added the proposed DWS-UNet into the baseline, as shown in Fig.4. The network architecture was described in detail in supplementary materials. We used the SGD optimizer, where the total number of epochs, the initial learning rate, and the weight decay were set to 30, $2e-3$,

Table 1. **Ablation 1: contribution of each component.** Our method consists of four components (*i.e*. Baseline: RetinaNet [19], GIoU-Localization: the localization loss [29], $\mathcal{L}_{loc}$, in (8), NIoU-Assignment and HN3A-Masking: the importance of each anchor, M_{HN3A}, in (8)). We conducted experiments for four cases using six evaluation metrics. The best results were written in boldface.

	case1	case2	case3	case4
Baseline	✓	✓	✓	✓
GIoU-Localization		✓	✓	✓
NIoU-Assignment			✓	✓
HN3A-Masking				✓
AP	0.455	0.485	0.515	**0.522**
$AP^{.50}$	-	0.880	0.895	**0.897**
$AP^{.75}$	0.389	0.492	0.544	**0.556**
AR_{300}	0.530	0.563	0.595	**0.601**
$AR_{300}^{0.50}$	-	0.923	0.933	**0.935**
$P^{R=.5}$	0.544	0.740	0.807	**0.816**

Table 2. **Ablation study 2: distance functions.**

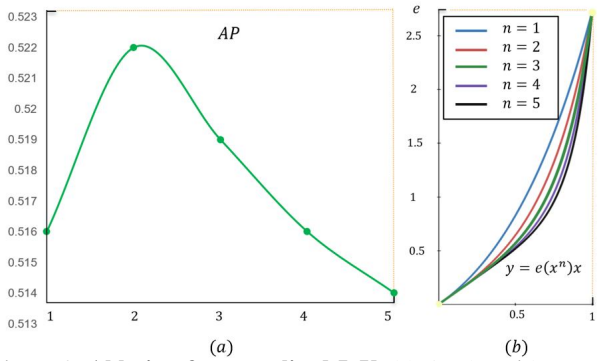
Distance	AP	$AP^{.50}$	$AP^{.75}$	AR_{300}	$AR_{300}^{0.50}$	$P^{R=0.5}$
WHD [30]	0.504	0.889	0.525	0.594	0.931	0.786
Our AWHD	**0.522**	**0.897**	**0.556**	**0.601**	**0.935**	**0.816**

and $1e-4$, respectively. The batch size was set to 1, but the learning rate decay was not used. For a fair comparison with conventional methods, data augmentation was not performed. As in [12, 15], a resolution was resized so that minimum and maximum sizes were 800 and 1333, respectively. We developed our method using windows 10 64-bit platform with Intel CPU i7 3.60 GHz and one NVIDIA GeForce GTX 2080 Ti. Python 3.7 and Pytorch 1.7.0. were used.

Evaluation metrics and datasets. For comparisons, we adopted six evaluation metrics from [12, 15, 20] (*i.e*. $\mathbf{AP}$: average precision at IoU= .50 : .05 : .95, $\mathbf{AP^{.50}}$: average precision at IoU= .50, $\mathbf{AP^{.75}}$: average precision at IoU= .75, $\mathbf{AR_{300}}$: average recall at IoU= .50 : .05 : .95 in the maximal 300 number of objects, $\mathbf{AR_{300}^{0.50}}$: average recall at IoU= .50, and $\mathbf{P^{R=0.5}}$: precision-recall curve at recall= 0.5 for IoU= 0.75). For densely-packed object detection and ablation studies, we used the SKU-110K dataset [12], which consists of 8233 training data, 588 validation data, and 2941 testing data. During the training time, we excluded erroneous data that could cause the loss explosion. We also used additional benchmarks, Web Market [36], Holoselecta [8], and GAPG-GP [9].

4.2. Ablation Study

Performance of each component. Table 1 shows the contribution of each component (*i.e*. Retinanet [19], GIoU-Localization [29], NIoU-Assignment, and HN3A-Masking) of our method to the success of densely-packed object detection. This experiment was conducted using the SKU-110K test dataset by adding four-component one by one. As shown in Table 1, our normalized IOU assignment and HN3A masking methods considerably improved the detection accuracy in terms of all evaluation metrics, which verifies the effectiveness of the proposed components.



(a) (b)

Figure 8. **Ablation 3: normalized IoU.** (a) AP (y-axis) according to different values of n (x-axis) in $\exp(M_{nia}^n)M_{nia}$ of (7), (b) penalty values on hard negative anchors (y-axis), which is computed by $\exp(M_{nia}^n)M_{nia}$, according to the IoU values (x-axis).

Table 3. **Ablation 4: hard negative-aware anchor attention.** Performance of M_{nia}^n in (7) according to different values of n.

n	AP	$AP^{.50}$	$AP^{.75}$	AR_{300}	$AR_{300}^{0.50}$	$P^{R=0.5}$
1	0.516	0.896	0.550	0.598	0.934	0.806
2	**0.522**	**0.897**	**0.556**	**0.601**	**0.935**	**0.816**
3	0.519	0.897	0.553	0.600	0.934	0.814
4	0.516	0.894	0.549	0.597	0.933	0.808
5	0.514	0.894	0.544	0.599	0.933	0.801

Performance of distance functions. Fig.3 shows the effectiveness of the proposed AWHD. As shown in Fig.3, the AWHD produced more accurate object center area map qualitatively compared the conventional WHD. In addition, our AWHD quantitatively outperformed the conventional WHD in terms of all evaluation metrics, as shown in Table 2. These results demonstrate that object center areas should be accurately detected for precise densely-packed object detection because anchor-based object detection is performed mainly based on the center areas.

Performance of the proposed hard negative attention. Fig.8 (a) shows effectiveness of the proposed penalty term, $\exp(M_{nia}^n)M_{nia}$, in (7) for hard negative anchors. When $n=2$, our method produced the best densely-packed object detection results in terms of AP, which demonstrates that the proper penalty on hard negative anchors is necessary for accurate detection. Fig.8 (b) illustrates $\exp(M_{nia}^n)M_{nia}$ according to different values of IoU and n. If n is set to large values, we assign small values to anchors with relatively small IoU values. In other words, we strengthen the criteria for hard negative anchors so that only good anchors can be considered. Table 3 evaluates the performance of M_{nia}^n according to n in terms of all evaluation metrics.

Table 4. **Comparison using the SKU-110K dataset.** The red and blue colors denote the best and second best results, respectively.

Method	backbone	AP	$AP^{.50}$	$AP^{.75}$	AR_{300}	$AR_{300}^{0.50}$	$P^{R=0.5}$
Faster-RCNN [28]	resnet50	0.045	-	0.010	0.066	-	0
YOLO9000 [26]	darknet19	0.094	-	0.073	0.111	-	0
MaskRCNN [13]	resnet50	0.403	0.742	0.396	0.465	0.778	-
RetinaNet [19]	resnet50	0.455	-	0.389	0.530	-	0.544
Goldman *et al*. [12]	resnet50	0.492	-	0.556	0.554	-	0.834
Kant *et al*. [15]	resnet50	0.521	0.891	0.562	0.596	0.931	-
Ours	resnet50	0.522	0.897	0.556	0.601	0.935	0.816

Table 5. **Comparison using other densely-packed object detection benchmark datasets.**

Dataset	Method	AP	$AP^{.50}$	$AP^{.75}$	AR_{300}	$AR_{300}^{0.50}$
WebMarket [36]	Goldman *et al*. [12]	0.383	0.773	0.332	0.491	0.855
	Kant *et al*. [15]	0.403	0.813	0.340	0.551	0.954
	Ours	0.453	0.879	0.408	0.583	0.974
Holoselecta [8]	Goldman *et al*. [12]	0.454	0.835	0.447	0.581	0.955
	Kant *et al*. [15]	0.384	0.705	0.368	0.524	0.843
	Ours	0.431	0.859	0.353	0.574	0.992
CAPG-GP [9]	Goldman *et al*. [12]	0.431	0.684	0.519	0.481	0.721
	Kant *et al*. [15]	0.510	0.777	0.616	0.572	0.816
	Ours	0.510	0.822	0.597	0.648	0.971

Figure 9. **Qualitative detection results** of the proposed method on the SKU-110k dataset.

4.3. Comparison with State-of-the-art Methods

We used the SKU-110K dataset, which is a representative dataset for densely-packed object detection. We compared recent densely-packed object detection methods as well as state-of-the-art object detection methods.

Comparisons using the SKU-110K dataset. Table 4 shows densely-packed object detection results on the SKU-110K dataset and demonstrates that our method outperforms other state-of-the-art methods (*i.e*. Faster-RCNN [28], YOLO9000 [26], and MaskRCNN [13]) including recent densely-packed object detection algorithms (*i.e*. Goldman *et al*. [12] and Kant *et al*. [15]) in terms of evaluation metrics. In particular, our method considerably surpasses our baseline, RetinaNet [19]. Fig.9 shows qualitative detection results of the proposed method. Although there were a large number of objects and their sizes were very small, our method accurately detected them under real-world environments (*e.g*. marts, convenience stores, wholesale, and retail stores). Please note that for a fair comparison we used the same experimental settings (*i.e*. same image resolution and no data augmentation).

Comparisons using other densely-packed object detection datasets. Table 5 shows object detection results on three densely-packed object detection benchmark datasets. For this experiment, all compared methods including our method were trained using the SKU-110K dataset to follow the same experimental settings in [15] and test resolution is set to 800 $\times$ 800. As shown in Table 5, our method consistently outperforms other densely-packed object detection methods in terms of most evaluation metrics for various benchmark datasets.

5. Conclusion

In this paper, we proposed advanced weighted Hausdorff distance (AWHD) and hard negative-aware anchor (HNAA) attention for densely-packed object detection. Despite the high object density and small-size objects, our AWHD accurately estimated a center area map and the proposed HNAA attention accurately obtained the relative importance of each anchor to give a penalty to hard negative anchors. Experimental results demonstrate that our method using the AWHD and HNAA attention significantly outperforms other state-of-the-art detection methods in several densely-packed object detection datasets. The proposed AWHD and HNAA can be incorporated into existing detection networks to improve detection accuracy.

References

[1] Hedy Attouch, Roberto Lucchetti, and Roger J-B Wets. The topology of the ρ-hausdorff distance. *Annali Mat. Pura. Appl.*, 160(1):303–320, 1991.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv:2004.10934*, 2020.

[3] Yuanqiang Cai, Longyin Wen, Libo Zhang, Dawei Du, and Weiqiang Wang. Rethinking object detection in retail stores. In *AAAI*, 2021.

[4] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object detection via region-based fully convolutional networks. In *NIPS*, 2016.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] M-P Dubuisson and Anil K Jain. A modified hausdorff distance for object matching. In *ICPR*, 1994.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. PASCAL VOC2008. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[8] Klaus Fuchs, Tobias Grundmann, and Elgar Fleisch. Towards identification of packaged products via computer vision: Convolutional neural networks for object detection and image classification in retail environments. In *ICIT*, 2019.

[9] Weidong Geng, Feilin Han, Jiangke Lin, Liuyi Zhu, Jieming Bai, Suzhen Wang, Lin He, Qiang Xiao, and Zhangjiong Lai. Fine-grained grocery product recognition by one-shot learning. In *ACM Multimedia*, 2018.

[10] Ross Girshick. Fast R-CNN. In *ICCV*, 2015.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[12] Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. In *CVPR*, 2019.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*, 2017.

[15] Sonaal Kant. Learning gaussian maps for dense object detection. *arXiv:2004.11855*, 2020.

[16] Davood Karimi and Septimiu E Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on medical imaging*, 39(2):499–513, 2019.

[17] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *CVPR*, 2019.

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[21] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018.

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[23] Greg Nichols. Retail robots coming to these grocery stores. https://www.zdnet.com/article/retail-robots-coming-to-these-grocery-stores/, 2021.

[24] Apoorva Raghunandan, Pakala Raghav, HV Ravish Aradhya, et al. Object detection algorithms for video surveillance applications. In *ICCSP*, 2018.

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[26] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *CVPR*, 2017.

[27] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv:1804.02767*, 2018.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *NIPS*, 2015.

[29] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019.

[30] Javier Ribera, David Guera, Yuhao Chen, and Edward J Delp. Locating objects without bounding boxes. In *CVPR*, 2019.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[32] Mingxing Tan, Ruoming Pang, and V Le Quoc. EfficientDet: Scalable and efficient object detection. *arXiv:1911.09070*, 2019.

[33] Nishchal K Verma, Teena Sharma, Shreedharkumar D Rajurkar, and Al Salour. Object identification for inventory management using convolutional neural network. In *AIPR*, 2016.

[34] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-yolov4: Scaling cross stage partial network. In *CVPR*, 2021.

[35] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *CVPR*, 2018.

[36] Yuhang Zhang, Lei Wang, Richard Hartley, and Hongdong Li. Where's the weet-bix? In *ACCV*, 2007.

[37] Chunjiang Zhao, Wenkang Shi, and Yong Deng. A new hausdorff distance for image matching. *Pattern Recognition Letters*, 26(5):581–586, 2005.

[38] Jianan Zhao, Fengliang Qi, Guangyu Ren, and Lin Xu. PhD learning: Learning with pompeiu-hausdorff distances for video-based vehicle re-identification. In *CVPR*, 2021.