Spatial Self-Distillation for Object Detection with Inaccurate Bounding Boxes

Di Wu*, Pengfei Chen*, Xuehui Yu*, Guorong Li, Zhenjun Han†, Jianbin Jiao
University of Chinese Academy of Sciences

Abstract

Object detection via inaccurate bounding boxes supervision has boosted a broad interest due to the expensive high-quality annotation data or the occasional inevitability of low annotation quality (e.g., tiny objects). The previous works usually utilize multiple instance learning (MIL), which highly depends on category information, to select and refine a low-quality box. Those methods suffer from object drift, group prediction and part domination problems without exploring spatial information. In this paper, we heuristically propose a Spatial Self-Distillation based Object Detector (SSD-Det) to mine spatial information to refine the inaccurate box in a self-distillation fashion. SSD-Det utilizes a Spatial Position Self-Distillation (SPSD) module to exploit spatial information and an interactive structure to combine spatial information and category information, thus constructing a high-quality proposal bag. To further improve the selection procedure, a Spatial Identity Self-Distillation (SISD) module is introduced in SSD-Det to obtain spatial confidence to help select the best proposals. Experiments on MS-COCO and VOC datasets with noisy box annotation verify our method’s effectiveness and achieve state-of-the-art performance. The code is available at https://github.com/ucas-vg/PointTinyBenchmark/tree/SSD-Det.

1. Introduction

Object detection [18, 38, 58, 30, 57] relying on large-scale datasets like MS-COCO[31] has significantly progressed and achieved good performance. However, accurate bounding box annotations are expensive and challenging in natural contexts [12]. Especially in many professional scenarios, it is difficult to label accurate annotations without domain knowledge (e.g., agricultural crop observation and medical image processing) [12, 1]. As shown in Fig. 1a in some complex datasets, the human annotators may also annotate inaccurate bounding boxes due to the inherent ambiguities[20] of objects. In addition, labelling with a detector or weak signal[11] (e.g., point) is much cheaper but brings more inaccuracy. Therefore learning robust detectors with inaccurate bounding boxes[12, 4, 35, 53, 25] is a practical and meaningful task and has boosted a broad interest.

To use the inaccurate annotations, most related methods [12, 11] refine the inaccurate annotations as Fig. 1a shows, and then train a detector head or re-train a detector with the refined box as the new supervision. There are two main steps during refining: 1) **Bag Construction:** For each object, obtaining some proposals around the inaccurate an-

---

* Equal contribution.
† Corresponding authors. (hanzhji@ucas.ac.cn)
notated bounding box to form the object-level proposal bag; 2) **Proposal Selection:** Selecting the top-\(k\) proposals with the highest classification confidence from each bag and then weighting average them to obtain the refined box.

During the proposal selection, they usually utilize multiple instance learning (MIL) [10] supervised by category information to choose the proposals with high classification confidence from the constructed bags. However, they pay less [12] or no [11] attention to mining spatial information, leading to the following problems as shown in Fig. 1b: (1) **Object Drift:** For each object, some proposals in the constructed bag do not have a high IoU with the original object but with another nearby object. These proposals are not spatially adjacent to the original object but still have high classification confidence, as the rightmost (the yellow box) proposal of \(O_2\) shows in the left-bottom corner of Fig. 1b. Only the category confidence is relied on for selecting proposals for \(O_2\), and the rightmost proposal will be selected as the refined box. It means the refined box drifts to another object (Fig. 1b (i)), reducing the recall; (2) **Group Prediction:** Most works [7, 11, 56] select the top-\(k\) proposals by classification confidence and then weight average them as the refined box, causing the group prediction problem, as shown in Fig. 1b (ii); (3) **Part Domination:** The detector often focuses on the object’s semantic region, which can statistically represent the category (e.g. the face). As shown in Fig. 1b (iii), the high classification confidence of the animal is in the discriminant part (the face) rather than the entire object as mentioned by [44, 39].

To address these problems, we propose a **Spatial Self-Distillation** based detector (SSD-Det) to integrate the spatial cues into the bounding box refinement. SSD-Det has two important components: the Spatial Position Self-Distillation (SPSD) module for the bag construction step and the Spatial Identity Self-Distillation (SISD) module for the proposal selection step. To construct high-quality proposal bags, SPSD utilizes a neighborhood sampler to generate a balanced and flexible initial proposal bag for each object and then trains a regressor with the supervision of the annotated inaccurate bounding boxes. Finally, high-quality proposal bags are constructed with proposals corrected by the regressor. The mechanism behind SPSD is that the network learns the spatial information from the reliable samples, e.g. those low-noise annotations, in the dataset and then guides the noisy samples to produce high-quality proposals, as shown in Fig. 2. In addition, to further combine the category information and the spatial information, an interactive structure is implemented by alternately using SPSD to mine spatial cues and MIL to utilize the category information. With SPSD and the interactive structure, a high-quality proposal bag can be constructed. Experiments on MS-COCO show that SPSD can significantly improve the mean/max IoU between objects and proposals (about 18/10 points, Fig. 5) in the constructed bag. Instead of selecting proposals by classification confidence, we have proposed the SISD module in the proposal selection step. We use it to obtain each proposal’s spatial confidence by predicting the IoU with the object and combining the IoU with classification confidence to select the top-\(k\) proposals. It is worth mentioning that SISD is an object-related IoU predictor, which means that the predicted IoU may be different for the same proposal that appears in different objects’ bags. Accordingly, it guarantees that SISD can better handle object drift and group prediction problems. Experiments on MS-COCO and VOC datasets verify the effectiveness of our method and bring state-of-the-art performance. The contributions are as follows:

1) We further investigate the inaccurate-box supervised object detection tasks and propose an end-to-end training SSD-Det that combines the spatial and the category information in an interactive fashion.

2) We utilize an SPSD module to generate higher-quality proposals sampling through statistic-guide spatial position distillation, raising the upper bound of the refinement.

3) To add spatial cues to classification confidence, we also introduce an SISD module to select a proposal belonging to the object rather than the category.

4) The performance of our proposed SSD-Det improves the mean average precision (AP) of the best previous method (e.g. over 10 AP on 40% noisy MS-COCO) and achieves state-of-the-art under various noise rate box supervision on MS-COCO and VOC datasets.

2. Related Work

2.1. Object Detection

Classic object detection [15, 38, 37, 32, 30, 3, 43, 57] is supervised by an accurate bounding-box. One-stage detectors utilize anchors as the sliding-window, such as YOLO [37], SSD [32], and RetinaNet [30]. Two-stage detectors mine spatial information to predict proposals (e.g. selective search [47] in Fast R-CNN [15] or RPN in Faster R-CNN [38]) and conduct classification and bounding-box
regression with filtered proposals sparsely. Transformer-based (i.e., DETR [3], Deformable-DETR [62], and SwinTransformer [33]) detectors utilize global information for better representation. Sparse R-CNN [43] combines a transformer’s advantages and CNNs for detection.

2.2. Weakly-Supervised Object Detection (WSOD)

WSOD trains object detectors with image tag supervision. Only with the category annotation, the majority of previous methods treat each image as a bag and candidate proposals as instances. They follow the multiple instance learning (MIL) pipeline [1, 44, 46, 7, 48], which highly depends on category information. However, the MIL loss function leads to a non-convex optimization problem; thus, MIL solutions are usually stuck into the local minima. Context information [24, 51], spatial regularization [1, 9, 48], and optimization strategy [44, 48, 46] are proposed to address the problems. SPE [28] introduces Transformer into WSOD and uses attention to generate proposals. SD-LocNet [59] tackles the initialized noisy object locations in WSOD and proposes a self-directed localization network to identify the noisy object instances. [44, 46, 7] use the pseudo label for classification’s iterative refinement. However, we use the pseudo box as a better self-distillation teacher. [54, 39] conduct regression to move the proposals, whereas we conduct regression to distill for better box construction. In this work, we also formulate box correction as a MIL problem.

2.3. Semi-Supervised Object Detection (SSOD)

Semi-supervised learning in object detection can be roughly categorized into two groups: consistency based [22, 45] and pseudo label based [27, 36, 41, 49, 64, 52]. [52] presents an end-to-end SSOD approach with two simple techniques named soft teacher and box jittering to facilitate the efficient leverage of the teacher model. Both Soft Teacher [52] and SSD-Det obtain pseudo-labels from candidate boxes, adopting distillation, box jitter and classification scores weighting policy. However, they are different: 1) SSOD selects candidate boxes from the teacher model’s detection results, while SSD-Det generates them with a generative approach, due to no any accurate supervised data for ensuring high-quality boxes in the detection results. 2) [52] is based on FixMatch [40] and requires maintaining two networks for teacher-student distillation structure. In contrast, our approach only needs a multi-head detector with a shared backbone for self-distillation. 3) Box Jitter: [52] calculates box variance from jittering for result selection, while we aim to generate candidate boxes that combine SISD-predicted IoU and classification scores, selecting boxes closer to the ground truth. 4) We use classification scores as weights for the next stage’s loss, while [52] employs them as a criterion for selecting reliable samples.

2.4. Learning with Noisy Annotations

Training CNNs under noisy labels has been an active research area. Previous research focuses on the classification task, and develops various techniques to deal with noisy labels, such as sample selection [17, 23] for training, label correction [42, 34], and robust loss functions [60, 14] against noisy labels. Recently, many efforts [4, 25, 35, 53, 11, 12] have been devoted to the object detection task. On the one hand, Simon et al. [4] first investigates the impact of different types of label noise on object detection. They propose a per-object co-teaching method to alleviate the effect of noisy labels. On the other hand, [53] proposes a meta-learning framework for noisy annotations consisting of noisy category labels and bounding boxes. [11, 12] utilize object-level MIL to refine the inaccurate box. OA-MIL [12] constructs proposal bags through label assignment in a discriminant style. P2BNet [11] originally conducts point-supervised object detection tasks. However, it can be seen as the box correction in its refinement stage. It uses hand-craft anchors to generate proposal bags. Our method inherits the generative style of P2BNet and conducts spatial distillation to mine spatial information.

2.5. Knowledge Distillation

Knowledge distillation (KD) [21] aims to learn compact and efficient student models guided by excellent teacher networks. It is first applied to object detection in [5], in which hint learning and KD are both used for multi-class object detection. Recently, many efforts [26, 50, 8, 16, 55] aim to mimic the feature. [61] shows that localization knowledge is more important and proposed a localization distillation method. We also transfer the spatial knowledge from reliable labeled instances to correct inaccurate bounding boxes (shown in Fig. 2) in a self-distillation manner.

3. Methodology

This work aims to learn a robust detector with inaccurate bounding boxes. Instead of training a detector with the original inaccurate bounding box, we follow most related works [11, 12] that design a branch to refine the inaccurate bounding box and then train the detector head or detector with the refined bounding box. The most important part is how to design a refining policy. We first design a two-stage basic box refiner (gray region in Fig. 3) as a naive solution that modified from [11]. Then, SPSD and SISD are proposed and added to further mine the spatial cues for box refinement, yielding SSD-Det. Therefore, the overall loss function is formulated as:

\[ L = L_{\text{Basic}} + \alpha_1 \cdot L_{\text{SPSD}} + \alpha_2 \cdot L_{\text{SISD}} + \alpha_3 \cdot L_{\text{Det}} \]  

where \( \alpha_1, \alpha_2 \) and \( \alpha_3 \) are set as 0.25, 0.25 and 4 respectively. \( L_{\text{Det}} \) denotes the loss of detector or detection head. During inference, only the detector or the detection head is used.
Figure 3: The framework of SSD-Det. It contains basic box refiner, SPSD module, SISD module and a detector head. Neighborhood sampler is adopted to extract with a backbone network. And then through RoIAlign [18] and two fully connected (fc) layers, features are extracted and denote as \( F \). Then the refine annotation is used to train a detection head or detector. Following [11], we design a two-stream structure as a MIL classifier to select the best proposal for box refinement. Giving an image with inaccurate box annotation, for each object, \( B \) is a bag of proposals (bounding boxes) that are generated around its inaccurate annotation by a sampler policy (e.g., selective search [47], edge box [63], neighborhood sampler in Sec. 3.2). Meanwhile a feature map is extracted with a backbone network. Then the refine annotation is used to train a basic box refiner for MIL training. The selected proposals are average weighted as the refined box and the instance score. The bag score \( S \) is obtained by the summating of the \( P \) proposal boxes’ proposal scores.

\[
O_{cls} = f_{cls}(F) \in \mathbb{R}^{P \times K}; [S_{cls}]_{pk} = e^{[O_{cls}]_{pk}} / \sum_{k=1}^{K} e^{[O_{cls}]_{pk}},
\]

\[
O_{ins} = f_{ins}(F) \in \mathbb{R}^{P \times K}; [S_{ins}]_{pk} = e^{[O_{ins}]_{pk}} / \sum_{p=1}^{P} e^{[O_{ins}]_{pk}},
\]

\[
S = S_{cls} \odot S_{ins} \in \mathbb{R}^{P \times K}; \hat{S} = \sum_{p=1}^{P} [S]_{p} \in \mathbb{R}^{K}.
\]

where \([\cdot]_{pk}\) is the value at row \( p \) and column \( k \) in the matrix.

Where the basic box refiner has two similar stages. The loss of stage I (termed \( L_I \)) adopt the MIL paradigm with the form of cross-entropy (CE) loss, defined as:

\[
L_I = CE(\hat{S}, e) = -\sum_{k=1}^{K} c_k \log(\hat{S}_k) + (1 - c_k) \log(1 - \hat{S}_k)
\]

where \( e \in \{0, 1\}^K \) is the one-hot category label. And each object’s proposals with the top-\( k \) highest proposal score \( S \) are weighted to obtain the refined box.

The stage II takes the refined box of stage I as input and performs fine refining with a similar structure as stage I. Differently, the focal loss is adopted in stage II instead of cross entropy loss. In order to cooperate with focal loss, the classification branch uses the sigmoid \( \sigma(x) \) instead of softmax function and we sample some negative samples \( N \) to further suppress the background. With the bag score \( \hat{S} \) and the negative sample scores \( S_{neg} \), the loss is:

\[
L_{II} = \left< c^T, \hat{S}^* \right> \cdot FL(\hat{S}, e) + \sum_{N} \beta \cdot FL(S_{neg}^s, c_{neg})
\]

where FL is the focal loss [30], \( \hat{S}_j \) represents the bag score predicted by stage I. \( \left< c^T, \hat{S}_j \right> \) represents the inner product...
of the two vectors, meaning the predicted bag score of the ground-truth category. $\beta$ is the average of $\left(c_j^T, S_j^T \right)$. They are used to weight each object’s FL for stable training. The overall loss function of the basic refiner here is:

$$L_{\text{basic}} = L_1 + \alpha_{\text{II}} \cdot L_{\text{II}}$$

(5)

where $\alpha_{\text{II}}$ are the weightloss of the two stages.

During training, the refined box of stage II is used as supervision for a detection head or detector. After training, the basic box refiner will be removed, leaving a well-trained detection head or detector. In this way, we can train a detector under inaccurate annotations.

### 3.2. Spatial Position Self-Distillation (SPSD)

Like most MIL paradigm methods, basic box refiner has two main components: bag construction and proposal selection. And the main idea is to use classification information to guide the refining. In this paper, we add spatial information to improve refining. Specifically, SPSD is proposed to use spatial information to enhance bag construction.

Bag construction aims to obtain proposals for each object, while proposal selection is to select the proposals from the object bag. Then, the refined box is averaged over the selected proposals. Therefore, the quality of the proposals in constructed bag determines the upbound of refining. The bag construction can be implemented in a variety of ways. In this paper, the basic box refiner adopts a naive neighborhood sampler for bag construction. basic box refiner adopts a naive neighborhood sampler for bag construction.

**Neighborhood Sampler.** Proposals around the inaccurate box are sampled to construct an object bag. For each inaccurate box $b^* = (b^*_x, b^*_y, b^*_w, b^*_h)$, its scale and aspect ratio with $s$ and $v$ are adjusted and its positions $o_x, o_y$ are jittered to obtain the diverse proposal $b = (b_x, b_y, b_w, b_h)$:

$$b_w = v \cdot s \cdot b_w, \quad b_h = 1/v \cdot s \cdot b_h, \quad b_x = b^*_x + b_w \cdot o_x, \quad b_y = b^*_y + b_h \cdot o_y$$

(6)

These proposals $b$ are used to construct the positive proposal bag $B$ to train the MIL classifier. Thanks to the hand-craft sampling way, the number of proposals in different objects’ proposal bags is controllable and balanced. However, the hand-craft neighborhood sampler strategy is difficult to set hyper-parameters, and the sampling space is discrete. For example, when the jitter region is small, the optimization space of refining is limited, while when it is large, more background will be introduced. Hence, we propose the SPSD module to mine spatial information for high-quality proposal bag construction.

**Statistically Guided Adaptive Sampling.** Instead of simply using a neighborhood sampler, we adopt a statistically guided adaptive sampling by adding SPSD modules into the basic box refiner. Taking the constructed proposal bag $B$ of hand-craft neighborhood sampler as input, the RoI features of proposals in $B$ are extracted and fed into the two shared fc layers to obtain $F$. Then a regression fc layer $f_{\text{dis}}$, supervised by the inaccurate annotated bounding box $b^*$, is introduced to predict the adaptive proposal bag $B^{\text{dis}} = f_{\text{dis}}(F) \in \mathbb{R}^{P \times 4}$, in which the proposals are closer to the object. Later, $B^{\text{dis}}$ as the constructed proposal bag is fed into stage I of basic box refiner. In order to combine category and spatial information, we implement an interactive structure by alternately using SPSD to mine the spatial cues and using MIL in basic box refiner to utilize the category information. Specifically, the refined bounding box $b^*$ of stage I that selected by the classification confidence is used to supervision of a new SPSD module for stage II. Similar as stage I, The new SPSD takes proposal bag $B^{\text{dis}}$ of hand-craft neighborhood sampler as input. Through the RoI align and the two shared fc layers, the feature $F$ is extracted. An extra fc layer $f_{\text{dis}}$ is then utilized to conduct further regression. Different with stage I, the obtained $B^{\text{dis}}$ is supervised by the refined $b^*$. The loss function of the spatial distillation for adaptive sampling can be defined as $L_{\text{SPSD}}$ in Eq. 7.

$$L_{\text{SPSD}} = \frac{1}{P} \left\{ \sum_{p=1}^{P} L_4([B^{\text{dis}}]_p, b^*) + \sum_{p=1}^{P} L_1([B^{\text{dis}}]_p, \hat{b}^*) \right\}$$

(7)

where the $L_1$ is the L1 loss function for loose restrictions.

The idea behind SPSD is that the dataset with inaccurate annotation still has many reliable, high-quality boxes and inaccurate boxes. Supervised by the high-quality boxes statistically, the network can guide those proposals sampled around the inaccurate bounding box to regress to the ground truth. With the self-distillation mechanism, SPSD learns the semantic-spatial correspondence knowledge from the reliable samples in the dataset and then propagates the knowledge to produce high-quality proposals.

**Adaptive Negative Sampling.** Negative samples are introduced in Stage II to better suppress the background. With the sampled $B^{\text{dis}}$, we can adaptively sample the negative samples with a small IoU (set smaller than 0.3 by default) with all positive proposals in all bags, to compose the negative sample set $N$ for stage II.

### 3.3. Spatial Identity Self-Distillation (SISD)

The basic box refiner selects the proposals only depending on classification confidence during proposal selection. To select the proposal which has high classification confidence and is also spatially close to the object from the bag, we propose a SISD module to predict the IoU between proposals and their corresponding object. Afterwards, through the combination between the IoU and the classification confidence, top-$k$ proposals are selected. In SISD, we design an Object Relevance Enhancement (ORE) module to distinguish different objects’ features with the same RoI region. And an identity predictor is designed to predict each proposal’s IoU with the object.
Object Relevance Enhancement (ORE). ORE enhances object-relevant features, making SISD an object-relevant IoU predictor. ORE allows the predicted IoU to be different for the same proposal in other objects’ bags. In addition, we integrate the feature of the bag’s corresponding object into the proposal feature, making the feature of different bags’ proposals distinct. That is the so-called ORE. For a proposal bag $B$, the feature $F$ is obtained through the RoI align and two fc layers. It is worth mentioning that the two fc layers do not share the parameters with those in the refiner since the optimization goals are contradictory. To represent the feature of the bag’s corresponding object, $F^+ \in \mathbb{R}^{1 \times D}$ is calculated by averaging features of $P$ proposals in proposal bag $B$. The object feature $F^+$ is broadcast into $\mathbb{R}^{P \times D}$, and then added to the proposal features to obtain the object-relevant features $F^* = F + F^+$.

**Spatial Identity Prediction.** By a following identity fc layer, $U \in \mathbb{R}^{P \times 1}$ is predicted. The pseudo label $T \in \{0, 1\}$ is IoU between proposals in $B$ and the merged box $b^*$ of stage I. For better optimization, the linear normalized $T' = (T - 0.5) / 0.5 \in (-1, 1)$ is utilized as supervision. The object function of the identity predictor is identified in:

$$L_{SISD} = \text{smooth}_{L1}(U, T')$$

where the $\text{smooth}_{L1}$ represents the smooth L1 loss. The predicted spatial confidence $U'$ is obtained by normalizing the $U$. Finally, $S^* = U' \cdot S$ is used to select the top-$k$ proposals for merging as the refined boxes.

### 4. Experiment

#### 4.1. Experimental Settings

**Datasets and Evaluation Metrics.** For experimental comparisons, two publicly available datasets are used for object detection with inaccurate bounding boxes: MS-COCO [31] and PASCAL VOC 2007 [13]. **MS-COCO** (2017 version) has 118k training and 5k validation images with 80 common object categories. **PASCAL VOC 2007** is one of the most popular benchmarks in generic object detection with 20 classes.

**Evaluation Metric.** We use mean average precision (mAP).
Table 3: Modules ablation of SPSD, SISD and Re-Train on MS-COCO validation set (without) and test set (with test). The Re-Train means we generate the pseudo label by SSD-Det and re-train a Faster R-CNN detector.

<table>
<thead>
<tr>
<th>Methods (w/o SISD)</th>
<th>AP</th>
<th>AP50</th>
<th>AP75</th>
<th>APs</th>
<th>APm</th>
<th>AP1s</th>
<th>AP1m</th>
<th>AP1l</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighborhood Sampler</td>
<td>24.6</td>
<td>52.0</td>
<td>20.1</td>
<td>14.3</td>
<td>28.2</td>
<td>31.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPSD (II) w/ weighted</td>
<td>26.0</td>
<td>53.3</td>
<td>22.5</td>
<td>15.6</td>
<td>29.4</td>
<td>33.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPSD (II) w/ weighted</td>
<td>26.3</td>
<td>53.4</td>
<td>22.5</td>
<td>15.6</td>
<td>29.3</td>
<td>33.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPSD (I+II) w/ weighted</td>
<td>27.2</td>
<td>53.7</td>
<td>24.7</td>
<td>15.9</td>
<td>30.3</td>
<td>35.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Different setting of SPSD.

Table 5: Different ORE strategies of SISD.

Table 6: Number of SPSD module.

mAP@[.5,.95] and (mAP@.5) for MS-COCO and VOC. The {AP, AP50, AP75, APs, APm, AP1s, AP1m, AP1l} is reported for MS-COCO and AP50 for VOC.

4.2. Comparison with State-of-the-Art

We compare our method with several state-of-the-art approaches [20, 17, 59, 58, 12] on MS-COCO and VOC 2007 datasets. We denote Clean-FasterRCNN and Noisy-FasterRCNN as FasterRCNN models trained under clean (original annotations) and noisy annotations with the default setting, respectively.

MS-COCO Dataset. Table 1 shows the comparison results on the MS-COCO. Inaccurate bounding box annotations significantly deteriorate the vanilla Faster R-CNN’s detection performance. Co-teaching and SD-LocNet only slightly improve the detection performance, especially under 40% box noise. That indicates that small-loss sample selection and sample weight assignment can not tackle noisy box annotations well. KL Loss slightly improves the performance under 20% and 40% box noise. By treating an object as a bag of instances, OA-MIL is somehow robust to noisy bounding boxes and performs better than other methods. Nevertheless, the previously-mentioned label assignment bag construction limits its ability to handle heavy noise. Our approach is more robust to noisy bounding boxes. It outperforms other methods by a large margin under high box noise levels and significantly boosts the detection performance across all metrics. For example, under 40% box noise, the end-to-end SSD-Det achieve 27.6 AP and 53.9 AP50, 9.0 and attains 11.3 point improvement compared with state-of-the-art method OA-MIL, respectively. Also, through re-training on FasterRCNN, the performance further reaches 29.3 AP and 54.8 AP50. With the backbone of ResNet-101, the performance achieves consistent improvement. On MS-COCO test set, our method also achieves state-of-the-art performance.

VOC 2007 Dataset. Table 2 shows the comparison results on the VOC 2007 test set. Co-teaching, SD-LocNet and KL Loss, can not address inaccurate bounding box annotations well. OA-MIL improves the performance on different noisy datasets. Our approach obtains further improvements to 77.10, 74.80, 71.50, 66.90 AP50 on 10%, 20%, 30 % and 40 % noisy box datasets, respectively.
Figure 4: Qualitative detection results on COCO validation. Previous methods miss objects and face part prediction problems. Our method misses fewer objects, and the bounding box quality is better, especially for small or overlapped objects.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Box Refiner+Re-Train</th>
<th>SSD-Det</th>
<th>SSD-Det+Re-Train</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AP/AP₅₀</td>
<td>AP/AP₅₀</td>
<td>AP/AP₅₀</td>
</tr>
<tr>
<td></td>
<td>29.0 / 54.4</td>
<td>27.6 / 53.9</td>
<td>29.3 / 54.8</td>
</tr>
</tbody>
</table>

Table 7: Comparisons of end-to-end and Re-Train.

4.3. Ablation Study and Analysis

To further analyze SSD-Det’s effectiveness and robustness, we conduct more experiments on COCO val set if there are no other instructions. Except for Table 3, the noise level of these experiments is 40%.

**Ablation of Modules.** Ablation study of each component in our approach is given in Table 3, including: (i) Different stages of our basic box refiner, i.e. training object detector without the stage II (2-Ref), where the pseudo boxes predicted by the stage I are served as the supervision for training a parallel detector. (ii) SPSD, i.e. training without SPSD, where the object-bag is constructed directly by neighborhood sampling around the noisy ground-truth or the predicted pseudo boxes of the stage I. (iii) SISD. (iv) Re-Train with FasterRCNN (Re-Train).

**Effectiveness of SPSD.** SPSD further improves the detection performance on the MS-COCO, especially under high box noise levels, e.g. under 40% box noise level, SPSD boosts the performance from 24.6 to 27.2, as shown in Table 3 (row 3). In Table 4, we conduct further ablation on SPSD. With SPSD bag construction only in stage II, the performance increases by 1.4 AP. The performance further improves with the proposal score of stage I as weights. With SPSD in all stages, the AP reaches 27.2. Fig. 5 shows the bag quality. With SPSD, the mean IoU increases from 40.3 to 58.7 and the max and top-10 IoU increase to 78.3 and 75.1, which indicates a better upper bound of proposal selection. More high-quality proposals bring better optimization and easier proposals selection.

**Number of SPSD.** As shown in Table 6. When adding 3 SPSD, performance drops slightly, probably due to the accumulation of errors outweighing the performance gain from extra stages. Hence, 2 SPSD is our default setting.

**Effectiveness of SISD.** SISD is designed to select object-aware proposals in box selection. Under 40% and 20% box noise, the detection performance improves from 27.2 to 27.6 and 33.0 to 33.6, which verifies the effectiveness of the module, shown in Table 3. We also study the
Table 9: Breakdown of different problems during refinement (COCO under 40% noise level). s, m and l mean small, middle and large scale.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>our</th>
<th>Quality (Average IoU)</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>Part</td>
</tr>
<tr>
<td>Objects-F</td>
<td>✓</td>
<td>43.3</td>
<td>24.1</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>47.0</td>
<td>30.8</td>
</tr>
<tr>
<td>COCO-F</td>
<td>✓</td>
<td>48.1</td>
<td>25.4</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>48.2</td>
<td>32.1</td>
</tr>
<tr>
<td>Point-based Annotator</td>
<td>✓</td>
<td>55.6</td>
<td>30.4</td>
</tr>
<tr>
<td>COCO-P</td>
<td>✓</td>
<td>65.2</td>
<td>46.7</td>
</tr>
</tbody>
</table>

Table 10: Analysis of noisy annotations types and quality.

<table>
<thead>
<tr>
<th>Method</th>
<th>Drift Rate (%)</th>
<th>Group Rate (%)</th>
<th>Part Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>1.5</td>
<td>1.0</td>
<td>1.3</td>
</tr>
<tr>
<td>OA-MIL[12]</td>
<td>15.1</td>
<td>17.8</td>
<td>17.1</td>
</tr>
</tbody>
</table>

\[\text{IoU} = \frac{A(I)}{A(D + G - A(I))^2}, \text{IoG} = \frac{A(I)}{A(G)}, \text{IoD} = \frac{A(I)}{A(D)} \] (9)

where \(A(*)\) is area of box *, D and G are refined box and gt box respectively, and I is insertion between D and G.

We statistically count the proportion of three noise types of 'bad' refined boxes (having small IoU with gt) in Table 9:

(i) Drift: 'bad' refined box has a higher IoU with another nearby object. (ii) Group: 'bad' refined box has high IoG with multiple objects. (iii) Part: 'bad' refined box has a high IoD. Table 9 shows quantitative results for each noise type of baseline and ours. The drift, group, part problems reduce from 15.1%, 6.7%, 2.8% to 1.5%, 1.7%, 1.0%, respectively, demonstrating our improvement.

Experiments on Real-life Noisy Annotations. Real-life noisy annotations stem from: low-quality data (e.g., occlusion, blur), human annotator errors and automatic machine annotator limitations. Noise from human errors is quite subjective, since differences between annotators. For a more objective analysis, noisy annotations from machine annotator is used for experiments. Without loss of generality, Faster R-CNN well-trained on MS-COCO was applied to Objects365 images, yielding Objects-F dataset, and to COCO-val images, producing COCO-F dataset. P2BNet [11], a point-based annotator, was used on COCO-val images with point annotations, generating COCO-P dataset. SSD-Det effectively improves low-quality boxes. As shown in Table 10, with SSD-Det’s refinement, the average IoU increases for Objects-F (from 44.3 to 47.0), COCO-F (from 45.1 to 48.2) and COCO-P (from 55.6 to 65.2) datasets. Further, the proportion of reliable annotations increases, and noise categories’ frequency (Part, Oversize, and Shift) decreases for all datasets.

5. Conclusion

This paper investigates problems during refinement caused by solely using category information to select proposals. We also propose SSD-Det to mine spatial information in a self-distillation fashion. SSD-Det introduces the SPSD module to learn semantic-spatial correspondence knowledge with neighborhood sampler and an interactive structure to combine spatial information and category information, thus producing a high-quality proposal bag. SISD in SSD-Det is utilized to improve the proposal selection procedure by integrating object-relevant spatial confidence. Complete ablations on multiple datasets verify the effectiveness of SSD-Det.

References


[12] Liu et al. Robust object detection with inaccurate bounding boxes. In *ECCV*, 2022. 1, 2, 3, 4, 6, 7, 9


[17] Bo Han, Quanming Yao, and Xingrui Yu et al. Distilling object detection with sparse proposal evolution. In *ECCV*, 2022. 3


[38] Shaoqing Ren, Kaiming He, and Ross B. Girshick et al. Faster R-CNN: towards real-time object detection with region proposal networks. *TPAMI*, 2017. 1, 2, 6, 7


[44] Peng Tang and Xinggang Wang et al. Multiple instance detection network with online instance classifier refinement. In CVPR, 2017. 2, 3
[48] Fang Wan, Pengxu Wei, and Zhenjun Han et al. Min-entropy latent model for weakly supervised object detection. TPAMI, 2019. 3
[53] Youjiang Xu, Linchao Zhu, and Yi Yang et al. Training robust object detectors from noisy category labels and imprecise bounding boxes. TIP, 2021. 1, 3
[54] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In ICCV, 2019. 3
[56] Xuehui Yu, Pengfei Chen, and Di Wu et al. Object localization under single coarse point supervision. In CVPR, 2022. 2
[57] Xuehui Yu, Yuqi Gong, and Nan Jiang et al. Scale match for tiny person detection. In WACV, 2020. 1, 2
[59] Xiaopeng Zhang, Yang Yang, and Jiashi Feng. Learning to localize objects with noisy labeled instances. In AAAI, 2019. 3, 6, 7