Abstract
We introduce a box-level active detection framework that controls a box-based budget per cycle, prioritizes informative targets and avoids redundancy for fair comparison and efficient application.
Under the proposed box-level setting, we devise a novel pipeline, namely Complementary Pseudo Active Strategy (ComPAS). It exploits both human annotations and the model intelligence in a complementary fashion: an efficient input-end committee queries labels for informative objects only; meantime well-learned targets are identified by the model and compensated with pseudo-labels.

1. Introduction

Our contributions can be summarized as follows:
• We propose a box-level active detection framework, where we control box-based budgets for realistic and fair evaluation, and concentrate annotation resources on the most informative targets to avoid redundancy.
• We develop ComPAS, a novel method that seamlessly integrates model intelligence into human efforts via an input-end committee for challenging target annotation and pseudo-labeling for well-learned counterparts.
• We provide a unified codebase with implementations of active detection baselines and SOTAs, under which the superiority of ComPAS is demonstrated via extensive experiments.

2. Related Work

Active scoring functions.
Our localization informativeness is efficiently estimated between stochastic perturbations of candidates without dependency on model architecture.

Multi-model score ensemble.
Our input-based committee promotes diversity via stronger positional and color perturbations applied on more input members, and disagreement is efficiently analyzed between a reference and members.

Implementation and evaluation.
We suggest box level annotation, which is attempted but neither well explored [31] nor applicable to the in-domain task [29].
To this end, we present a strong pipeline that integrates both human annotations and machine predictions on the box-level.
To help advance reproducible research, we introduce a shared implementation of methods based on the same detector, train with similar procedures in a unified codebase, evaluate under the box-level criterion and support both labeled-only and mixed-supervision learning.

3. ComPAS for Box-level Active Detection

3.1. Problem Formulation

We actively select top-ranked informative bounding box proposals for annotators to identify the objects of interest within the candidate regions.

In the subsequent active acquisition cycles, both sparsely labeled and unlabeled images are evaluated, among which top-ranked boxes with low overlap with existing ground truths are prompted for labels.

As illustrated in Fig. 2, under the box-level active detection scenario, we propose a Complementary Pseudo Active Strategy (ComPAS), where the synergy between hard ground truth mining during active sampling (Sec. 3.2) and easy pseudo-label generation (Sec. 3.3) is exploited.

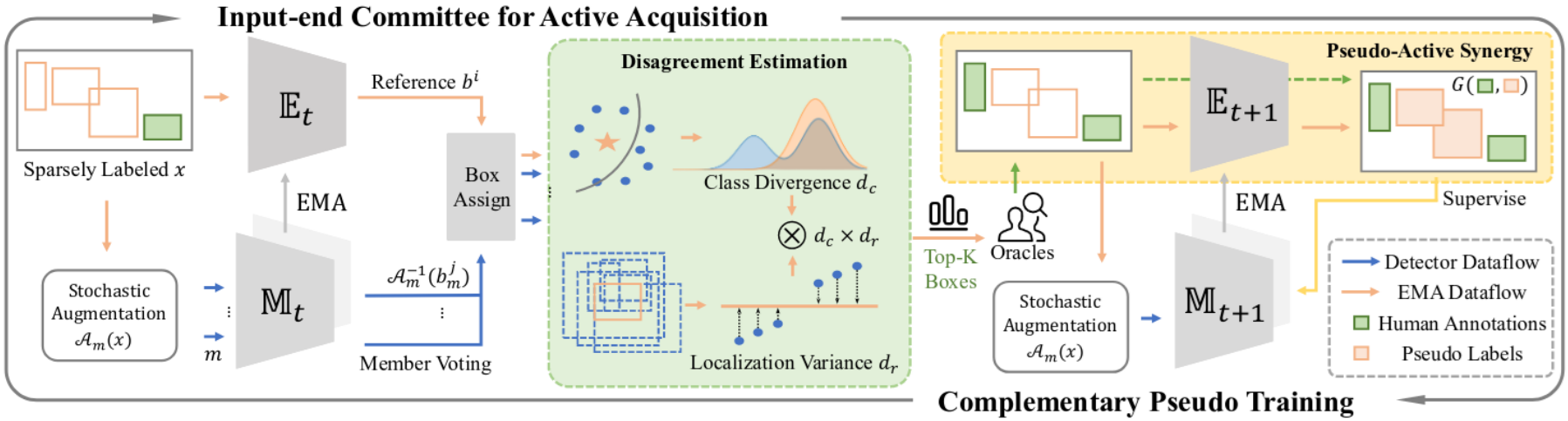

Figure 2. Overview of our ComPAS pipeline for box-level active detection, which iterates between active acquisition via the input-end committee and complementary training based on pseudo-active synergy. Only workflow of sparsely labeled images is shown for generality.

3.2. Active Acquisition via Input-end Committee

We propose to introduce invariant transformations on the input-end.
The posterior disagreement is thus estimated amongst multiple stochastic views of the input, which can be considered as committee members.
Based on the chairman predictions as a reference, measuring disagreement between it and all other member hypotheses can effectively reduce the assignment complexity.

$$\theta'_{tr} = \alpha\theta'_{tr-1} + (1-\alpha)\theta_{tr}, \qquad (1)$$

Disagreement on classification.

$$d_c^{ij} = -\mathbb{E}_{\mathbf{q}^i}[\log \mathbf{q}^j]. \qquad (2)$$

disagreement about box bi is aggregated among M committee members :

$$d_c^i = \frac{1}{M}\sum_m^M \left(\frac{1}{k_{mi}}\sum_j^{k_{mi}} d_c^{ij}\right), \qquad (3)$$

A larger value indicates higher disagreement amongst the input-end committee over a box candidate.
It shows that the current model cannot consistently make invariant label predictions under varying degrees of image perturbations, and thus it should be queried for human annotations.

Disagreement on localization.
disagreement over the location of bi is measured based on the chairman re-calibrated boxes:

$$d_r^i = \frac{1}{4}\sum_k^4 \hat{\sigma}_k(\{\mathbb{E}^{reg}(\mathcal{A}_m^{-1}(b_m^j))\}). \qquad (4)$$

Overall, for the box-level detection task, our scoring function is formulated as follows:

$$d^i = d_c^i \times d_r^i, \qquad (5)$$

3.3. Sparse- and Mixed-Supervision Training
Despite the significant label absence, as described in Sec. 3.2, the silver lining is that human annotations have been provided for targets that the previous detector fails to interpret, leaving the easier ones to be concerned about. We find the pseudo-label generation complementary to it, where targets with confident model predictions are kept for selftraining, while challenging targets with uncertain predictions are filtered out. With both active sparse training and pseudo-label generation, we can reduce noise incurred by missing labels, as well as alleviate the error accumulation of pseudo signals. To exploit labeled, sparsely labeled and optionally unlabeled images, we adopt the SOTA pseudolabel generation scheme inspired by [18, 30, 33].

Supervised loss for labeled images.

$$\mathcal{L}_l = \frac{1}{N_l}\sum_i \mathcal{L}_{cls}(x_l^i, y_l^i) + \mathcal{L}_{loc}(x_l^i, t_l^i), \qquad (6)$$

Pseudo-label generation.
The data batch is appended with randomly sampled sparse or unlabeled images if available. $\{b^i\}$

Pseudo-active synergy for sparse images.
For a sparsely labeled image x_s , the pseudo active synergy is exploited as follows:

$$G(y_s, \hat{y}_{sc}) = y_s \cup \{\hat{y}^i_{sc} \mid IoU(\hat{b}^i_{sc}, b^j_s) \leq \lambda_g, \forall b^j_s \in b_s\}, \tag{7}$$

The supervision quality for sparse images is thus enhanced after the completion:

$$\mathcal{L}_s = \frac{1}{N_s}\sum_i^{N_s} \mathcal{L}_{cls}(\mathcal{A}(x^i_s), G(y^i_s, \hat{y}^i_{sc})) + \mathcal{L}_{loc}(\mathcal{A}(x^i_s), G(t^i_s, \hat{t}^i_{sr})), \tag{8}$$

Mixed-supervision with unlabeled images.

$$\mathcal{L}_u = \frac{1}{N_u}\sum_i^{N_u} (\mathcal{L}_{cls}(\mathcal{A}(x^i_u), \hat{y}^i_{uc}) + \mathcal{L}_{loc}(\mathcal{A}(x^i_u), \hat{b}^i_{ur})), \tag{9}$$

Overall training objectives.
objective function for labeled-only setting : $\mathcal{L}_l + \frac{N_s}{N_l}\mathcal{L}_s$

objective function for mixed-supervision setting : $\mathcal{L}_l + \frac{N_s}{N_l}\mathcal{L}_s + \frac{N_u}{N_l}\mathcal{L}_u$

5. Conclusion

In this paper, we reveal the pitfalls of image-level evaluation for active detection and propose a realistic and fair box-level evaluation criterion. We then advocate efficient box-level annotation, under which we formulate a novel active detection pipeline, namely Complementary Pseudo Active Strategy (ComPAS) to exploit both human annotations and machine intelligence. It evaluates box informativeness based on the disagreement amongst a near-free input-end committee for both classification and localization to effectively query challenging targets. Meantime, the detector model addresses the sparse training problem by pseudo-label generation for well-learned targets. Under both labeled-only and mixed-supervision settings on VOC0712 and COCO datasets, ComPAS outperforms competitors by a large margin in a unified codebase.