Robust Object Detection in Remote Sensing Imagery with Noisy and Sparse Geo-Annotations

Maximilian Bernhard
LMU Munich
Munich, Germany
bernhard@dbs.ifi.lmu.de

Matthias Schubert
LMU Munich
Munich, Germany
schubert@dbs.ifi.lmu.de

ABSTRACT

Recently, the availability of remote sensing imagery from aerial vehicles and satellites constantly improved. For an automated interpretation of such data, deep-learning-based object detectors achieve state-of-the-art performance. However, established object detectors require complete, precise, and correct bounding box annotations for training. In order to create the necessary training annotations for object detectors, imagery can be georeferenced and combined with data from other sources, such as points of interest localized by GPS sensors. Unfortunately, this combination often leads to poor object localization and missing annotations. Therefore, training object detectors with such data often results in insufficient detection performance. In this paper, we present a novel approach for training object detectors with extremely noisy and incomplete annotations. Our method is based on a teacher-student learning framework and a correction module accounting for imprecise and missing annotations. Thus, our method is easy to use and can be combined with arbitrary object detectors. We demonstrate that our approach improves standard detectors by 37.1% AP$_{50}$ on a noisy real-world remote-sensing dataset. Furthermore, our method achieves great performance gains on two datasets with synthetic noise. Code is available at https://github.com/mxbh/robust_object_detection.

1 INTRODUCTION

Over the last years, deep-learning-based object detectors have established themselves as an essential and highly valuable tool for the automated analysis of remote sensing imagery. Unfortunately, object detectors require a large number of annotated training samples and the quality of annotations is crucial for a high performance [7, 11, 14]. Therefore, annotations are usually created manually by human annotators. As this process is time-consuming and might require expert knowledge, the availability of sufficient annotations often becomes an obstacle in practice.

Thus, the optimal usage of already existing annotations is key to reducing labeling efforts and improving the applicability of object detectors. In geospatial applications, georeferenced imagery can be combined with geolocalized annotations from a different source. However, this procedure is usually faulty, leading to two common types of annotation errors: First, aligning GPS annotations with georeferenced imagery often results in certain misalignments and displacements between images and annotations [6, 17]. Second, the completeness of GPS annotations is often an issue as not all objects of interest might have been registered, e.g., objects on private ground in urban areas. Both effects can be observed in Figure 1.
To overcome the problems posed by these annotation error types, we present a unified framework for training object detectors with extremely noisy and sparse, i.e., incomplete, supervision. At the core of our approach is a correction module that takes model predictions and potentially noisy annotations as input and produces refined targets for stable training. Thus, it operates in an unsupervised way without needing clean annotations. We employ a teacher-student learning framework, in which a student is supervised by targets that have been corrected by a teacher network. In doing so, we achieve a robust training scheme that is able to cope with extremely low annotation quality. We conduct experiments on a real-world dataset and two other datasets with synthetic noise.

2 RELATED WORK

In [7, 11], methods for training object detectors under class label and bounding box noise were proposed. Furthermore, [1, 5, 10] present approaches for dealing solely with inaccurate bounding boxes. More precisely, [1] assumes a very specific setting where only point annotations are available. We extend this method by enabling end-to-end training, individual box coordinate correction, and adding a mechanism to deal with incomplete annotations.

The problem of Sparsely Annotated Object Detection (SAOD), i.e., object detection with missing annotations, was addressed in [14, 15, 18]. Here, we combine the pseudo-labelling scheme of [14] with our box correction method in order to deal with sparse and inaccurate annotations simultaneously.

Although being relevant in practice, the problem of object detection with extremely noisy and sparse annotations was hardly addressed in the past. Only recently, a method including mechanisms for these annotation error types was proposed in [16], but it was only developed and evaluated in the standard setting on the COCO dataset [8], which is generally considered clean. In contrast, we investigate extremely noisy settings, where standard methods do not suffice to cope with the annotation noise and fail. Moreover, [4, 19] present approaches for dealing with noisy and partial labels for semantic segmentation in remote sensing applications, whereas we are interested in the task of object detection.

3 PROBLEM SETTING

Given a dataset consisting of images $I \in \mathbb{R}^{3 \times H \times W}$, our goal is to train an object detector $\Phi$ that is able to predict the true class label $l = 1..L$ and a minimal bounding box $b = (x_1, y_1, x_2, y_2)$ for every object of interest in an image. Furthermore, the detector produces a probability score $s \in [0, 1]$ indicating its confidence for every predicted instance. Thus, the detector output is a set of triplets of bounding boxes and class labels and confidence scores, i.e., $\Phi(I) = \{ (b_p, \hat{p}, \hat{s}) \}_{p=1..N_{preds}}$. Instead of the true boxes and labels $\{(b_t, l_t)\}_{t=1..N_{true}}$, only a subset consisting of imprecise boxes $\{(\tilde{b}_t, \tilde{l}_t)\}_{t=1..N_{targets}}$ is available for supervision during training. That is, $N_{targets} \neq N_{true}$ as some target boxes are missing and, additionally, the coordinates of the available boxes do not describe the true extents of objects.

4 METHOD

We propose a training framework, which employs a correction module that takes unrefined targets as well as predictions of a detector $\Phi$ for a given image as input. Guided by the predictions, the correction module transforms noisy targets into a new and more reliable set of targets. As we address two types of annotation errors, missing annotations and imprecise boxes, our solution comprises two submodules, each tackling one type of error. In the following, we explain our general training framework. Afterward, we describe the aforementioned correction submodules.

4.1 Training Framework

We perform a form of teacher-student training. This is a common practice that reduces self-confirmation bias when models (partly) supervise themselves. In our case, self-supervision is employed to mitigate the effect of annotation noise. However, when self-supervision is erroneous, plain training may lead to an amplification and accumulation of errors, ultimately resulting in bad generalization and unstable training. An asymmetric teacher-student architecture such as ours is an effective means to counter these problems. In principle, our teacher-student framework is similar to the one proposed in [9]. Nonetheless, we additionally integrate our correction module to the framework. This is necessary as we solve a different task than [9], which was proposed for semi-supervised object detection. An overview of our training pipeline can be seen in Figure 2.

For every image $I$, we start by creating a weakly augmented version $I_T$ and a strongly augmented version $I_S$ of it. Thereby, we assume $I_S = g \circ h(I_T)$, where $h(\cdot)$ denotes a photometric transformation that can consist of operations like color jitter, blurring, or brightness modification. In contrast, $g(\cdot)$ denotes a geometric transformation such as rotation or flipping. As weak augmentations, i.e., for creating $I_T$, we solely employ random resizing and flipping.

We feed $I_T$ and $I_S$ into a teacher detector $\Phi_T$ and student detector $\Phi_S$, respectively. Thereafter, the teacher predictions $\hat{\Phi}_T(I_T) = \{(\hat{b}_p, \hat{l}_p, \hat{s}_p)\}_{p=1..N_{preds}}$ as well as the unrefined, noisy annotations $\{(b_t, l_t)\}_{t=1..N_{targets}}$ are used to create corrected targets...
C = \{(c_t, l_t)\}_{t=1...N_{targets}} by applying our proposed correction module. Based on these refined targets, the loss for the student detector is computed, i.e. \( L_{\text{student}} = L(\Phi\langle l_s, g(C)\rangle) \). Here, \( L \) denotes the loss function defined by the detector architecture. Also, note that we have to geometrically align the corrected targets \( C \) with the outputs of \( \Phi_{s} \) as they were produced based on different image versions. While \( \Phi_{s} \) is continuously updated with \( L_{\text{student}} \) and a suitable variant of SGD, the teacher model \( \Phi_{T} \) is an exponential moving average (EMA) of the student. That is, after every iteration, we update every weight parameter \( \theta_T \) of the teacher as a convex combination of itself and its corresponding student parameter \( \theta_S \), i.e. \( \theta_T \leftarrow a \theta_T + (1 - a) \theta_S \), where \( a \in [0, 1] \) is a keep rate. After training, the teacher \( \Phi_{T} \) is used for inference. Before we start with our teacher-student training, a detector is trained in a standard supervised manner on the unrefined data to provide a good initialization for the teacher and the student model.

4.2 Box Correction

To correct potentially noisy boxes, we apply a novel box correction algorithm to the unrefined targets and the teacher predictions. The main idea is to determine predicted boxes that are sufficiently overlapping with an unrefined target box and then replace it with a weighted average of the overlapping predicted boxes. More precisely, we start by selecting the noisy boxes \( T^{(i)} = \{i = 1...N_{targets} : l_i = 1\} \) and teacher predictions \( P^{(i)} = \{p = 1...N_{preds} : l_p = 1\} \) that have a certain class label \( l \) as we handle all classes separately. Next, we apply an iterative local averaging scheme that is similar to k-Means where the centroids correspond to the corrected boxes. We initialize the corrected boxes \( c_{t,i} \) as copies of the unrefined target boxes \( b_l \) for \( i \in T^{(i)} \). In each iteration, we assign each predicted box \( \hat{b}_p \) for \( p \in P^{(i)} \) to a corrected box \( c_{t,i} \) if it meets two conditions (linked with a logical conjunction):

\[
J_{c_{t,i}} = \{ p \in P^{(i)} | i = \arg \min_{p \in P^{(i)}} \delta(c_{t,i}, \hat{b}_p) \land \delta(b_l, \hat{b}_p) \leq d \}
\]

First, the corrected box \( c_{t,i} \) has to be closer to the predicted box \( \hat{b}_p \) than any other corrected box \( c_{t,j} \). As a suitable distance measure for rectangular bounding boxes, we use \( \delta(b_l, \hat{b}_p) = 1 - \text{IoU}(b_l, \hat{b}_p) \in [0, 1] \), where \( \text{IoU}(\cdot, \cdot) \) denotes the intersection over union. Let us note that if this choice does not suit the characteristics of the dataset, another measure such as GloIoU [13] can be used. The second condition for the assignment of a predicted box \( \hat{b}_p \) to a corrected box \( c_{t,i} \) is that it is sufficiently close to the original, unrefined target box \( b_l \), i.e. their distance \( \delta(b_l, \hat{b}_p) \) may not exceed a threshold \( d \in \mathbb{R}^+ \). This design choice ensures that very distant predicted boxes (possibly false positives) do not impact the corrected boxes.

After the assignment of predicted boxes, the corrected boxes \( c_{t,i} \) are updated as the weighted average of their assigned predicted boxes, i.e. \( c_{t,i} = \{\sum_{p \in J_{c_{t,i}}} w_p \cdot \hat{b}_p\}_{p=1...N_{preds}} \), where \( w_p \) is the softmax of \( \{\hat{b}_p\}_{p \in J_{c_{t,i}}} \). The average is computed for each of the four box coordinates \( \square = x_1, y_1, x_2, y_2 \) separately. To obtain the weights, we softmax the scores of the assigned boxes (note that, in contrast to the rest of the paper, we use raw logit scores instead of probability scores here). Thus, predicted boxes with high confidence receive larger weights. Moreover, we use the softmax temperature \( \gamma \) as an additional hyperparameter that further allows controlling the impact of low-scoring boxes. We iterate assignments and updates until the corrected boxes do not change anymore and output the final corrected boxes denoted as \( C_B = \{(c_t, l_t)\}_{t=1...N_{targets}} \).

4.3 Sparse Label Correction

To address the problem of missing annotations causing false negatives during training, we employ a pseudo-labeling scheme that adds confident predictions to the incomplete set of available targets.

The inputs for our pseudo-label generation are the teacher predictions \( \Phi_T(I_T) \) for an image and the output of the box correction algorithm \( C_B \). In the first step, all the predictions with a predicted probability \( \hat{p} \) of less than a specified threshold \( \tau \in [0, 1] \) are removed. Next, non-maximum suppression (NMS) is performed on the remaining predicted boxes to remove redundant predictions. For this further reduced set of predicted boxes, the IoUs with the target boxes \( c_{t,i} \) are computed. All the boxes with a sufficiently large overlap (i.e. \( \text{IoU} > 0.5 \)) with any of the target boxes are removed. Finally, the remaining boxes are added to the new set of target boxes. More formally, the extended and final set of corrected targets \( C \) arises as

\[
C = C_B \cup \{(b_p, l_p) \mid \exists_{b_p} : (b_p, l_p, \hat{p}) \in \text{NMS}(\Phi_T(I_T)) \land \hat{p} \geq \tau; \\
\forall(c_t, l_t) \in C_B : l_t = l_p \land \text{IoU}(b_p, c_t) > 0.5\}
\]

where \( C_B \) is the output of the box correction module (consisting of boxes and labels). This new set of targets \( C \) is used as supervision for the student model \( \Phi_S \).
the GPS locations do not always match the imagery well (see Figure 1). The dataset consists of 3,536 tiles of 512 by 512 pixels. For validation and testing, clean annotations were created. As training on Edmonton Trees dataset was subject to relatively high variance, we report all scores for this dataset as the average and standard deviation over five runs. Furthermore, we ran experiments on the commonly used multi-class object detection datasets NWPU VHR-10 [2] and Pascal VOC [3]. Since the available annotations for these datasets can be considered clean, we introduce synthetic noise. For the bounding box noise, we follow [7] and displace every horizontal box coordinate by a number of pixels randomly chosen from the interval \([-\text{w}N_h, +\text{w}N_h]\), where \(\text{w}\) denotes the pixel width of the bounding boxes and \(N_h\) is the box noise level. Analogously, every vertical box coordinate is moved by a random number of pixels in the interval \([-\text{h}N_v, +\text{h}N_v]\), where \(\text{h}\) is the height of the box. We report results for the three different box noise levels \(N_h = 0\), \(N_h = 20\%\) and \(N_h = 40\%\). Furthermore, we sparsify the annotations by randomly removing box annotations. In the settings \(N_h = 0\%\) and \(N_h = 50\%\), the corresponding fraction of annotations is dropped. Additionally, we introduce another extreme level of sparse annotations \(N_v = \infty\), where only one annotation per image is kept.

5.2 Results

In Table 1, we provide the results for the different datasets and noise levels in our experiments. When comparing our training method with the standard training of Faster R-CNN [12], we observe large gains in every setting. Interestingly, the differences between \(N_h = 50\%\) and \(N_h = \infty\) are mostly rather small for our method, while they are substantial for standard training. On Edmonton Trees with real-world annotation noise, the improvement of our method is particularly remarkable, achieving an average of 79.8\% and a maximum of 81.2\% \(AP_{50}\) over five runs. These values indicate that our method is indeed capable of ensuring robust training in real-world applications.

Furthermore, we applied our method in the clean setting \((N_h = 0\%, N_v = 0\%)\). For NWPU VHR-10, we observe a small improvement of 1.6 points in \(AP_{50}\). However, when we removed the correction module in the teacher-student training, we observed a \(AP_{50}\) gain of 1.8 points. For Pascal VOC, the performance dropped after initialization with the vanilla model — independent of whether the correction module was used or omitted. Thus, we cannot conclude that generating corrected pseudo-labels with our correction module leads to improvements when clean annotations are available.

6 CONCLUSION

In this paper, we proposed a method for training object detectors in extremely noisy settings with incomplete and imprecise bounding box annotations. Its modular design and its effectiveness on both simulated and real-world annotation noise make it valuable in many practical scenarios where clean annotations are not available. We are convinced that the remote sensing community benefits from our work, as our method allows robust training of object detectors without the necessity of manually labeling large amounts of images, therefore removing a considerable barrier to employing object detection in practice.

ACKNOWLEDGMENTS

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content. Additionally, we thank Guillermo Castilla for the fruitful discussion in the early phase of this project.

REFERENCES

[1] Maximilian Bernhard and Matthias Schubert. 2021. Correcting Imprecise Object Locations for Training Object Detectors in Remote Sensing Applications. Remote Sensing 13, 24 (2021). https://doi.org/10.3390/rs13244962
[2] Gong Cheng, Peicheng Zhou, and Junwei Han. 2016. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing 54, 12 (2016), 7405–7415. https://doi.org/10.1109/TGRS.2016.2601622
[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. International journal of computer vision 88, 2 (2010), 303–338.
[4] Simone Fobi, Terence Conlon, Jayant Taneya, and Vijay Modi. 2020. Learning to Segment from Misaligned and Partial Labels. Association for Computing Machinery, New York, NY, USA, 286–290. https://doi.org/10.1145/3378393.3402254
[5] Jiabo He, Sarah Erfani, Xingjun Ma, James Bailey, Ying Chi, and Xian-Sheng Hua. 2021. Alpha-IoU: A Family of Power Intersection over Union Losses for Bounding Box Regression. Advances in Neural Information Processing Systems 34 (2021).
[6] Michael L. Hughes, Patricia F. McDowell, and W. Andrew Marcus. 2006. Accuracy assessment of georectified aerial photographs: Implications for measuring lateral channel movement in a GIS. Geomorphology 74, 1 (2006), 1–16. https://doi.org/10.1016/j.geomorph.2005.07.001
[7] Junnan Li, Caiming Xiong, Richard Socher, and Steven Hoi. 2020. Towards noise-resistant object detection with noisy annotations. arXiv preprint arXiv:2003.01285 (2020).
[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In European conference on computer vision. Springer, 740–755.
[9] Yeu-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zuoqi Kira, and Peter Vajda. 2021. Unbiased teacher for semi-supervised object detection. arXiv preprint arXiv:2102.09480 (2021).
[10] Jiafeng Mao, Qiqi Yu, and Kiyoharu Aizawa. 2021. Noisy localization annotation refinement for object detection. IEEE Transactions on Information and Systems 104, 9 (2021), 1478–1485.
[11] Jiafeng Mao, Qiqi Yu, Yoko Yamakata, and Kiyoharu Aizawa. 2021. Noisy Annotation Refinement for Object Detection. arXiv preprint arXiv:2110.10456 (2021).
[12] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Stephen Lin, Han Hu, and Xiang Bai. 2021. Bootstrap Your Object Detector via Mixed Training. Advances in Neural Information Processing Systems 34 (2021).
[13] Maximilian Bernhard and Matthias Schubert. 2021. Correcting Imprecise Object Locations for Training Object Detectors in Remote Sensing Applications. Remote Sensing 13, 24 (2021). https://doi.org/10.3390/rs13244962
[14] Armand Zampieri, Guillaume Charpiat, Nicolas Girard, and Yuliya Tarabalka. 2018. Multimodal image alignment through a multiscale chain of neural networks. CoRR abs/1806.06986 (2018). arXiv:1806.06986 http://arxiv.org/abs/1806.06986
[15] Zhe Wu, Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. 2018. Soft Sampling for Robust Object Detection. CoRR abs/1806.06986 (2018). arXiv:1806.06986 http://arxiv.org/abs/1806.06986
[16] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Stephen Lin, Han Hu, and Xiang Bai. 2021. Bootstrap Your Object Detector via Mixed Training. Advances in Neural Information Processing Systems 34 (2021).
[17] Armand Zampieri, Guillaume Charpiat, Nicolas Girard, and Yuliya Tarabalka. 2018. Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing. In Proceedings of the European Conference on Computer Vision (ECCV). 657–673.
[18] Han Zhang, Fangyi Chen, Zhiqiang Shen, Qiqi Hao, Chenchen Zhu, and Marius Savvides. 2020. Solving Missing-Annotation Object Detection with Background Recalibration Loss. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 1888–1892.
[19] Stefano Zorzzi, Ksenia Bittner, and Friedrich Fraundorfer. 2020. Map-Repair: Deep Self-supervised Learning for Sparsely Annotated Object Detection. arXiv preprint arXiv:2005.09450 (2020).
[20] Stefano Zorzzi, Ksenia Bittner, and Friedrich Fraundorfer. 2020. Map-Repair: Deep Self-supervised Learning for Sparsely Annotated Object Detection. arXiv preprint arXiv:2005.09450 (2020).
[21] Ying Chang, Qi Wang, Zhe Wu, and Kiyoharu Aizawa. 2021. Recalibration Loss. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 658–666.
[22] Tiancai Wang, Tong Yang, Jiale Cao, and Xiangyu Zhang. 2020. Co-mining: Self-supervised learning for sparsely annotated object detection. arXiv preprint arXiv:2012.01950 (2020).
[23] Zhe Wu, Navaneeth Bodla, Bharat Singh, Mahyar Najibi, Rama Chellappa, and Larry S. Davis. 2018. Soft Sampling for Robust Object Detection. CoRR abs/1806.06986 (2018). arXiv:1806.06986 http://arxiv.org/abs/1806.06986
[24] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Stephen Lin, Han Hu, and Xiang Bai. 2021. Bootstrap Your Object Detector via Mixed Training. Advances in Neural Information Processing Systems 34 (2021).
[25] Maximilian Bernhard and Matthias Schubert. 2021. Correcting Imprecise Object Locations for Training Object Detectors in Remote Sensing Applications. Remote Sensing 13, 24 (2021). https://doi.org/10.3390/rs13244962