Learning to segment from misaligned and partial labels

Simone Fobi
Columbia University
sf2786@columbia.edu

Terence Conlon
Columbia University
tmc2180@columbia.edu

Jayant Taneja
University of Massachusetts Amherst
jtaneja@umass.edu

Vijay Modi
Columbia University
modi@columbia.edu

ABSTRACT

To extract information at scale, researchers are increasingly applying semantic segmentation techniques to remotely-sensed imagery. While fully-supervised learning enables accurate pixelwise segmentation, compiling the exhaustive datasets required is often prohibitively expensive, and open-source datasets that do exists are frequently inexact and non-exhaustive. In this paper, we present a novel and generalizable two-stage framework that enables improved pixelwise image segmentation given misaligned and missing annotations. First, we introduce the Alignment Correction Network to rectify incorrectly registered open source labels. Next, we demonstrate a segmentation model – the Pointer Segmentation Network – that uses corrected labels to predict infrastructure footprints despite missing annotations. We demonstrate the transferability of our method to lower quality data sources by applying the Alignment Correction Network to correct OpenStreetMaps building footprints, and we show the accuracy of the Pointer Segmentation Network in predicting cropland boundaries in California. Overall, our methodology is robust for multiple applications with varied amounts of training data present, thus offering a method to extract reliable information from noisy, partial data.

CCS CONCEPTS

• **Computing methodologies** → *Artificial intelligence*; *Computer vision*; *Computer vision problems*; • **Applied computing** → *Physical sciences and engineering*.

KEYWORDS

segmentation; misaligned and missing labels; open source data

ACM Reference Format:
Simone Fobi, Terence Conlon, Jayant Taneja, and Vijay Modi. 2020. Learning to segment from misaligned and partial labels. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '20), June 15–17, 2020, , Ecuador*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3378393.3402254

COMPASS '20, June 15–17, 2020, , Ecuador

ACM ISBN 978-1-4503-7129-2/20/06…$15.00
https://doi.org/10.1145/3378393.3402254

Figure 1: Types of label noise present in open source data. Building footprints are the class of interest

1 INTRODUCTION

Processing remotely-sensed imagery is a promising approach to evaluate ground conditions at scale for little cost. Algorithms that intake satellite imagery have accurately measured crop type [12],[9], cropped area [6], building coverage [15] [14], urbanization [1], and road networks [4] [16]. However, successful implementation of image segmentation algorithms for remote sensing applications depends on large amounts of data and high-quality annotations. Although some global ground truth datasets like OpenStreetMaps (OSM) offer large amounts of labels for use at no cost, their annotations suffer from multiple types of noise [10] [2], including: *Missing or omitted annotations*, defined as objects being present in the image and not existing in the label [10]; *misaligned annotations* occur when annotations are translated and/or rotated from their true position [13]; and *incorrect annotations* – annotations that do not directly correspond to the object of interest in the image. Figure 1 presents examples of these types of label noise. For applications such as measuring building or field area that are useful in downstream analyses of wealth, crop yield and more, high noise levels decrease the ability to segment images successfully.

To address issues of misaligned and omitted annotations, and to extract information from imperfect data, we present a simple and generalizable method for pixelwise image segmentation. First, we address annotation misalignment by proposing an Alignment Correction Network (ACN). With a small number of images and human verified ground truth annotations, the ACN learns to correct misaligned labels. Next, the corrected open source annotations are used to train the Pointer Segmentation Network (PSN), a model which takes in a point location and identifies the object containing that point. Learning associations from a representative point is a widely acknowledged method of object detection [3]. By *pointing out* an object of interest, our network ignores other instances that may not have corresponding annotations, therefore preventing performance

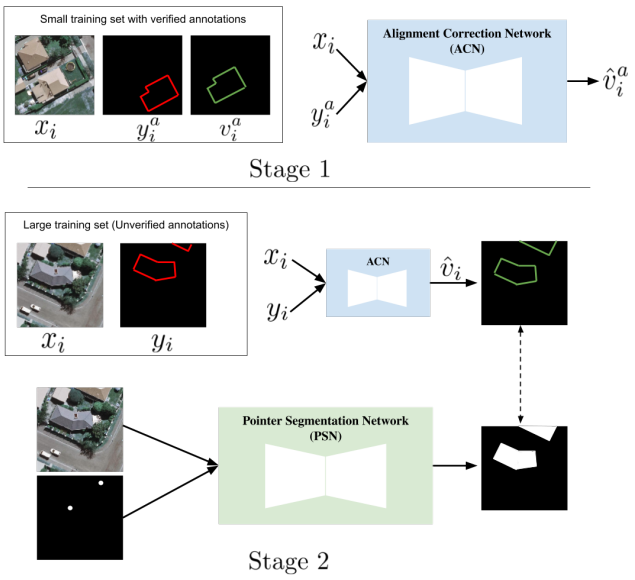

Figure 2: Summary of our two-stage approach to segment from noisy annotations. Stage 1: The ACN uses an image (x_i) and label (y_i^a) with a single misaligned annotation to predict a realigned annotation $\hat{v}_i^a$. Random shifts between ±10 pixels are applied to v_i^a to obtain y_i^a. The network is trained with a small set of images (x) and verified ground truth annotations (v). Stage 2: A large noisy training set is first realigned with the ACN. Realigned, incomplete annotations are used for supervision. The PSN uses selected points from available instances, x_i and $\hat{v}_i$ to learn segmentation.

degradation caused by annotation-less instances. Our two-step approach does not require a custom loss to handle noisy labels. While our approach cannot replace large amounts of careful annotations, it can complement existing open source datasets and algorithms, reduce the cost of obtaining large amounts of full annotations, and allow researchers to extract information from imperfect datasets. For details about our work and related literature, please see the full version of our paper here: https://arxiv.org/abs/2005.13180

2 METHODS

In this work, we propose a novel two-stage approach to learn segmentation from noisy labels. See Figure 2. Binary cross-entropy loss is used for all networks. Both ACN and PSN use a baseline architecture (lightUNet) shown in the full version of our paper, albeit with the number of input channels modified.

2.1 Alignment Correction Network

Given inputs of an image x_i and a label y_i^a containing one misaligned instance a, the ACN outputs a label $\hat{v}_i^a$ containing the predicted, realigned annotation. When multiple misaligned instances are present in an image, the instances are corrected independently. We choose this approach for two reasons: it allows instances within an image to have varying degrees of translation error and also enables the network to be robust to incomplete labels with missing instances. A small dataset of images (x) and carefully verified labels (v) is used to train the ACN. Noisy labels y_i^a are generated by applying random shifts to v_i^a.

2.2 Pointer Segmentation Network

This network learns to segment an image using only m available annotations. The PSN takes as inputs an image x_i and a single channel of points specifying selected instances to be segmented; it outputs a segmentation mask only for the selected instances. We specify the fraction of instances to be used for training using a parameter α, where α is the number of selected instances divided by the number of available instances. By including a single channel containing the points $p_i(\alpha)$, our PSN segments only instances that are associated with the points. This approach offers two benefits: first, learning is simplified to specify instances of interest; second, the network can be trained without custom the losses to handle noise. The model is trained by randomly picking α for every image in each epoch. At inference time, all instances of interest are specified using points.

To sequentially test the two-stage approach, the ACN is used to correct a training dataset that is then inputted to the PSN for object segmentation. A longer discussion of our methodology is available in our full report.

3 DATA

Three separate datasets (all described below) are used to train and test ACN and PSN performances. During training and testing, we only use images that contain labels.

The Aerial Imagery for Roof Segmentation (AIRS) dataset is used to establish baseline performances for both the ACN and PSN. The AIRS dataset covers most of Christchurch ($457km^2$), New Zealand and consists of orthorectified aerial images (RGB) at a spatial resolution of 7.5 cm with over 220,000 building annotations; the images are split into a training set T_{set} and a validation set V_{set}[5]. Other than basic filtering to remove images with <10% building cover, we fully preserve the training and validation sets.

OSM provides open-source building footprints for many parts of the world; unfortunately, label quality is highly heterogeneous. In order to test the performance of the ACN on these incomplete and misaligned building footprints, we pair OSM annotations for Kenya [7] with selected DigitalGlobe tiles from Western Kenya (a box enclosed by 0.176 S, 0.263 S, 34.365 E, and 34.453 E) and closer to Nairobi (a box enclosed by 1.230 S, 1.318 S, 36.738 E, and 36.826 E). We generated human verified ground truth annotations for 500 of the image patches.

Crop maps and decameter imagery is used to demonstrate the flexibility of the PSN. The California Department of Water Resources provides a Statewide Cropping Map for 2016 [11]; we pair this shapefile with Sentinel-2 satellite imagery to predict cropland parcels [8].

4 RESULTS

For all model testing, we report the mean intersection over the union (mIOU), defined as the intersection of the predicted and true label footprints divided by the union of the same footprints, averaged across the dataset.

4.1 Alignment Correction Network

We evaluate the performance of the ACN on the AIRS V_{set}. We generate random translations between ± 10 pixels for the xy-axis and apply them to ground truth AIRS annotations, resulting in unique translation shifts for each object in an image. The introduction of noise through random translation yields a baseline mIOU of 0.55. The shifted annotations together with the images are fed into the ACN, and the corrected annotations are compared to the true annotations to drive the learning process. For the first set of tests, we report the mIOU on V_{set} when varying amounts of T_{set} data are used for training; random translations between ± 10 pixels are applied to all objects in V_{set}. When the ACN is trained with 800, 400 and 240 images, the corresponding mIOUs on all images in V_{set} are 0.81, 0.77 and 0.67 respectively, compared to the baseline of 0.55. This suggests that the ACN performs better when more images are used, but also that it can learn with only a couple hundred training images.

Using the ACN model trained with 400 images and random translation shifts between ± 10 pixels, we next evaluate the robustness of the ACN to varying levels of translation shifts. Table 1 shows mIOU before and after ACN correction for different ranges of translations shifts in V_{set}. Across all translation shifts, the ACN realigns shifted annotations, even for translations larger than those in the training dataset (>10 pixels).

Table 1: mIOU before and after ACN realignment, trained on 400 images.

	mIOU	
Translation Shift (± pixels)	Before ACN	After ACN
0 to 5	0.63	0.81
5 to 10	0.40	0.73
10 to 15	0.26	0.46
15 to 20	0.18	0.28

4.2 Pointer Segmentation Network

The performance of the PSN is compared against a baseline model (lightUNet) for segmention of every building instance in AIRS V_{set}. Both models are trained on partial but well-aligned images from AIRS T_{set}. Table 2 reports the performance of the lightUNet and the PSN with varying fractions of selected annotations (α): As α decreases, performance of the PSN remains robust, indicating that the network still learns the segmentation task despite missing annotations.

Table 2 compares the performance of the PSN when building centroids are used as inputs. When randomly generated (non-centroid) points within the building footprint are used with α = 0.7, we observe a drop in mIOU of the PSN from 0.89 to 0.83. This suggests that annotators should strive to extract points near the center of buildings to ensure better segmentation outcomes. Additionally, we evaluate how both networks handles heterogeneous (Het.) amounts of label completeness by sampling α from a random uniform distribution between 0 and 1; α is resampled for each image during every training epoch. Here we find that the PSN suffers no significant performance degradation for a random distribution of α. Figure 3 shows some outputs of PSN and lightUNet models when both are trained with α = 0.7. Although both networks are trained with missing annotations, generated annotations from the PSN appear more visually accurate.

Table 2: mIOU for all buildings in AIRS V_{set}.

	mIOU	
α	PSN	lightUNet
1	0.9	0.85
0.7	0.89	0.53
0.5	0.87	0.18
Het.	0.87	0.71

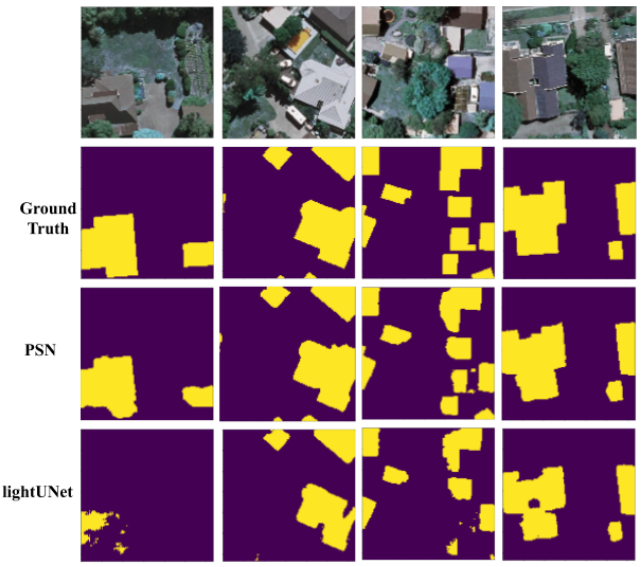

Figure 3: Annotations from PSN and lightUNet models when trained with α = 0.7. Predictions are made for all building instances in the image and are compared to the ground truth.

4.3 Sequential Testing

We next use the AIRS dataset to evaluate the performance of our two-stage methodology where the ACN and PSN are trained and tested sequentially. Using T_{set}, we establish two training datasets for the sequential process: T1, containing misaligned labels generated from the true T_{set}; and T2, containing ACN-corrected T1 labels.

The ACN model trained on 400 images generates T2. The noise present in both training datasets is captured by the mIOU listed in Table 3. The PSN and lightUNet models are trained with T1 and T2 using $\alpha = Het$, an

Table 3: Dataset noise and performance of segmentation architectures.

	mIOU
T1: Misaligned train data	0.57
PSN (trained on T1)	0.54
lightUNet (trained on T1)	0.17
T2: ACN-corrected train data	0.81
PSN (trained on T2)	0.79
lightUNet (trained on T2)	0.74

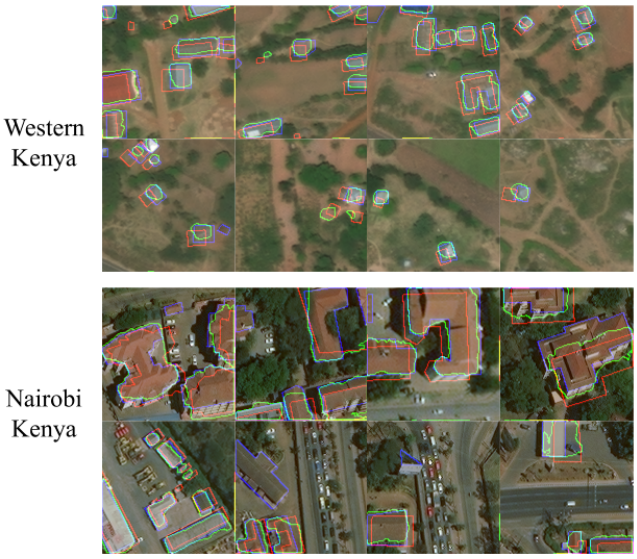

Figure 4: Hand-labelled ground truth annotations, OSM annotations and ACN-corrected annotations. The ACN is trained on 400 images from Western Kenya and Nairobi, and improves label alignment despite the noisier training data.

identical implementation of label withholding to that described in the previous section. The trained models are used to segment V_{set} images; we compare predicted annotations to the true annotations to attain the performance metrics reported in Table 3. The PSN performs significantly better than the lightUNet when trained on either misaligned labels (T1) or ACN-corrected labels (T2). Again, we find that with incomplete labels, regardless of alignment quality, the PSN outperforms the lightUNet. Moreover, in both training configurations, PSN mIOU performance nears that of the training dataset. As a result, we conclude that the PSN is able to predict object extents at a similar accuracy to that which exists in the training dataset.

4.4 ACN Application: Realignment of OSM Annotations

To confirm the performance of our realignment method on noisier images and labels, we tested the ACN on OSM building polygons in Kenya, a dataset containing considerable amounts of label misalignment. Of the 500 human-verified ground truth image labels generated for Kenya, we use 400 to train the ACN and 100 to validate. The extent of noise in the OSM labels is measured by comparing the labels to the human-verified ground truth labels: We calculate mIOUs of 0.30 and 0.31 for the train and validation data respectively. OSM training labels are used to train the ACN and the trained model is ran on the 100 validation labels. A 50% improvement in mIOU from 0.31 to 0.47 is observed on the 100 validation images. Figure 4 shows a sampling of ACN-corrected OSM annotations for images in the validation dataset. Overall, we find that the ACN is able to correct misaligned OSM annotations both in rural and urban regions. In rural Western Kenya, where buildings tend to be smaller, the ACN shifts OSM footprints to better align with the buildings. We observe that the noisier image quality makes it more difficult for the ACN to identify extremely small buildings. Sometimes when no buildings are present the model does not predict a correction. In more urbanized Nairobi, the ACN also improves the alignment of OSM annotations, albeit with some failure cases. Overall, our results suggest that the ACN is transferable to noisy datasets.

4.5 PSN Application: Cropland Segmentation

We also apply the PSN to the task of cropland segmentation using Sentinel-2 imagery and a 2016 California cropping map. Similar to previously described tests, the performance of the PSN and lightUNet in recreating field boundaries is measured for different values of α. At all fractions of available training data shown in the table, the PSN outperforms the lightUNet. After 40 training epochs, the PSN is able to predict all field boundaries for the test set across all values of α; in contrast, the performance of the lightUNet degrades significantly as field boundaries are withheld. Here, we find that the PSN can accurately predict cropland extents using only a subset of fields, and does not require the comprehensive set of training polygons that would be necessary for traditional segmentation networks.

Table 4: mIOU for all field boundaries in the test set.

	mIOU	
α	PSN	lightUNet
1	0.92	0.75
0.75	0.91	0.69

5 CONCLUSION

In this paper, we present a novel and generalizable two-stage segmentation approach that addresses common issues in applying deep learning approaches to remotely-sensed imagery with unreliable ground truth data. First, we present the Alignment Correction Network (ACN), a model which learns to correct misaligned instance annotations. We test the ACN on a set of alignment errors, including misalignment of the AIRS dataset and existing, substantial misalignment errors within the OSM Kenyan building footprint dataset. Overall, we find that the ACN significantly improves annotation alignment accuracy. We next introduce the Pointer Segmentation Network (PSN), a model which reliably predicts an object's extent using only a point from the object's interior. For all segmentation tasks and testing configurations – those which vary the fraction of available training annotations and those which change the location of the training point – the PSN outperforms a baseline segmentation model. Lastly, we sequentially link the ACN and PSN to demonstrate the ability of the combined networks to accurately segment objects having learnt from misaligned and incomplete training data. Taken together, we envision our proposed networks providing value to the community of researchers and scientists looking to extract information from widely-available satellite imagery and unreliable ground-truth datasets.

REFERENCES

[1] Rasha Alshehhi, Prashanth Reddy Marpu, Wei Lee Woon, and Mauro Dalla Mura. 2017. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing* 130 (2017), 139–149.

[2] Anahid Basiri, Mike Jackson, Pouria Amirian, Amir Pourabdollah, Monika Sester, Adam Winstanley, Terry Moore, and Lijuan Zhang. 2016. Quality assessment of OpenStreetMap data using trajectory mining. *Geo-spatial information science* 19 (2016), 56–68.

[3] Amy Bearman, Vittorio Ferrari Olga Russakovsky, and Li Fei-Fei. 2016. What's the Point: Semantic Segmentation with Point Supervision. *European conference on computer vision* (2016).

[4] Gabriel Cadamuro, Aggrey Muhebwa, and Jay Taneja. 2019. Street smarts: measuring intercity road quality using deep learning on satellite imagery. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS 2019)*. 145–154.

[5] Qi Chen, Lei Wang, Yifan Wu, Guangming Wu, Zhiling Guo, and Steven L. Waslander. 2018. Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. *arXiv preprint arXiv:1807.09532* (2018).

[6] Zhenrong Du, Jianyu Yang, Cong Ou, and Tingting Zhang. 2019. Smallholder crop area mapped with a semantic segmentation deep learning method. *Remote Sensing* 11, 7 (2019), 888.

[7] Humanitarian Data Exchange. 2020. HOTOSM Kenya Buildings (OpenStreetMap Export). Retrieved February 27, 2020 from https://data.humdata.org/dataset/hotosm_ken_buildings.

[8] Ferran Gascon, Catherine Bouzinac, Olivier ThÃĽpaut, Mathieu Jung, Benjamin Francesconi, JÃľrÃťme Louis, Vincent Lonjou, Bruno Lafrance, StÃľphane Massera, AngÃľlique Gaudel-Vacaresse, Florie Languille, Bahjat Alhammoud, FranÃğoise Viallefont, Bringfried Pflug, Jakub Bieniarz, SÃľbastien Clerc, LaÃńtitia Pessiot, Thierry TrÃľmas, Enrico Cadau, Roberto De Bonis, Claudia Isola, Philippe Martimort, and ValÃľrie Fernandez. 2017. Copernicus Sentinel-2A Calibration and Products Validation Status. *Remote Sensing* 9, 6 (2017). https://doi.org/10.3390/rs9060584

[9] Natalia Kussul, Mykola Lavreniuk, Sergii Skakun, and Andrii Shelestov. 2017. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geoscience and Remote Sensing Letters* 14, 5 (2017), 778–782.

[10] Volodymyr Mnih and Geoffrey E. Hinton. 2012. Learning to label aerial images from noisy data. In *Proceedings of the 29th International conference on machine learning*.

[11] California Department Of Water Resources. 2020. 2016 California Statewide Agricultural Land Use Map. Retrieved February 27, 2020 from https://gis.water.ca.gov/app/CADWRLandUseViewer/.

[12] Rose M Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. 2019. Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2019), 75–82.

[13] John E Vargas-Munoz, Sylvain Lobry, Alexandre X. Falcao, and Devis Tuia. 2019. Correcting rural building annotations in OpenStreetMap using convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing* 147 (2019), 283–293.

[14] G. Wu and Z. Guo. 2019. GeoSeg: A computer Vision Package for Automatic Building Segmentation and Outline Extraction. *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium* (2019).

[15] Yongyang Xu, Liang Wu, Zhong Xie, and Zhanlong Chen. 2018. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sensing* 10, 1 (2018), 144.

[16] Yongyang Xu, Zhong Xie, Yaxing Feng, and Zhanlong Chen. 2018. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sensing* 10, 9 (2018), 1461.