Interactive Segmentation for Disaster Relief Mapping

Muhammed Razzak and **Kris Sankaran**
Montréal Institute for Learning Algorithms
{razzakmu, kris}@mila.quebec

Abstract

Annotation of aerial imagery of disaster relief sites is a critical and time sensitive task, for disaster relief organisations. Manual annotation requires large teams of volunteers and a significant time commitment. Despite the improved performance of segmentation models they still struggle, particularly in new environments. Thus manual annotation is still the standard as these organisations need accurate and verified information. This paper explores the use of points as a supervisory signal in an interactive correction scenario to obtain accurate semantic segmentation of large scale aerial imagery: (1) we show that a few points increases mIoU performance by 15% and (2) it works on and provides larger performance increase for out-of-distribution data.

1 Introduction

Disaster relief organisations, such as the American Red Cross and Médecins sans Frontières, need accurate maps of disaster affected regions to reach those in need, prioritise resources, determine local rally points, and set evacuation routes. In many cases, maps of these areas do not exist or are inaccurate and so need to be generated rapidly. These organisations currently leverage an army of volunteers to complete the mapping [Dittus *et al.*, 2017].

Currently, the workflow developed to map certain objects (e.g. buildings) using computer vision involves doing segmentation on aerial imagery, followed by converting the segmentation masks into polygons for each object. During the initial stage of our collaboration with American Red Cross, to help speed up mapping of disaster relief areas, a number of insights with regards to the mapping process were gleaned: (1) areas requiring mapping are fairly heterogeneous and likely differ from any previously acquired training data; (2) given the critical nature of the task and current state of segmentation (particularly in the context of heterogeneous areas), volunteers would still be required to validate and correct the generated maps before they could be uploaded and used. Thus, the question has become how could we better leverage current automated segmentation models and the volunteers' interactions to speed up the mapping process. We focus here on interactive segmentation in response to this problem.

While interactive segmentation has been well studied in the traditional computer vision settings, the effectiveness of interactive semantic segmentation for this large scale remote sensing setting is unclear (where there is more than simply a few objects to label per image). Furthermore, the utility of interactivity for out-of-distribution data is also unclear, in both the traditional and remote sensing setting.

Our contribution is summarised as follows:

- In section 4, we provide a refined framework for experimentation, that addresses the real-world problem faced.
- In section 5.1, we show that a small number of interactive corrections (9 or less) improves segmentation masks in large scale semantic segmentation, with qualitative results being particularly compelling.
- In section 5.2, we show that interactive corrections improves out-of-distribution performance significantly, with mIOU increasing from 0.66 to 0.76 from 9 interactions.

2 Related Work

Interactive Segmentation: Until the advent of deep learning, graphical models formed the basis of interactive segmentation. The seminal pieces of literature in the area is work of [Boykov and Jolly, 2002], which uses graph cuts to segment objects in images based on labelled pixels provided by user, and GrabCut [Rother *et al.*, 2004], which segments from bounding boxes by iteratively updating the model. More recent methods make use of fully convolutional networks. iFCN [Xu *et al.*, 2016] makes use of positive and negative clicks, in the spirit of [Boykov and Jolly, 2002], to guide the segmentation. This deep learning pipeline was extended to GrabCut, with Deep GrabCut [Xu *et al.*, 2017]. Other forms of supervision have been employed with similar success, such as bounding boxes [Dai *et al.*, 2015], scribbles [Lin *et al.*, 2016] and extreme points [Maninis *et al.*, 2018].

The current state-of-the-art have moved beyond just providing an initial supervisory signal towards a corrective approach, with one [Mahadevan *et al.*, 2018] or more models [Benenson *et al.*, 2019] used in the pipeline.

An alternative approach was developed in Polygon-RNN [Castrejon *et al.*, 2017] and further developed in [Acuna *et al.*, 2018]. These methods predict a polygon outlining of the object to be segmented. The polygons can then interactively be corrected. Of note, all these methods perform single or few object segmentation. We study multi-object segmentation on the scale of 10 to 100 of objects per image.

In the context of remote sensing or aerial imagery, interactive segmentation has largely been unexplored. GrabCut has been applied to remote sensing images with some success [Yang *et al.*, 2017], however the no interactive methods involving a deep learning pipeline have been tested.

Semantic Segmentation: There is significant body of literature pertaining to semantic segmentation for aerial imagery. The progress in the field has largely followed the progress in the computer vision field, with the use of deep fully convolutional networks being pervasive [Long *et al.*, 2015]. Notably, the architectures in widespread use are encoder-decoder networks [Ronneberger *et al.*, 2015; Badrinarayanan *et al.*, 2017]. In order to create finer segmentation masks the use of CRFs were introduced [Chen *et al.*, 2018]. These models have been adapted for large scale semantic segmentation[Maggiori *et al.*, 2017; Bastani *et al.*, 2018; Iglovikov *et al.*, 2018], where breaking an image into patches for processing is common strategy.

Out-of-Distribution Adaption: Performance degradation of deep learning models, when they are used on out-of-distribution data is well-documented. Given the large quantities of labelled data required to train deep learning models, tackling the domain adaption problem has been at the forefront of the vision community recently with a number of different approaches being taken [Wang and Deng, 2018]. In remote sensing community this problem has also surfaced. The performance discrepancy in segmentation models between different cities, weather and seasonality is evident [Maggiori *et al.*, 2017]. We show quantitatively this degradation, and that interactivity can significantly improve a models out-of-distribution performance, given cheap additional signals at test time.

3 Models

We assess two models: U-Net [Ronneberger *et al.*, 2015] as a baseline and our own. We made a conscious decision against testing certain algorithms that are considered state-of-the-art, such as DEXTR [Maninis *et al.*, 2018] and RIS-Net [Liew *et al.*, 2017], or baselines in the field [Rother *et al.*, 2004; Xu *et al.*, 2017] as they are ill-suited to this task. These algorithms require bounding boxes to be drawn around the objects and/or multiple interactions per object. For large scale annotation of buildings/bridges, this sort of interactivity does not provide a significant reduction in annotation time, due to the numerous initial interactions required for each object.

The focus of this paper is to assess the utility of the interactive correction algorithm. As such we use the simple encoder-decoder architecture, U-Net, as both the baseline and the backbone for our network. This architecture could be replaced by a more recent and performant one, such as Deeplabv3 [Chen *et al.*, 2018] or the DenseNet Tiramisu [Jegou *et al.*, 2017].

3.1 U-Net

The U-Net architecture [Ronneberger *et al.*, 2015] has developed into the baseline for any semantic segmentation task, due to its simplicity and effectiveness. We use this as baseline for this task too. The U-Net architecture used is almost a replica of the original net. Two minor adaptions are made: (1) batch normalisation is included and (2) we only have 5 steps in the encoder the last of which has 512 filters.

3.2 Proposed Model

Our proposed model is guided by the following: (1) the principle that less manual annotations (clicks) is better, (2) that the performance of current segmentation models is of decent quality and (3) that corrections should take less time than clicking on each object for manual segmentation. We used the insights gained [Mahadevan *et al.*, 2018] and [**?**; Benenson *et al.*, 2019].

Overall Model As opposed these works however, we do not have the user provide any initial interaction – no bounding box, extreme points or initial clicks. We have a base network provide the initial segmentation mask, and the use a second network to incorporate the corrections. This is not dissimilar from what [Benenson *et al.*, 2019] propose; the difference being that their base network takes in a bounding box input. To incorporate the output from the base network and the clicks from the user, we concatenate the inputs.

Training Strategy The base network is trained prior using the same data. We then adopt the iterative training strategy initially proposed in [Mahadevan *et al.*, 2018] and further established in [Benenson *et al.*, 2019] to train the correction network, where corrections are iteratively added. This strategy involves the use of multiple rounds of corrections for each image, with the error backpropogated each round. We use the insights from [Benenson *et al.*, 2019] to train with 3 rounds of corrections, with 3 clicks provided in each round. Further clicks in each round or more rounds do improve performance further, but the increase is marginal.

Simulated Click Strategy The simulated clicks are placed at the largest error regions in that order (i.e. largest 3 error regions have clicks each round). This is done by comparing the output, from the previous round, with the groundtruth. The pixels at which the errors are located are grouped together using connected component labelling. The 3 largest of these pixel clusters are selected and the centroid of each used as a click.

Click Encoding As per our results in our initial experimentation (which we have not included here) and the results of [Benenson *et al.*, 2019], the choice of click encoding has little impact at best and at worse degrades performance. As such, in place of Guassians or Euclidian distance encoding for this model, we simply use a binary disk to indicate the click. A 3x3 binary disk is placed on the centre of each click, with the relevant positive or negative encoding.

Loss We optimise using a soft Dice loss.

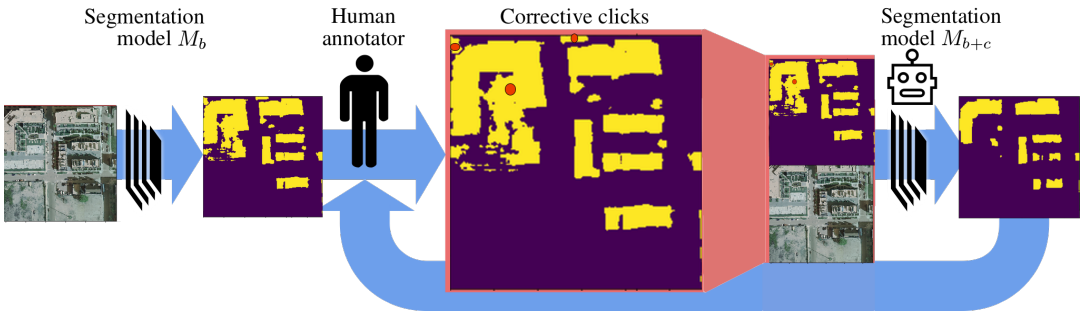

Figure 1: The proposed model: M_b indicates the base network and M_{b+c} indicates the correction network. This examples shows 3 negative clicks being used, but this could be any combination of positive and negative clicks. Adapted from [Benenson *et al.*, 2019].

4 Experiments

We conduct a comparison of the two models on a single dataset to assess the utility of the interactivity. We try to replicate the scenario that a volunteer mapper would likely be placed in, if they are assisted by a computer vision algorithm during mapping. In addition, we specifically design an experiment to investigate out-of-distribution performance.

4.1 Datasets

We perform our experiments on the INRIA Aerial Image Labeling Dataset [Maggiori *et al.*, 2017], which contains pixel-wise semantic labelling of buildings for 5 cities. Each city has 36 5000x5000 pixel tiles. The imagery is RGB aerial imagery with 0.3 meter resolution. For training and testing purposes we generate 100 576x576 pixel patches from each tile.

For our in-distribution experiments, we use random sampling from all fives cities to provide test and validation subsets, with an 80/20 split.

For the out-of-distribution experiments, we use four of the five cities as the training data and the remaining one (Vienna) as a test of out-of-distribution performance. It has been previously noted that performance degrades significantly on this sort of out-of-distribution data, which we also confirm in our results.

Both training and evaluation are done on the patches.

4.2 Training

Each model was trained for 30 epochs with a batch size of 8 patches, with the Adam optimiser.

4.3 Evaluation

The evaluation is the same as the training procedure of our model. The (robot) user provides 3 clicks each round, for a total of three rounds. The method places a click at the centroid of error region, mimicking a user's behaviour under the assumption that the user clicks in the middle of the region of greatest error.

We evaluate performance on the mean Intersection over Union (mIoU) metric.

5 Results

We split the results section into in distribution and out-of-distribution results.

5.1 In-Distribution

Model	Number of Clicks	mIoU
U-Net	0	0.893
Interactive	9	0.921

Table 1: Performance of models on in-distribution validation subset. Interactivity shows a minor improvement quantitatively.

The results show that the interactivity, quantitatively, provides a marginal increase in performance over the base network. And while this improvement is still significant, we can see this translate into more qualitatively pleasing results, as illustrated by figure 2.

5.2 Out-of-Distribution

Model	Number of Clicks	mIoU
U-Net (OOD)	0	0.66
Interactive (OOD) - simple	9	0.76
Interactive (OOD) - normal	9	0.72
Human	330	> 0.95

Table 2: Performance of models on validation subset, where the distribution differs from the training subset. Interactivity shows a significant improvement quantitatively.

Clearly the iterative clicks improve the resulting segmentation masks significantly. In fact, the out-of-distribution performance increase from the clicks exceeds that of the in-distribution. Prior to this result, it was unknown whether interactive models would perform on out-of-distribution data. We show here that it not only works, but provides more useful signal than in the in-distribution setting.

We test a number of configurations involving the base and correction networks. We report on two here. Namely, one where the corrections network is an additional U-Net as described earlier and one where it a U-Net with substantially less filters (we divided the number of filters by 4). We found that both networks yield substantial improvements in mIoU, with the corrections network with less filters performing better in this case (given better hyperparameter optimisation, we think the performance of the corrections network with more filters is likely to improve). But nonetheless it shows that a

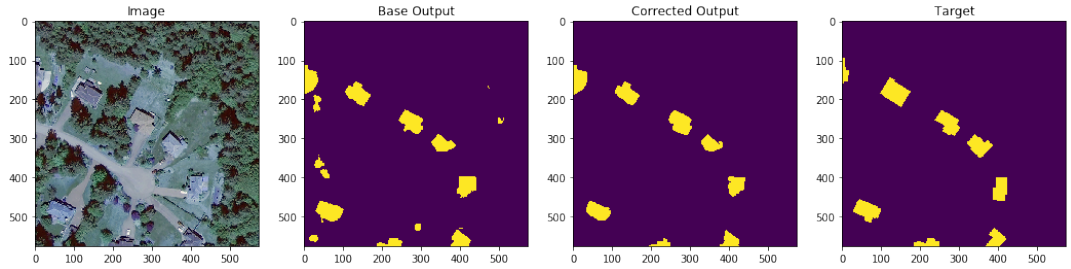

Figure 2: Example showing the performance for in-distribution data. The segmentation masks are smoother and more correct.

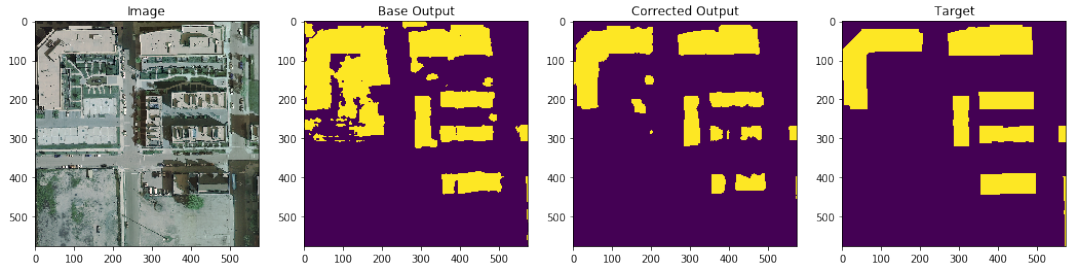

Figure 3: Example showing the performance for out-of-distribution data. A round of clicks shows a significant improvement with the segmentation error in the top left.

simple corrections network, which is substantially less computationally intensive, can provide a meaningful performance increase with minimal additional computation.

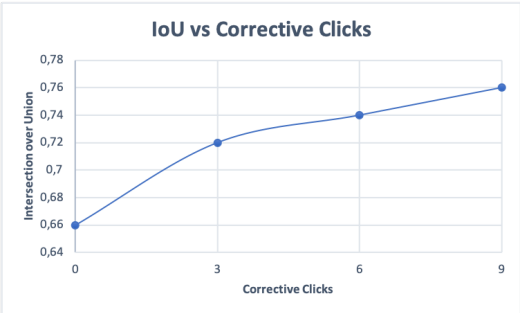

Figure 4: Effects of number of clicks on mIoU per region on average out-of-distribution

Figure 4 shows the effect of number of clicks on mIoU during testing. This model was trained with 3 rounds of 3 clicks. We see that the first round of clicks provides the largest increase in performance, with further rounds providing more smaller increases. This is likely due to the clicks in the first round removing the largest error regions.

Lastly, it is useful to put the number of clicks used in perspective. By counting the number of corners across polygons, we estimate that it would take on average 330 clicks for a human to segment each patch. Our interactive method is able to provide segmentation masks with a relatively high mIoU, with a substantially less clicks (we show the performance at 9 clicks). This would substantially reduce the time required from volunteers to map disaster relief sites.

6 Conclusion

In this paper, we investigated the utility of interactive corrections for semantic segmentation of buildings from aerial imagery, for fast disaster relief mapping, complementing, rather than attempting to automate, existing efforts. We show that the proposed interactive approach achieved an improvement of 15% on out-of-distribution inputs. Furthermore, we show a qualitative improvement in the segmentation masks produced, which make the masks more appropriate for uploading to mapping databases such as OpenStreetMaps.

6.1 Further Work

We believe this work provides a suitable foundation for further development segmentation algorithms that speed up annotation and rapidly adapt to new environments:

Continual and/or Online Learning We have observed from the experiments on out-of-distribution performance that a few points can allow the model to correct itself in a new environments. If the model could use this information to update itself, this would allow the model to adapt to new environments permanently. This would mean that the model gradually improves over as annotations arrive.

Active Learning Certain samples provide more information to the model that the others. To adapt to new environments, an algorithms that could obtain samples from the new data that contain the most information would be tremendously useful. We could then have the user provide the segmentation mask and use these to update the model.

Lastly, a more comprehensive ablation study on the correction network will be performed, with respect to the model and training strategy. The use of CRFs and better segmentation backbones (such as DeepLab) should provide far better absolute results.

Acknowledgements

This work was in part enabled by the support provided by Compute Canada and Calcul Quebec.

References

[Acuna *et al.*, 2018] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient Interactive Annotation of Segmentation Datasets with Polygon-RNN++. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 859–868. IEEE, jun 2018.

[Badrinarayanan *et al.*, 2017] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[Bastani *et al.*, 2018] Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. RoadTracer: Automatic Extraction of Road Networks from Aerial Images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4720–4728. IEEE, jun 2018.

[Benenson *et al.*, 2019] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. mar 2019.

[Boykov and Jolly, 2002] Y.Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 105–112. IEEE Comput. Soc, 2002.

[Castrejon *et al.*, 2017] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating Object Instances with a Polygon-RNN. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4485–4493. IEEE, jul 2017.

[Chen *et al.*, 2018] Liang Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[Dai *et al.*, 2015] Jifeng Dai, Kaiming He, and Jian Sun. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[Dittus *et al.*, 2017] Martin Dittus, Giovanni Quattrone, and Licia Capra. Mass Participation During Emergency Response. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, pages 1290–1303, New York, New York, USA, 2017. ACM Press.

[Iglovikov *et al.*, 2018] Vladimir Iglovikov, Selim Seferbekov, Alexander Buslaev, and Alexey Shvets. TernausNetV2: Fully Convolutional Network for Instance Segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 228–2284. IEEE, jun 2018.

[Jegou *et al.*, 2017] Simon Jegou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1175–1183. IEEE, jul 2017.

[Liew *et al.*, 2017] Junhao Liew, Yunchao Wei, Wei Xiong, Sim-Heng Ong, and Jiashi Feng. Regional Interactive Image Segmentation Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2746–2754. IEEE, oct 2017.

[Lin *et al.*, 2016] Di Lin, Jifeng Dai, and Kaiming He. ScribbleSup : Scribble-Supervised Convolutional Networks for Semantic Segmentation The Chinese Univeristy of Hong Kong. Technical report, 2016.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440. IEEE, jun 2015.

[Maggiori *et al.*, 2017] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229. IEEE, jul 2017.

[Mahadevan *et al.*, 2018] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Iteratively Trained Interactive Segmentation. may 2018.

[Maninis *et al.*, 2018] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool. Deep Extreme Cut: From Extreme Points to Object Segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 616–625. IEEE, jun 2018.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[Rother *et al.*, 2004] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "GrabCut". *ACM Transactions on Graphics*, 23(3):309, aug 2004.

[Wang and Deng, 2018] Mei Wang and Weihong Deng. Deep Visual Domain Adaptation: A Survey. feb 2018.

[Xu *et al.*, 2016] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. Deep Interactive Object Selection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 373–381. IEEE, jun 2016.

[Xu *et al.*, 2017] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. Deep GrabCut for Object Selection. jul 2017.

[Yang *et al.*, 2017] Y. Yang, H. Li, Y. Han, and F. Yu. Research on Method of Interactive Segmentation Based on Remote Sensing Images. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W7:961–964, sep 2017.