Large-scale semantic classification: outcome of the first year of Inria aerial image labeling benchmark

Bohao Huang, Kangkang Lu, Nicolas Audebert, Andrew Khalel, Yuliya Tarabalka, Jordan Malof, Alexandre Boulch, Bertrand Le Saux, Leslie Collins, Kyle Bradbury, et al.

LARGE-SCALE SEMANTIC CLASSIFICATION: OUTCOME OF THE FIRST YEAR OF INRIA AERIAL IMAGE LABELING BENCHMARK

Bohao Huang[1], *Kangkang Lu*[2], *Nicolas Audebert*[3,4], *Andrew Khalel*[5], *Yuliya Tarabalka*[6], *Jordan Malof*[1]
Alexandre Boulch[3], *Bertrand Le Saux*[3], *Leslie Collins*[1], *Kyle Bradbury*[1], *Sébastien Lefèvre*[4], *Motaz El-Saban*[5]

[1] Duke University; [2] NUS; [3] ONERA; [4] Univ. Bretagne-Sud, IRISA; [5] Raisa energy; [6] UCA, Inria
Email: yuliya.tarabalka@inria.fr

ABSTRACT

Over the recent years, there has been an increasing interest in large-scale classification of remote sensing images. In this context, the Inria Aerial Image Labeling Benchmark has been released online in December 2016. In this paper, we discuss the outcomes of the first year of the benchmark contest, which consisted in dense labeling of aerial images into building / not building classes, covering areas of five cities not present in the training set. We present four methods with the highest numerical accuracies, all four being convolutional neural network approaches. It is remarkable that three of these methods use the U-net architecture, which has thus proven to become a new standard in image dense labeling.

Index Terms— Classification benchmark, aerial images, deep learning, convolutional neural networks, U-net.

1. INTRODUCTION

The problem of large-scale semantic labeling is of paramount importance in remote sensing. It consists in the assignment of a thematic label to every image pixel. A large variety of classification methods have been proposed, ranging from the classification of individual pixels to the incorporation of multi-scale spectral-spatial features, in particular automatically learned with convolutional neural networks [1, 2].

One of the current challenges consists in designing methods that generalize to different areas of the earth and can take into account the important intra-class variability encountered over large geographic extents. To evaluate the generalization capabilities of classification techniques, the Inria Aerial Image Labeling (IAIL) Benchmark has been proposed and released online at mid December 2016 [3]. The benchmark images cover varied urban landscapes over two different continents, ranging from dense metropolitan districts (e.g., San Francisco's financial district) to alpine resorts (e.g,. Lienz in Austrian Tyrol). The reference data comprises *building* and *not building* classes. Contrary to all previous datasets, the training and test sets have been split by city, i.e. the classifier performance is evaluated on the set of cities not present in the training set. The test set reference data has not been publicly released, and a contest has been launched online to classify data from five test cities.

Train	Tiles*	Total area	**Test**	Tiles*	Total area
Austin, TX	36	81 km^2	Bellingham, WA	36	81 km^2
Chicago, IL	36	81 km^2	San Francisco, CA	36	81 km^2
Kitsap County, WA	36	81 km^2	Bloomington, IN	36	81 km^2
Vienna, Austria	36	81 km^2	Innsbruck, Austria	36	81 km^2
West Tyrol, Austria	36	81 km^2	East Tyrol, Austria	36	81 km^2
Total	180	405 km^2	Total	180	405 km^2

Table 1: IAIL dataset statistics. *Tile size: 5000^2 px. (0.3 m/pixel).

In this paper, we discuss the outcomes of the first year of the Inria benchmark contest. We first briefly recall the composition of the dataset and give its use statistics. We then present four winning methods with the highest numerical accuracies, all four being convolutional neural network approaches. Finally, we conclude about the state-of-the-art in large-scale semantic labeling.

2. INRIA DATASET AND STATISTICS

The IAIL benchmark dataset[1] is composed of 360 color (3-band RGB) orthorectified images, with the spatial resolution of 30 cm/pixel, with the total coverage of 810 km^2 (every image size is 5000^2 pixels). The images have been acquired during several flight campaigns over different urban areas of the Unites States and Austria. The reference data was created by rasterizing shapefiles of public domain official building footprints, and is composed of two classes: *building* and *not building*. An example of a close-up from the IAIL dataset with the corresponding reference data is shown in Fig. 2.

Table 1 summarizes the regions included in the dataset and their distribution into training and test sets. The amount of training and test data is the same. The split was done in such a way that both sets contain landscapes from the United States and Europe, as well as both high-density (e.g., Vienna/Innsbruck) and low-density (e.g., Kitsap/Bloomington) settlements. Only training reference data has been publicly released. More information about the IAIL dataset can be found in [3].

[1] project.inria.fr/aerialimagelabeling

During the first year after the IAIL benchmark release, the dataset has been downloaded more than 800 times, by researchers from all continents, with the approximately equal distribution from public and private institutions. The 16 submissions with the classification results on the test set have been received and automatically evaluated, using two performance measures:

• *Intersection over union* (IoU) of the building class, i.e., the number of pixels labeled as building in both the prediction and the reference, divided by the number of pixels labeled as pixel in the prediction or the reference.

• *Accuracy*, i.e., the percentage of correctly classified pixels.

These measures are computed for each of the regions individually (e.g., Innsbruck, San Francisco) and also for the overall test set. The following section presents four approaches which achieve the highest performance in terms of these criteria.

3. METHODOLOGIES

3.1. U-net with novel training/test strategy (Applied Machine Learning Lab AMLL at Duke University)

We used the original U-net architecture from [4], with a single major modification; we used half as many filters at each layer (see Fig. 1). Therefore, for example, we used 32 filters instead of 64 in the first-level convolutional layers, 64 filters instead of 128 filters in the second-level layers, etc. When training the U-net, we used the standard input and output patch sizes of 572×572 and 388×388, respectively. However, as we will explain, it was beneficial to modify these sizes during label inference.

The design of the IAIL dataset, in which no training data is available for all five testing cities, makes overfitting much more likely compared to other benchmark problems in remote sensing segmentation. We do not yet have experimental evidence, we hypothesize that the reduced number of filters in our U-net model (i.e., reduced learning capacity) reduced our risk of overfitting to the training data, even after optimizing it for good performance on our validation data.

3.1.1. Network training

Pre-processing was comprised of padding all of the image tiles symmetrically with 92 pixels on each side, and then computing a global mean (i.e., a single RGB value) which was subtracted from each pixel in the imagery dataset.

We created two datasets with the available labeled image tiles from each city: a training dataset comprised of tiles 6 through 36 from each city, and a validation dataset comprised of the remaining tiles. Within the training tiles, we extracted 572×572 patches on a uniform grid, with 92 pixels of overlap between neighboring patches. We pooled together the patches extracted from each training tile to create a training dataset

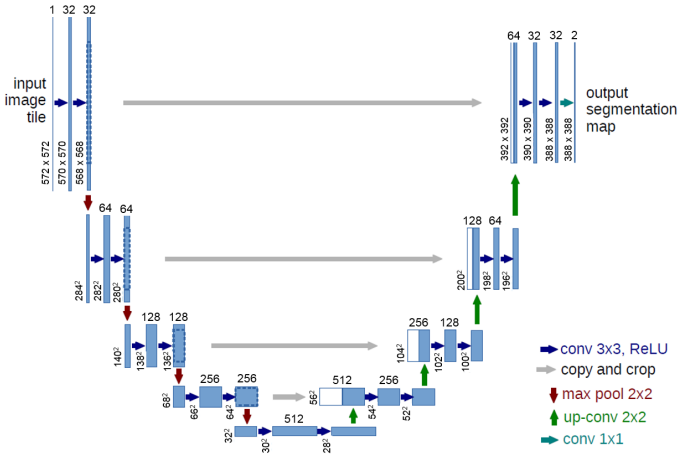

Fig. 1: U-net architecture designed by AMML.

of patches. We found that extracting training patches on a uniform grid, with relatively little overlap between patches, yielded networks with much better performance than, for example, sampling patches from random locations. We provide comprehensive experimental support for this training strategy in [5].

During network training we used a standard stochastic gradient descent with a cross-entropy objective function. We used the Adam optimizer with an initial learning rate of $1e{-}3$, and a momentum of 0.9. We trained our network for 100 epochs, where each epoch consisted of 8000 mini-batches, and the learning rate was reduced to $1e-4$ after 60 epochs.

Minibatches consisted of 5 training patches, which were drawn randomly from our patch training dataset. Online augmentation was applied randomly to the input patches, including vertical/horizontal flips and 0/90/180/270 degrees rotations.

3.1.2. Label inference

We observed that the U-Net model produces relatively poor predictions at the edge of its output. In order to mitigate this problem, we increased the input size of the U-Net to 2636×2636 during label inference, which was the maximum size that could be supported by our 1080 Ti GPU. When compared to using the original input patch size, we found that this approach improved our performance, while also reducing our overall inference time. We provide comprehensive experimental support for this inference strategy in [6].

3.2. Dual-resolution U-net (NUS)

We proposed a dual-resolution U-net [4] architecture, along with soft Jaccard loss. To fit large images to GPU memory, we have to cut them into patches. However, these patches will bring artifacts along boundary region during testing. To solve this issue, we use a pair of dual-resolution images as input.

In detail, we crop one high-resolution 384×384 patch from the original image. Then we crop a 768×768 patch with the same center and downsample it to a twice lower resolution (384×384) image. Features from both high and low resolution patches are extracted by U-Net, then score maps for each resolution are computed. A weight map is further learnt to merge score maps from different resolutions. This weight map determines, for each pixel, how much the network relies on different resolution inputs. To summarize, the final result is a weighted sum of dual-resolution score maps.

By using the proposed dual-resolution architecture, one can better train and predict large or along-patch-boundary buildings, and prevent from artifacts when merging patches.

Besides that, the employed loss function is a combination of sigmoid cross-entropy ($sigmCE$) and a soft Jaccard loss introduced in [7]. Jaccard, or IoU, index is commonly used as an evaluation metric for segmentation tasks, and intuitively, it would be useful to compute a loss with Jaccard index. To make it differentiable, we use prediction scores and corresponding ground truth to compute a soft Jaccard index ($l_{soft-IOU}$), as defined in [7], and use it as part of the loss function. Our loss function is computed as:

$$L_{NUS} = L_{sigmCE} - \log l_{soft-IOU}. \quad (1)$$

During training, we extracted 384×384 patches from images and use vertical/horizontal flips for data augmentation. We used the Adam optimization algorithm, with a base learning rate of $1e-3$, a momentum of 0.9 and "poly" learning rate policy. We firstly finetuned channel numbers of the original U-net to better fit our dataset (channels of the modified U-net are: 32, 64, 128, 128, 256, 128, 128, 64, 32). Then, we integrated all the modules of the proposed architecture and trained our network from scratch for 30 epochs.

3.3. Signed distance transform regression (ONERA)

In this method, we use a standard fully convolutional network: SegNet [8], which we adapt to include spatial context in the optimization process. Indeed, the standard semantic segmentation loss function is the averaged classification error - the cross-entropy, which is computed on all pixels regardless of their location. To constrain the network to learn for each pixel the spatial dependencies from its neighbours, we add a regularization loss computed on the Euclidean signed distance transform (SDT) [9].

For each ground truth mask, we compute its Euclidean SDT. Therefore, we obtain a continuous representation of the ground truth, that assigns to each pixel its distance to the nearest building border. We also slightly alter the network so that it has two outputs: the standard classification and a regression of the SDT. This can be seen either as a multi-task framework or an additional regularization term in the form of an $L1$ penalty on the inferred distances. Assuming that $Z_{seg}, Z_{dist}, Y_{seg}, Y_{dist}$ respectively denote the output of the segmentation softmax, the regressed distance, the ground truth segmentation labels and the ground truth distances, the final loss to be minimized is:

$$L_{ONERA} = NLLLoss(Z_{seg}, Y_{seg}) + \lambda L1(Z_{dist}, Y_{dist}), \quad (2)$$

where $NLLLoss$ is a negative log-likelihood loss function, and λ is an hyper-parameter that controls the strength of the regularization.

This significantly improves the network predictions. Indeed, the inferred maps present now a better connectivity and generally smoother borders. This stems from the regularization using the SDT regression, that constrains the network to learn the class of each pixel, but also how far it is located from the building edge. It therefore reduces the influence of ambiguous spectrometry and reinforces the impact of the spatial context.

Our SegNet is trained on 384×384 patches with stochastic gradient descent, using pre-trained VGG-16 weights for the encoder and random He initialization [10] for the decoder.

3.4. Stacked U-Nets (Raisa Energy)

We started by dividing the original tiles into smaller patches, which are fed as training data to our model. Our model is based on the U-Net architecture [4]. The U-Net architecture consists of a contracting path followed by an expanding path. This setup empowers the model with accurate localization capabilities so it can be utilized for precise segmentation tasks. Instead of using a single U-Net, our model uses a stack of two U-Nets arranged end-to-end. The second network works as post-processor for the previous one to enhance its predictions.

Since IoU and accuracy are the performance evaluation metrics, we used a loss function that combines both binary cross entropy and a differential form of Intersection-over-Union (IoU) [7] to focus on our objective. In addtion, data augmentation in the form of basic rotations and reflections is used at both training and inference times, as it led to a more robust model with better results on the validation set.

The output of our model is a 2D dense activation map representing scores for individual pixels whether it contains a building or not. These activation maps are concatenated together to reconstruct the whole tile. We used reflections at the the tile edges to obtain better prediction at the most outer patches. These reflections remove the discontinuity effect of the edges.

4. EXPERIMENTS

Both numerical and visual classification comparisons for all IAIL benchmark submissions can be found on `https://project.inria.fr/aerialimagelabeling/leaderboard/`. Table 2 summarizes the numerical results for the four approaches described in the previous section, and compares them with the results published in the original benchmark paper [3]. Fig. 2 shows a sample of classification

Table 2: Numerical evaluation on test set.

Method	Bellingham		Bloomington		Innsbruck		San Francisco		East Tyrol		Overall	
	IoU	Acc.	IoU	Acc.	IoU	Acc.	IoU	Acc.	IoU	Acc.	IoU	Acc.
AMLL	67.14	96.64	65.43	96.73	72.27	96.66	**75.72**	**91.80**	74.67	97.70	**72.55**	**95.91**
NUS	**70.74**	**97.00**	66.06	96.74	**73.17**	**96.75**	73.57	91.19	**76.06**	**97.81**	72.45	95.90
ONERA	68.92	96.94	**68.12**	**97.00**	71.87	96.72	71.17	89.74	74.75	97.78	71.02	95.63
Raisa	68.73	96.79	60.83	96.23	70.07	96.31	70.64	89.52	74.76	97.64	69.57	95.30
Inria [3]	56.11	95.37	50.40	95.27	61.03	95.37	61.38	87.00	62.51	96.61	59.31	93.93

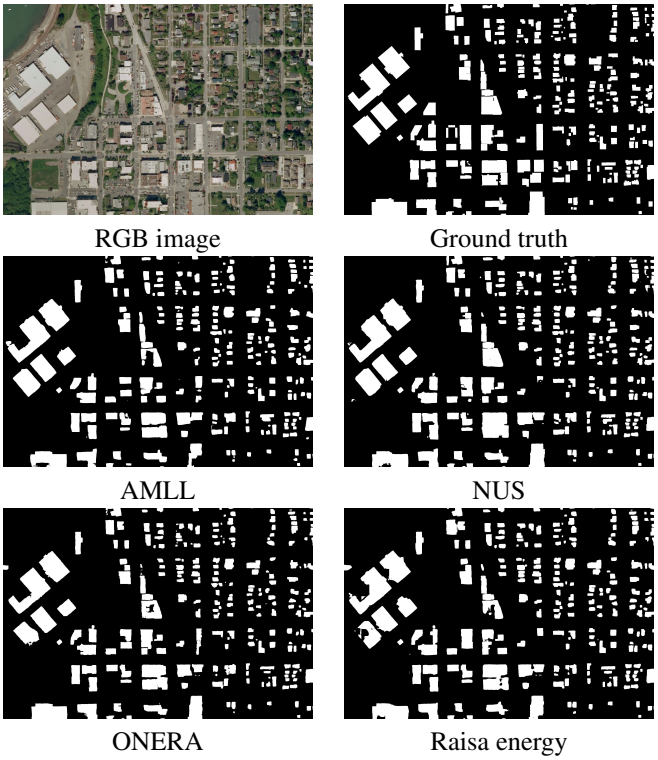

Fig. 2: Sample of classification results.

maps. The methods proposed by the AMLL lab and the NUS, both based on the U-net architecture, yielded the highest and comparable accuracies. The AMLL method performed particularly well on the dense urban areas, while the NUS architecture yielded the best performance on the less populated areas.

5. CONCLUDING REMARKS

From the outcomes of the first year of the IAIL benchmark contest, we can conclude on the following:

• The U-net architecture has shown the highest performance and has thus proven to be well suited for image dense labeling.

• A loss function must be carefully designed. It has been proven that combining both the averaged classification error and either a differential form of IoU, or an SDT-based regularization loss improves segmentation performance.

• A good choice of training/inference strategies boosts classification results, yielding the winning performances.

6. REFERENCES

[1] Michele Volpi and Devis Tuia, "Dense semantic labeling of sub-decimeter resolution images with convolutional neural networks," *IEEE TGRS*, vol. 55/2, 2017.

[2] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution aerial image labeling with convolutional neural networks," *IEEE TGRS*, vol. 55/12, 2017.

[3] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *IGARSS*, 2017.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[5] B. Huang et al., "Sampling training images from a uniform grid improves the performance and learning speed of deep convolutional segmentation networks on large aerial imagery," in *IGARSS*, 2018.

[6] B. Huang et al., "Increase the input image size of convolutional segmentation networks during label inference to improve their performance and speed on large aerial imagery," in *IGARSS*, 2018.

[7] Gellert Mattyus, Wenjie Luo, and Raquel Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in *ICCV*, Oct 2017.

[8] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation," *IEEE TPAMI*, vol. 39/12, 2017.

[9] Q. Z. Ye, "The signed Euclidean distance transform and its applications," in *ICPR*, 1988.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE CVPR*, 2015, pp. 1026–1034.