SEMCITY TOULOUSE: A BENCHMARK FOR BUILDING INSTANCE SEGMENTATION IN SATELLITE IMAGES

R. Roscher[1,2,*], M. Volpi[3], C. Mallet[4], L. Drees[1], J. D. Wegner[5]

[1] Institute of Geodesy and Geoinformation, University of Bonn, Germany - (ribana.roscher, ldrees)@uni-bonn.de
[2] Institute of Computer Science, University of Osnabrueck, Germany
[3] Swiss Data Science Center, ETHZ and EPF Lausanne, Switzerland - michele.volpi@sdsc.ethz.ch
[4] Univ. Gustave Eiffel, IGN-ENSG, LaSTIG/Strudel, Saint-Mande, France - clement.mallet@ign.fr
[5] EcoVision Lab, ETH Zurich, Switzerland - jan.wegner@geod.baug.ethz.ch

Commission II

KEY WORDS: Benchmark, machine learning, instance segmentation, buildings

ABSTRACT:

In order to reach the goal of reliably solving Earth monitoring tasks, automated and efficient machine learning methods are necessary for large-scale scene analysis and interpretation. A typical bottleneck of supervised learning approaches is the availability of accurate (manually) labeled training data, which is particularly important to train state-of-the-art (deep) learning methods. We present `SemCity Toulouse`, a publicly available, very high resolution, multi-spectral benchmark data set for training and evaluation of sophisticated machine learning models. The benchmark acts as test bed for single building instance segmentation which has been rarely considered before in densely built urban areas. Additional information is provided in the form of a multi-class semantic segmentation annotation covering the same area plus an adjacent area 3 times larger. The data set addresses interested researchers from various communities such as photogrammetry and remote sensing, but also computer vision and machine learning.

1. INTRODUCTION

The automatic interpretation of aerial and satellite images is of considerable research interest in the areas of remote sensing, machine learning (ML) and computer vision. One of the most active areas is the development of efficient and scalable learning algorithms. The "big data" regime is now dictated by the exponential increase in data availability, with frequent improvements on spatial, spectral and temporal resolutions, and Earth observation (EO) missions lasting longer and guaranteeing data continuity.

For the purpose of fostering research and operational practices, the community at large proposed various benchmarks (Bakuła et al., 2019), which deal with different tasks such as land cover classification (Demir et al., 2018), road detection (Mnih, 2013), instance segmentation (Waqas Zamir et al., 2019), and object recognition (Humanity & Inclusion, 2018, Van Etten et al., 2018).

Challenges arise when the amount of labeled information which is necessary to build a well generalizing interpretation model is limited or not representative enough of operational scenarios. In general, it is common to have access to a vast amount of unlabeled data, but only a tiny fraction of it is accurately labeled. This setting drives the ML practitioners in EO applications to perform research in the areas of active learning, weakly supervised or few shot learning, transfer learning, and self-taught learning (Wurm et al., 2019, Bettge et al., 2017, Crawford et al., 2013). On a parallel line, lots of efforts are devoted to methods which are able to efficiently use large amounts of labeled data, in particular since the success of deep learning, which is avid in terms of annotations. Still, deep learning has now become the go-to technique for most image processing tasks, a success also translated to the EO community, where the adaptation of deep learning techniques from the computer vision community became one of the most active research areas (Ma et al., 2019). However, unlabeled data contains invaluable information, but its extraction is far from trivial and requires dedicated methods often dealing poorly with unbalanced data and rare classes (Oliver et al., 2018).

The development and analysis of methods able to deal efficiently with large-scale data sets in EO is still connected to many open research questions such as imbalanced classes, non-uniform atmospheric effects, and varying spatial and spectral resolutions. Furthermore, most recent methods are developed independently to EO data peculiarities: A proper exploitation of geographical and spatial invariance, the use of object sizes in a consistent metric space, or known land cover reflectances are mostly ignored. These aspects can be used to regularize the learning process and, to some extent, improve the sample complexity and cope with rare and unbalanced classes.

One obstacle to analyze, optimize, and evaluate suitable modern machine learning methods is the missing availability of public benchmark data sets and the lacking diversity of implemented tasks on the same data. Moreover, various shortcomings can occur: While many benchmark data sets address realistic questions and already refer to large areas, some of them lack in accuracy or do not reflect space-borne settings (e.g., multispectral imagery), which would not favour scientific developments over engineering efforts. The following list covers some recent benchmarks related to ours:

crowdAI Mapping Challenge. This benchmark challenge deals with building instance segmentation (Humanity & Inclusion, 2018). The training set is composed of more

*Corresponding author

Figure 1. 1 out of 16 tiles of the `SemCity Toulouse` data set (876 px × 863 px, area covered ≈3 km^2). Illustrated are the satellite image (near infrared-green-blue), semantic segmentation annotation, and building instance annotation (from left to right).

than 200.000 300 px × 300 px tiles, but only in RGB channels. The areas covered are mainly residential areas, where buildings are homogeneous in size, background discriminative and instances clearly separated one from each other.

SpaceNet. Several Space Net data sets have been released and used for benchmark challenges. The first and second challenges aim at the detection of building footprints from Worldview-II and Worldview-III high-resolution 8-band images. The building footprint annotations are initially derived semi-automatically and subsequently improved manually. Buildings are marked individually, where each street address is defined as building. However, building corners are only accurate to 5 px.

DeepGlobe. This data set contains three challenges: road detection, building detection, land cover classification (Demir et al., 2018). The building detection data set is based on the SpaceNet data set, and contains 24.586 images of size 650 px × 650 px from the WorldView-III sensor with a spatial resolution of 31 cm in the panchromatic band and 1.24 m in all other bands. Adjoining buildings are marked as single buildings, and thus the data set is more detailed than previous building data set. Since the building footprint is annotated, the appearance of the building in the image deviates from the annotation, depending on the height of the buildings and the viewing angle. The annotations are, however, not consistently performed throughout the data set.

INRIA Aerial Image Labeling data set. The semantic segmentation data set contains building/non-building annotations, where training and test set are split over cities in Europe and America (Maggiori et al., 2017). The RGB- or RGB-Infrared imagery has a spatial resolution of 10 cm-30 cm, covering a total area of 810 km^2. The benchmark does not distinguish between building instances, such that adjoining buildings are marked as one segment.

ISPRS Vaihingen and Potsdam. The multi-class semantic segmentation data sets contain 4-band RGB-Infrared imagery with a high spatial resolution of 9 cm and 5 cm, respectively (Rottensteiner et al., 2013). Vaihingen covers an area of 1.5 km^2 with small detached buildings, and Potsdam covers an area of 3.5 km^2 comprising the historic city and dense settlements. Both data sets are smaller than the other mentioned data sets and do not contain building instances.

Open Cities AI Challenge. This semantic segmentation data set consists of RGB images with varying spatial resolution, in which building footprints in 10 cities across Africa are annotated (GFDRR, 2020). The quality varies significantly across the data sets (marked as tier 1 and tier 2) with some annotations having a displacement of several meters.

iSAID . This data set is a multi-class instance segmentation benchmark (Waqas Zamir et al., 2019). It contains 2.806 high-resolution aerial images taken from the DOTA (Xia et al., 2018) data set with over 655.000 annotated instances from 15 classses. This benchmark covers many classes that are not of global interest (e.g., baseball courts), instances are clearly identifiable, and the class "building" is not part of the benchmark.

Benchmarks are an invaluable contribution to the community and pushed many boundaries in terms of research. The more tasks and data characteristics are covered, the more ML methods can be compared and more limitations can be identified. The above-mentioned data sets and the many others not mentioned here, enable an objective and comprehensive comparison of different ML approaches and a thoughtful development of novel methods. Some of the data sets were offered in the form of challenges and have thus led to further progress in the development of new, more efficient methods (Bakuła et al., 2019, Rottensteiner et al., 2014). The provision of winning solutions as open-source software promotes further progress, as it was done for example in the SpaceNet Challenge 1[1]. We advocate that experiments should be released as open source code to foster reproducibility and development.

In this paper, we introduce the `SemCity Toulouse` benchmark, which aims at complementing existing benchmarks, and to foster the development of machine learning methods which address the aforementioned open research questions. The benchmark focuses on building instance segmentation, where additional information is provided in the form of a multi-class semantic segmentation annotation covering the same area plus an adjacent area 3 times larger. Although this is a traditional task, our benchmark focuses on a densely populated area, the city of Toulouse (France), showing nicely organized residential areas and old town structures. Each single building instance is independently annotated, so no large blocks of multiple housing are depicted in the training set. This issue is of particular interest when cadastral archives need to be updated or, in turn, when

[1] `https://spacenetchallenge.github.io/Challenges/Challenge-1.html`

those are used to generate training data. Our data set comes directly from the Worldview-II sensor, providing 8 spectral channels and therefore going beyond the use of standard RGB imagery and opening the doors to multi-spectral deep learning.

2. DATA SETS

In this section, we discuss the study site, imagery characteristics, the annotated sets, and the annotation process which was conducted to derive the labels.

2.1 Study site

Our study site is a $50\,\mathrm{km}^2$ area covering the greater city center of Toulouse, France. Toulouse is the capital of the Occitanie region and located in the southwest of France in the département Haute-Garonne. It has over 480.000 inhabitants and is the fourth largest city in France. Our study city covers the river Garonne and neighborhoods including the old city, several residential areas, sparsely built areas, and work/industrial districts.

2.2 Imagery

The original data set contains a $16\,\mathrm{bit}$ 8-band Worldview-II satellite image from April 2011 with a ground sample distance of $50\,\mathrm{cm}$ for the panchromatic band and $2\,\mathrm{m}$ for the other bands. The $2\,\mathrm{m}$-resolution image has a spatial extent of $3643\,\mathrm{px} \times 3560\,\mathrm{px}$. In the data set, we provide the geotiff image in original resolution and a pansharpened image produced with the band-dependent spatial detail algorithm (Garzelli et al., 2007)[2], resulting in a spatial resolution of $50\,\mathrm{cm}$.

We split the image into 16 tiles of equal size, as illustrated in Fig. 2, where one tile has a size of $3504\,\mathrm{px} \times 3452\,\mathrm{px}$ in the panchromatic band and a size of $876\,\mathrm{px} \times 863\,\mathrm{px}$ in all other bands, covering an area of about $3\,\mathrm{km}^2$. In addition to the image data, two annotation sets are provided for each tile in the resolution of $0.5\,\mathrm{m}$. The high resolution panchromatic images, the pansharpened images, and additional tools like Google StreetView and OpenStreetMap[3] were used as a basis for visual inspection. In the following, we introduce the building instance segmentation task in more detail.

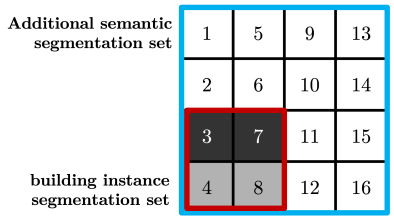

Figure 2. Splitting of training (light gray) and test tiles (dark gray) for the building instance segmentation task. Further areas with semantic segmentation annotations (white) are provided.

2.3 Single Building Instance Segmentation

The annotation are given as indexed geotiff images with consecutive indices for each instance. The set covers the tiles 3, 4, 7, and 8 and contains building instances, where an individual building is defined as an object with a street number. In this benchmark, buildings are annotated as they can be seen in the image rather than building footprints, to guarantee a precise match between annotation and input imagery. The number of building per tile is given in Tab.1.

Table 1. Number of building instances per tile.

Tile	# buildings
3	1.273
4	1.755
7	3.464
8	2.963
Total	9.455

Since building instances are not uniquely defined across different benchmarks, the following list provides more details about the annotation process.

Types of buildings We include all types of residential buildings, shops, office buildings, department stores, discount stores, shopping centers, as well as buildings and halls of industrial in the data set. However, a few large building structures mainly from industrial districts which cannot certainly be divided into single building instances are excluded. Also, in the case of dense building structures, especially in the old city centre, building structures cannot be certainly assigned to an instance, for example due to the limited resolution, and excluded from the instance data set. Overall, about 5% of all buildings are excluded from evaluation. In case they can be generally identified as building, they are included in the semantic segmentation annotations. Three examples of included building types can be seen in Figure 3.

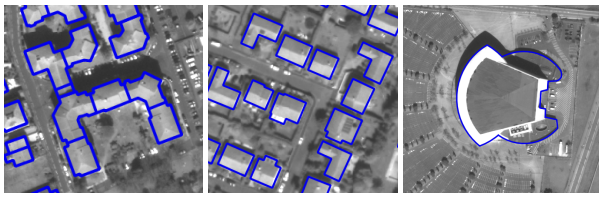

Figure 3. Various kinds of buildings included in the benchmark: residential buildings, detached houses, and Zénith Toulouse Métropole (from left to right).

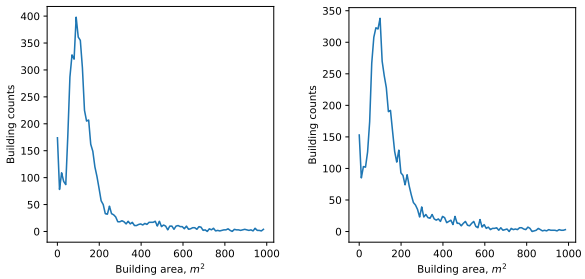

Figure 4. Distribution of buildings given their surface, in m^2, for the training and test sets. Values larger than $1000\,\mathrm{m}^2$ have been clipped out.

The distribution of buildings regarding their size in m^2 is given in Fig. 4.

Shadows and occlusions We exclude buildings affected by the following two cases: First, buildings which are mostly covered by other objects such as trees, and second, buildings

[2] https://www.pansharp.com/applications/panfusion/
[3] openstreetmap.org.

which are covered largely by the shade of a neighbouring house. The latter case occurs in the vicinity of tall houses or terraced houses that lie on a north-south axis and are of different heights. Due to the unfavorable position of the sun in the South, entire buildings in these rows are often covered with shadow. However, as long as the shape of the house and the separation of the instances is recognizable, these houses are not excluded. Examples are illustrated in Fig. 5.

Figure 5. Buildings that are clearly in the shadow of other objects and whose outlines or separation are not visible to the annotator (indicated in red) are not included in the instance segmentation data set and excluded for evaluation. Buildings which can be identified though, for example, covered in shadow are included to foster models which are robust to shadows.

Small independent man-made objects Small independent man-made objects such as free-standing garages are not defined as building instances. If garages connect directly to a residential building, they form a common instance with this building. Examples are illustrated in Fig. 6.

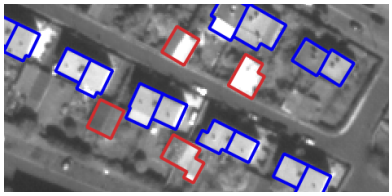

Figure 6. Small independent man-made objects (indicated in red) are not defined as building in the instance segmentation data set and marked as void. Therefore, they are not used for evaluation.

Prefab housing estate and apartment buildings Long building blocks such as prefab housing estates and apartment buildings are divided into building instances if identifiable with Google Maps. The division is based on the assignment of house numbers and visually visible borders in the image. In case that the building cannot be divided into separate instances and at the same time we have evidence that it has not only one house number, the instance is excluded from the annotated set. Examples are illustrated in Fig. 7.

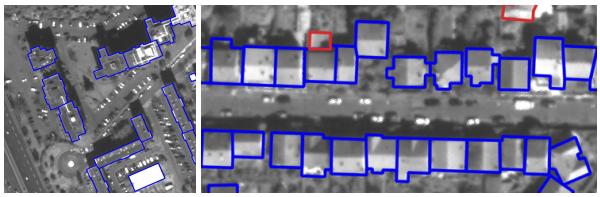

Figure 7. Apartment buildings and their instance annotation provided in the data set (indicated in blue). Red marked objects are not identifiable and excluded from evaluation.

2.4 Semantic Segmentation

Each of the 16 image tiles is paired with a manually annotated land cover map, which can be used as ancillary data for building instance segmentation, e.g., by learning multiple tasks on the same data. The annotation is an indexed tiff image with the following eight classes:

- impervious surface: street, pavement, concrete;
- building: all kinds of buildings and building complexes;
- pervious surface: grass, low vegetation, soil;
- high vegetation: trees;
- car: vehicles of all kinds;
- water: natural water bodies, swimming pools;
- sport venues: all kind of event and sport venues such as horse race tracks, tennis courts;
- void: all pixels, which could not be assigned to any of the above mentioned classes.

The total amount of annotated pixels without the class void is about 12.5 Mio., where the class-wise percentage of labels is given in Tab. 2.

Table 2. Number of labeled samples (in percentage) in the semantic segmentation ancillary data.

class	# labeled samples
impervious surface	23%
building	23%
pervious surface	30%
high vegetation	16%
car	2%
water	3%
sport venues	3%

The annotations are provided pixel-wise, where all pixels are assigned to a class `void` if they do not belong to any of the considered classes or if they cannot be assigned with a high certainty. Examples image patches are given in Fig. 8.

Figure 8. Sample panchromatic image patches with annotation (color coded as described in Tab. 2)

Please note, since the annotation was performed by photo-interpretation, small errors are unavoidable.

2.5 Intention behind the provided data sets

Tasks which can be considered with the provided data sets will revolve around the transfer of knowledge between scenes of different characteristics (e.g., old town and suburbs) and learning with limited labeled data. Special focus should be laid on scalable algorithms which can deployed to large-scale data processing.

In detail, we encourage the development and evaluation of approaches which use transfer learning between EO specific models for common tasks such as instance segmentation. Here, a relevant research question is how models, learned on other benchmark data sets, perform on our data set. The differences between the benchmark data sets can be exploited to answer

questions about the influence of i) a different number of spectral channels, ii) different spatial resolutions, and iii) different geographical scales and locations. Furthermore, most of these questions can also be addressed by a multi-task approach, in which building instance segmentation is supported by land cover semantic segmentation.

We further encourage the development and evaluation of methods which use both labeled and unlabeled data, or various kinds of labeled data. The latter one can be approached by using OpenStreetMap data or ancillary data, which we provide with multi-class semantic labels. A promising direction is multi-task learning, where different kinds of tasks such as instance segmentation and semantic segmentation are solved jointly and both tasks regularize each other, resulting in a higher generalization ability and a higher robustness of the learned models. Another promising direction is the pre-training of models which use labels, which are easier to acquire then instances. That means, we encourage the development of approaches which use different levels of label information.

3. STATE-OF-THE-ART AND INSTANCE SEGMENTATION BASELINES

3.1 Evaluation

`SemCity Toulouse` is designed for building instance segmentation, where individual building masks are compared to reference building masks.

For the evaluation of the building instance segmentation, we follow the evaluation metrics of the COCO data set (Lin et al., 2014) [4]. One essential measure is the intersection over union (IoU) score defined as:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}\,, \tag{1}$$

where TP are the true positives, FP the false positives, and FN are the false negatives. Moreover, we evaluate the result by means of precision p and recall r scores, which are defined as:

$$p = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad r = \frac{\text{TP}}{\text{TP} + \text{FN}}\,. \tag{2}$$

We summarize precision and recall scores by means of the F1 score for the class "buildings":

$$\text{F1} = \frac{2 \cdot p \cdot r}{p + r}. \tag{3}$$

Precision and recall are computed with respect to the building instances. We also report average precision (AP), average recall (AR) and average F1 (AF1), which are calculated over multiple IoU values as for the COCO benchmark. Specifically, $(\cdot)@0.5 - 0.95$ means the average of measure $(\cdot)$ over several IoU thresholds in the interval from 0.5 to 0.95 with step size 0.05. As standard for COCO, we use P@.5, P@.75 and AP@.5-.95; but we also use recall R@.5, R@.75 and AR@.5-.95 and F1@.5, F1@.75 and AF1@.5-.95. We report these metrics for two different, complementary settings. The former relates to the *detection* of an instance: a counter is simply increased when the predicted object mask IoU $\geq t$, where t is the IoU threshold with the reference mask. In the second setting, we retrieve objects as detected based on the mask IoU, but we then accumulate, for each object, the number of correctly segmented pixels. This way, we can recompute the metrics above but taking into account objects size. Obviously, the two correlate as IoU implicitly measures the segmentation accuracy, but, particularly for lower thresholds, we can appreciate how the model is correct in detecting reference instance boundaries.

[4] `http://cocodataset.org/#detection-eval`

3.2 Experimental setup

To create a first and realistic baseline, we use the benchmark in combination with the standard instance segmentation method Mask R-CNN (He et al., 2017). We use this method as it is widely accepted as a standard benchmark, and it provides the baseline metrics to which all consecutive developments should compare to. We use the `Matterport` implementation, available at `https://github.com/matterport/Mask_RCNN`. Before training, we re-tile original images into 512×512 patches. For training, we use a stride of 256, so every image overlaps with a direct neighbor 50%. We also use data augmentation, wrapped by the `imgaug` package in `Matterport`, available at `https://github.com/aleju/imgaug`. We use vertical and horizontal flipping with a probability of 0.5, and one rotation among 7 main directions. We also scale channel values by multiplying them by a factor sampled uniformly in $[0.9, 1.1]$.

We train on single GPU machine NVIDIA Tesla P100, as provided by Swiss Data Science Center on demand infrastructure. We adjust the architecture to accommodate 8 channels as inputs and the final branches to predict only one instance class. We select the remaining hyperparameters by searching for better F1 score by using a random subset of the tiled training data as a validation set (then merged to the training set for the final model). In this case, only non-negative maximum suppression thresholds and detection thresholds are tuned. We fine-tune for 100 epochs, each going over a random subset of 500 image patches, with a learning rate of 0.001 and balanced losses. The training is performed using Tensorflow's Stochastic Gradient Descent implementation (via the keras interface). The first 25 epochs are dedicated into training the first and the last layers, while the remaining 75 to fine-tune the whole model. The last 25 epochs are run with a learning rate reduced by 10. We use ResNet-50 as backbone, fine tuning weights transferred from COCO instance segmentation, and a batch size is 3. This baseline will provide a crude, first evaluation to be used for further comparisons.

3.3 Results

We show in Fig. 9 the different metrics at the predefined IoU levels. In panel (a), we show the precision at different IoU levels. It is clear that at lower IoU levels, the predicted instance masks cover less well the reference, as indicated by a lower precision. Notably, as reported in Tab. 3, at IoU ≥ 0.5 84% of the predictions are correct, which is good for a comparatively simple baseline. In panel (b), we show recall with the same IoU thresholds. This time, only roughly more than 60% of the object have been correctly detected, which in turn points at the fact that many instances present in the reference are missed. As for precision, the pixel-level evaluation shows that, in general, the number of correctly segmented pixels from the reference could be improved. Panel (c) shows a combination of the scores, the F1 (see Eq. 3). The object-level score at lower IoU is high, but decreases quickly as the IoU threshold

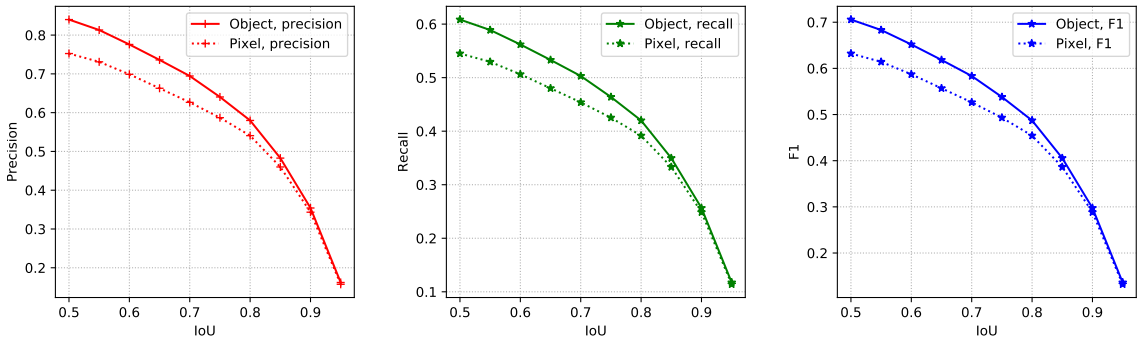

Figure 9. Panel (a) shows evaluations at different IoU scores for the precision, for both the object-level evaluation (solid line) and the pixel-based evaluation (dashed line). Panel (b) shows the recall of the model, with the same style as panel (a). Panel (c) shows the same for the F1 score.

	Global metrics			Precision		Recall		F1	
Evaluation	**AR**	**AP**	**AF1**	**P@0.5**	**P@0.75**	**R@0.5**	**R@0.75**	**F1@0.5**	**F1@0.75**
Object	0.440	0.608	0.511	0.840	0.640	0.608	0.464	0.706	0.538
Pixel	0.403	0.556	0.467	0.752	0.587	0.545	0.425	0.632	0.493

Table 3. Global evaluation metrics. AR, AP and AF1 are computed averaging values obtained in IoU 0.5 to 0.95 (cf. x-axis in Fig. 9).

level is increased, meaning that, also considering precision and recall, instance segmentation can be improved.

In Fig. 10, we show predictions of 8 example patches, summarizing situations encountered for this data set with Mask R-CNN. We also show some visual details of these results in Fig. 11. The first 4 panels in Fig. 10(a)-(d) and Fig. 11(a)-(d) show examples, where our method achieves good results. Over uniform background, uniform instance size – corresponding to the "peak" shown in Fig. 4 – and fairly homogeneous color, Mask R-CNN is able to detect most instances well enough. Note also that in these situations, instances are composed by single housing, so each instance mostly corresponds to an independent building. In panel Fig. 10(c), in the bottom middle, there is an elongated structure corresponding to several instances. In this case, the model recognizes correctly the overall structure, but overpredicts each single instance as no visual cues are available to separate them. On an opposite situation, for panel Fig. 10(d), upper part of the image – with detail in Fig. 11(a) – the model is able to correctly separate many instances belonging to 4 large building blocks, but obviously this time, visual cues are pointing at the right direction.

The last 4 panels of Fig. 10 show examples for incorrect predictions. Panels Fig. 10(e),(f) show situations in which visual cues and homogeneity in color and size are detrimental to the detection. In (e), many instances belonging to attached housing blocks are missed completely, while in Fig. 10(f), the instances are detected correctly (at IoU$\geq$ 0.5) but the correct shape is missed. This is probably due to the size of objects, which are relatively rare occurrences in the training set. Panel Fig. 10(g) show similar issues with large buildings, this time however the elongated structures are undersegmented. Some more examples of this are shown in Fig. 11(e)-(g), where the object is either very ambiguous, such as a car park (Fig. 11(e)), unusual shapes (Fig. 11(f)) or ambiguous in terms of number of instances, making the model predictions incorrect, as in Fig. 11(g) and (h).

We assume a more sophisticated, more targeted training can help to overcome this challenge. Panel (h) shows good detections illustrated in panels (a)-(d), but with clear false positives. The model confuses building with parking lots, often visually similar to large industrial complexes, with cluttered rooftops.

Table 3 shows a summary of the global metrics illustrated in Fig. 9. Overall, the baseline performs fairly well on simple "exemplar" situations, in which instances are spatially separated, homogeneous in size and in color and when a single instance correspond to a single building. Models start to struggle when instance size or appearance is rare, which is a common feature for industrial areas and densely build areas, particularly in old towns, where a building appears different from all the others. This can explain the fairly good scores at lower IoU, since at this level many instances are detected, but not accurately segmented. Obviously, as the IoU gets stricter, segmentation are also good, as IoU is a measure of quality in segmentation, and therefore pixel-level scores become more aligned to the object-level metric. Still, at high IoU, the quality of the detection needs improvement as many instances are missed (low recall) or and many masks are imprecise, since the IoU threshold excludes predictions (low precision).

We trained and evaluated a baseline method for instance segmentation, which is however, not satisfactory at a global level. With the provision of the data set, we leave the application of more sophisticated methods open to research. We argue that in order to solve problems related to complex structures and to multiple instances visually belonging to the same structure, different approaches are needed. We believe that some of these issues might be alleviated by smart data augmentation and explicit size modeling, but some of the issues that will be encountered in this benchmark will be only solved by the use of geographical regularities and spatial pattern, or by an additional usage of ancillary information available in the web. Furthermore, it is possible to see how high IoU scores correspond to low accuracy metrics. In general, instances are detected well macroscopically, but more sophisticated methods such as informed machine learning approaches exploiting geometrical constraints or spatial post-processing might significantly improve results.

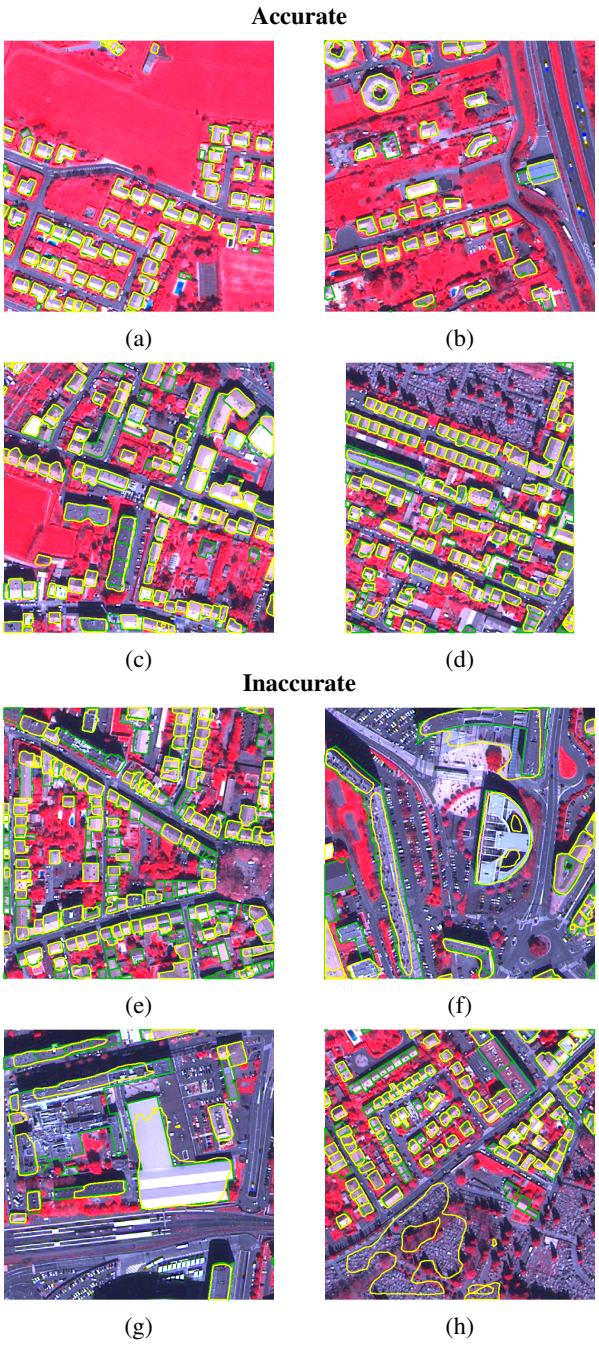

Figure 10. Some example predictions for the tiled test images for the baseline model. In yellow are shown predictions, while in green the reference samples. Note that only predictions with IoU ≥ 0.5 are shown.

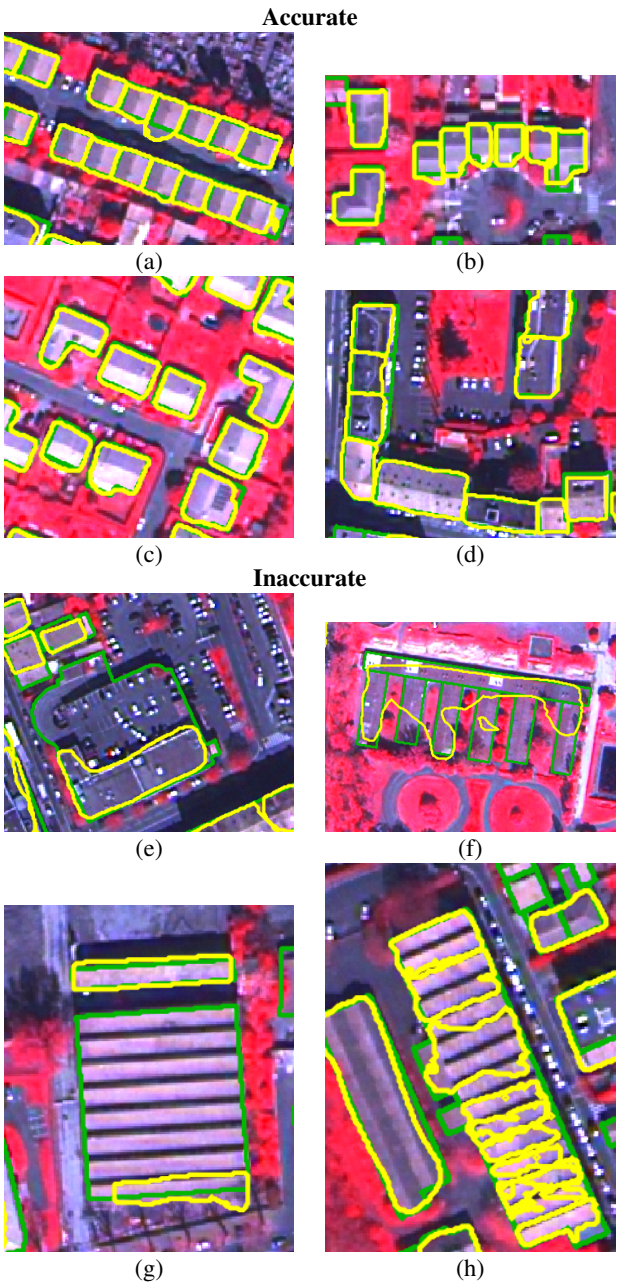

Figure 11. Some details of predictions. In yellow are shown predictions, while in green the reference samples. Note that only predictions with IoU ≥ 0.5 are shown.

4. CONCLUSIONS AND FUTURE PLANS

We introduced the `SemCity Toulouse` benchmark, a satellite-based data set for very high spatial resolution building instance segmentation. Beside building instances, we provide ancillary data in form of a multi-class land cover classification map, covering four times the area of the instances annotation map. This should give opportunity to take advantage of different types and levels of labels, if the number of actual labels for the intended task is limited. The data sets are available under `http://rs.ipb.uni-bonn.de/data/`.

We also discussed pros and cons of the baselines obtained by Mask R-CNN, which is a standard for instance segmentation tasks. In addition, we make suggestions on how the baseline solution can be improved. We hope that our project will enable and foster the evaluation and development of machine learning methods and bridges the gap between different research communities.

With the publication of this paper we will first provide the instance building segmentation and the semantic segmentation for the same area. Due to cross-checking by several annotators and minor improvements to ensure quality, the rest of the semantic segmentation will be published this fall. Due to this, the numbers in Tab. 2 can slightly change. Our further plans are to provide an additional point in time for change detection and to extend the benchmark by another city with different characteristics.

ACKNOWLEDGMENTS

We thank the International Society for Photogrammetry and Remote Sensing for funding the scientific initiative 'ISPRS Benchmark Challenge on Large Scale Classification of VHR Geospatial Data', in which this benchmark was created. We like to acknowledge the annotators who ensure the quality of the data set not only by manual labeling but also by independent cross-checking. Finally, we would like to thank the Swiss Data Science Center for offering a GPU to train the benchmark models.

REFERENCES

Bakuła, K., Mills, J., Remondino, F., 2019. A Review of Benchmarking in Photogrammetry and Remote Sensing. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, XLII-1-W2, 18.

Bettge, A., Roscher, R., Wenzel, S., 2017. Deep self-taught learning for remote sensing image classification. *ESA Big Data from Space*.

Crawford, M. M., Tuia, D., Yang, H. L., 2013. Active learning: Any value for classification of remotely sensed data? *Proceedings of the IEEE*, 101(3), 593–608.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raska, R., 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 172–17209.

Garzelli, A., Nencini, F., Capobianco, L., 2007. Optimal MMSE pan sharpening of very high resolution multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(1), 228–236.

GFDRR, 2020. Open cities ai challenge: Segmenting buildings for disaster resilience. `https://www.drivendata.org/competitions/60/building-segmentation-disaster-resilience/`. Accessed: 2020-01-23.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

Humanity & Inclusion, 2018. crowdAI Mapping Challenge - Building Missing Maps with Machine Learning. `https://www.crowdai.org/challenges/mapping-challenge`. Accessed: 2020-01-23.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft coco: Common objects in context. *European Conference on Computer Vision*, 740–755.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B., 2019. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152, 166–177.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 3226–3229.

Mnih, V., 2013. Machine Learning for Aerial Image Labeling. PhD thesis, University of Toronto, Canada.

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., Goodfellow, I., 2018. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in Neural Information Processing Systems*, 3235–3246.

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J. D., 2013. ISPRS Test Project on Urban Classification and 3D Building Reconstruction. , ISPRS - Commission III - Photogrammetric Computer Vision and Image Analysis, Working Group III / 4 - 3D Scene Analysis.

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J. D., Breitkopf, U., Jung, J., 2014. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, 256–271.

Van Etten, A., Lindenbaum, D., Bacastow, T. M., 2018. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*.

Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao, L., Xia, G.-S., Bai, X., 2019. isaid: A large-scale dataset for instance segmentation in aerial images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 28–37.

Wurm, M., Stark, T., Zhu, X. X., Weigand, M., Taubenböck, H., 2019. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150, 59–69.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2018. Dota: A large-scale dataset for object detection in aerial images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3974–3983.