Urban Building Classification (UBC) – A Dataset for Individual Building Detection and Classification from Satellite Imagery

Xingliang Huang[1,2]*, Libo Ren[1,2]*, Chenglong Liu[1,2], Yixuan Wang[3,5], Hongfeng Yu[1,2]
Michael Schmitt[5], Ronny Hänsch[4], Xian Sun[1,2]†, Hai Huang[5]†, Helmut Mayer[5]
[1]Aerospace Information Research Institute, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China [3]Technische Universität München, Germany
[4]German Aerospace Center, Germany [5]Universität der Bundeswehr München, Germany
{huangxingliang20, renlibo20, liuchenglong20}@mails.ucas.ac.cn,
ge35qej@mytum.de, hfyu@mail.ie.ac.cn, ronny.haensch@dlr.de, sunxian@aircas.ac.cn,
{michael.schmitt, hai.huang, helmut.mayer}@unibw.de

Abstract

We present a dataset for building detection and classification from very high-resolution satellite imagery with the focus on object-level interpretation of individual buildings. It is meant to provide not only a flexible test platform for object detection algorithms but also a solid basis for the comparison of city morphologies and the investigation of urban planning. In most current open datasets, buildings are treated either as a class of landcover in the form of masks or as simple objects defined by separate contours (footprints). Our dataset, instead, represents individual buildings using in-depth object-level descriptions concerning geometry as well as functionality. Buildings are treated as objects with individual ID and boundary. Adjacent building blocks are also separated according to house numbers making a subsequent high-level classification of individual buildings possible. The buildings are classified into predefined roof types, such as flat, gable and hipped roof as well as functional purposes, i.e., residential, commercial, industrial, public, and their sub-classes, e.g., single-family house, office building and school. In the first version of the dataset we provide selected urban areas from two cities: Beijing in China and Munich in Germany. It, therefore, (1) allows to verify algorithms that are not only valid for specific regions but also work robustly in spite of the diversity of cities on different continents with various land forms and styles of architecture and at the same time (2) provides the possibility to quantitatively compare the statistics and morphology of different cities. It is planned to extend the dataset by a continuous integration of various urban areas worldwide.

*Equal contribution.
†Corresponding author.

1. Introduction

Buildings are one of the most important components of urban areas. The investigation of buildings plays an essential role in urban planning, city administration, emergency management, tourism, etc. With the advent of deep learning techniques, the performance of building detection and classification in remote sensing data has been significantly improved. One key driver are the ever increasing remote sensing datasets [14], of which the building-related datasets are summarized in Section 2.

We propose a novel dataset with a specific focus on the object-level interpretation of individual buildings, which are represented with in-depth descriptions concerning both geometry as well as functionality. Based on this, the dataset provides the possibility to quantitatively compare different cities with regard to statistics and morphology. As shown in Figure 1, Beijing's buildings (rows 1 and 2) typically show a neat arrangement and a modern steel/concrete style, while buildings in Munich (rows 3 and 4) are mostly distributed along the historical streets and are of lower height.

Although we can readily identify geometrical information of buildings, e.g., contour and roof shape, in satellite imagery, it is usually substantially difficult to accurately identify the function of buildings. So we label the function using additional map information.

Data sources for cities such as OSM (OpenStreetMap) and Google Maps provide the basis for a large number of statistics on buildings worldwide. But different data sources have different definitions for building attributes, so it is difficult to combine multiple data sources to automate the labeling of buildings for remote sensing images. Moreover, many building attributes are missing in these data sources and they do not provide up-to-date,

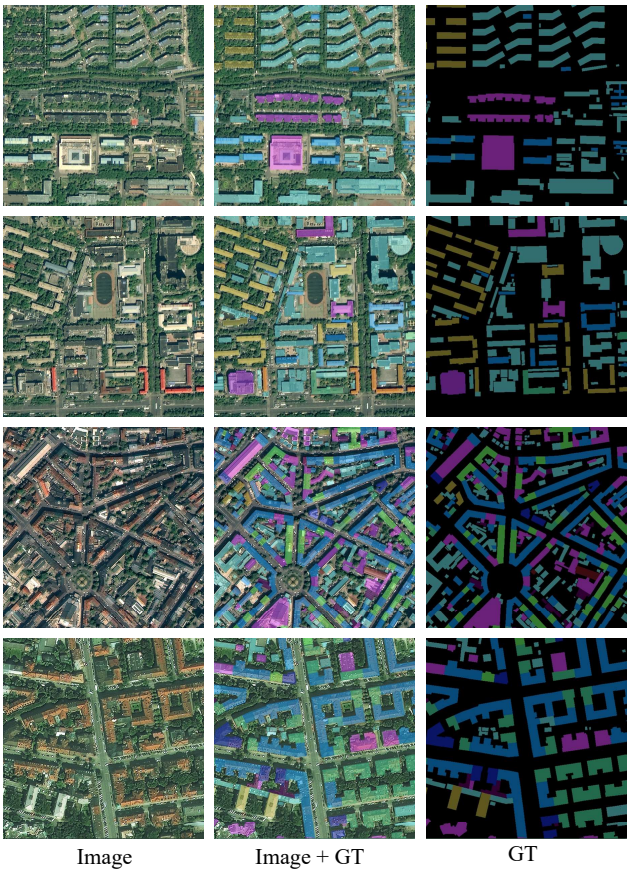


Figure 1. Example UBC data of Beijing (rows 1 and 2) and Munich (rows 3 and 4) with input images (left), ground truth (GT, right), and overlaid views (center). Colors of ground truth indicate various classes of roof type.

complete and accurate statistics of building locations and attributes. Further information on urban buildings can be provided by high resolution and low-cost remote sensing images. To learn large-scale statistics of urban buildings in remote sensing images with deep learning techniques, a building dataset with consistent attribute standards and global diversity is urgently needed.

Particularly, the main contributions of our work are:

- **In-depth annotation:** Instead of visual annotations, in this dataset boundaries as well as functions of individual buildings are labeled according to OSM and Google Maps. This means that adjacent buildings in a building block as for instance shown in Figure 4 can be correctly separated into different objects. The functions of buildings can often be more accurately derived from their attributes in the maps than from images.
- **Fine-grained categories:** We provide novel fine-grained categories concerning (1) building geometry with 25 roof types as well as (2) two levels of functional purposes with the five main classes: residential, commercial, industrial, public and other, and 36 subclasses, e.g., single-family house, office building and school.
- **Flexible Structure:** This dataset consists of explicitly separated subsets corresponding to different cities. The subsets can be combined to train general detectors and to test their robustness or employed separately to investigate and compare characteristics of different urban areas. The categories are given on various levels of both roof type and function. They can be used to define different setups for competition with varying amounts of data. Please note that it is planned to extend this dataset. We expect that along with the increasing size of the urban areas and the extended coverage of different classes, the advantage of our multi-level category definition will become even more apparent.

2. Related Work

Suitable datasets are critical for the development and evaluation of object detection and classification algorithms, especially deep neural network models. An increasing number of remote sensing datasets has been introduced in recent years with various data sources as well as target objects. The DOTA dataset [8] contains over 1.7 million instances of 18 classes with oriented bounding box annotations collected from 11,268 aerial images. It has, thus, greatly contributed to the development of detection algorithms for rotated objects in remote sensing data. The FAIR1M dataset [17] also consists of over one million instances of fine-grained objects in high-resolution remote sensing imagery, providing the community data with 5 categories and 37 sub-categories of ground targets. In the ISPRS Urban Modelling and Semantic Labeling Benchmark [13] multispectral imagery and airborne laserscanner data of Vaihingen and Potsdam, Germany, as well as Toronto, Canada, are meant for the detection of urban objects, such as buildings, roads, trees, as well as for 3D building reconstruction. The TorontoCity dataset [21] provides aerial imagery with about 10 cm ground resolution depicting around 400 thousands buildings. SpaceNet consists of a series of remote sensing datasets with various basic data including multispectral imagery and synthetic aperture radar (SAR) data and purposes such as building and road network extraction as well as classification. The SpaceNet 2 Challenge [20] contains 302,701 building footprints in 24,586 scenes, while SpaceNet 6 [16] is a multi-sensor all-weather mapping dataset, consisting of both optical and SAR imagery, aiming to map building footprints using multi-modal data. SpaceNet 7 Multi-Temporal Urban Development Challenge [19] is based on

Dataset	Classes	Instance Quantity	Modality	Resolution
SpaceNet 2 [20]	1	500k	RGB MSI	0.3 m
SpaceNet 6 [16]	1	4.8k	RGB SAR	0.5 m
Toronto City [21]	1	400k	RGB	0.05-0.1 m
GaoFen-3 Building [24]	1	(semantic segmentation)	RGB SAR	1 m
INRIA [11]	2	(semantic segmentation)	RGB	0.1-0.3 m
DSTL [15]	5	2k	RGB MSI	0.3 m
SemCity Toulouse [12]	6	9k	PAN	0.5 m
UBC	61	41k	RGB	0.5-0.8 m

Table 1. Comparison of building datasets

imagery collected by Planet Labs' Dove Satellites and contains around 500,000 buildings tracked over time.

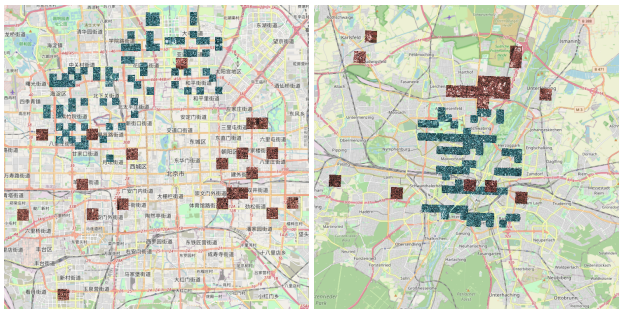
The INRIA aerial image labeling benchmark [11] consists of precisely registered cadastral records as well as 15 cm or 30 cm orthorectified imagery. It considers the two classes building and non-building, i.e., trees and roads. The DSTL Satellite Imagery Feature Detection dataset [15] provides multispectral satellite imagery in RGB as well as 16-bands with a resolution of 0.3 m. It employs a coarse classification of buildings, including residential and non-residential building, fuel storage facility, and fortified building. It comprises about two thousand building instances. The SemCity Toulouse benchmark [12] focuses on building instance segmentation. It provides multi-class semantic segmentation annotation including residential and office building, shop, department store, discount store, shopping center, as well as industrial building.

Related datasets include also the GaoFen-3 SAR dataset [24] for semantic segmentation of buildings. It is acquired in spotlight (SL) mode with high-resolution (1 m) and a wide swath (10 km) and covers urban as well as rural areas in, e.g., Hongkong, Berlin, and Shanghai. A comprehensive comparison of our dataset with selected remote sensing datasets containing buildings is given in Table 1.

3. Dataset

3.1. Satellite Imagery

The SuperView (or "GaoJing" in Chinese) satellites are commercial very high-resolution earth observation spacecrafts operated by Beijing Space View Tech Co Ltd. They are equipped with sensors collecting both panchromatic (0.5 m, Ground Sampling Distance – GSD) and multispectral (2 m) GSD imagery with a maximum scene size of 60 km $\times$ 70 km [5]. The Gaofen-2 high-resolution imaging satellites from the China National Space Administration (CNSA) are capable of collecting images with a GSD of 0.81 m in the panchromatic and 3.24 m in the multispectral bands with a swath width of 45 km [4]. Here, we chose panchromatic and multispectral data from SuperView and Gaofen-2 for urban areas of Beijing and Munich and obtained 4-band images (red, green, blue and near-infrared) at 0.5 m and 0.8 m by pan-sharpening. In the current version of the dataset only the three visible bands, i.e., RGB, are used. Possible multispectral as well as SAR data are scheduled for the further extension of the dataset (cf. Section 5). The whole dataset consists of 800 tiles with 600 $\times$ 600 pixels and 200 pixels overlap for adjacent tiles. The information about data coverage and instances is shown in Figure 2 as well as Table 2.



(a) Beijing (b) Munich

Figure 2. Data coverage in Beijing and Munich: Tiles from Superview (blue) and Gaofen-2 (red)

City	Source	Coverage (km^2)	Building Instances
Beijing	SuperView	20	10,105
	Gaofen-2	12.8	4,824
Munich	SuperView	20	19,352
	Gaofen-2	12.8	7305

Table 2. Data coverage and building instances for Beijing and Munich

3.2. Definition of Classes

Current datasets focus on the location, footprint extraction and segmentation of buildings. Classification, especially fine-grained classification of buildings is rare yet. The latter is of great interest for applications in city planning and

urban development analysis. We are aware that the definition of classes has a huge influence on the performance of the classification. To derive plausible classes which conform to common understanding as well as to what can be seen from remote sensing data, we first summarize the categories of the main popular data sources and standards: OpenStreetMap [23], CityGML (partially open) as well as Google Maps and derive our own classes taking the characteristics and limits of satellite imagery into account. One obvious advantage is that the existing geometrical and functional attributes of buildings in the above mentioned data sources can be easily mapped to our classes. This makes the current manual annotation easier and will allow for a (semi-) automatic annotation (cf. Section 5) in the future.

The definition the roof classes is given in Table 3. For the roof type, we use 25 fine-grained classes (including "other") based on the geometry of the roof. The fine classes are grouped into nine coarse classes based on geometrical similarities. The coarse classes are especially useful when the instances of the fine-grained classes are not enough for a stable training.

Coarse	Fine-grained
flat	flat roof
	flat roof HVAC*
	flat roof complex
shed	shed roof
gable	gable roof
	gable roof asymm*
	gambrel roof
	butterfly roof*
row	row roof shed*
	row roof gable*
	row roof arched*
hipped	multiple eave roof*
	hipped roof v1
	hipped roof v2
	half hipped roof*
	mansard roof
	pinnacle roof
arched	arched roof
	half arched roof*
revolved	dome*
	cone*
	cupola*
freeshape	freeshape surface*
	freeshape poly*
other	other

Table 3. Roof type classes. * indicates the fine-grained classes with few instances, which are merged into the class "other" in the following experiments

For the building functions, as shown in Table 4, we define five coarse classes, i.e., "residential", "commercial", "industrial", "public" and "other" which can be split into 36 fine-grained classes. Also inside each coarse class there is a fine-grained class named "other", e.g., "public other". In contrast to the "other" in the coarse classes, it is employed to label the instances which are not listed in the fine-grained classes, but still can be determined as belonging to one of the coarse classes. This happens quite often as the function classes hardly cover all possibilities.

Coarse	Fine-grained
residential	single-family house, multi-family house, row house, apartment high, apartment block, villa, garage, residential other
commercial	office building, retail and mall, hotel, parking house, restaurant, commercial other
industrial	power plant, warehouse, manufacturing, water treatment, industrial other
public	administration, gas station, education, stadium, sports hall, transportation, theatre, fire station, police station, military, church, mosque, temple, airport building, hangar, public other
other	other

Table 4. Building functionality classes

3.3. Annotation

The footprint of buildings are annotated with polygons. The footprints from OSM are used as basis and are manually refined/corrected according to the input images. If available, the roof type and function information is taken from OSM as well as Google Maps and is mapped to the predefined classes. Figure 4 shows one example.

We are aware of the heterogeneous quality of OSM [6] and noticed that the availability and quality of OSM data for Munich is substantially better than for Beijing. For the selected areas in relation to the corresponding final ground truth, the footprints of 89.7% of the buildings have been provided in the OSM dataset. 39.9% of the buildings in Munich have the function attribute and 22.2% roof type information. Yet, in Beijing, only for 27.6% of the building footprints have been included and not all of them are correctly located. Only 4.2% of the buildings have function information and there is no roof type information in the OSM data

for Beijing. A summary is given in Figure 3. Therefore, we have referred to information from Google Maps to manually improve the annotation of roof types and functions. I.e., the annotators visually check the roof shape as well as (heterogeneous) labels of buildings and manually interpret their roof types and functions according to the predefined categories in UBC. For difficult instances, particularly the buildings with multiple labels (cf. Section 3.4), additional Google Street View data (for Munich only) are optionally employed to assistant the (visual) interpretation. To ensure the quality of annotation, the results from annotators are examined by more experienced inspectors in two rounds: One complete check and one random check. Controversial labels are determined by consensus of multiple annotators and inspectors. The consistency of annotations are also ensured for the buildings in overlapping areas. As shown in Table 2, the UBC dataset provides altogether 41,586 building instances, including 14,929 for Beijing and 26,657 for Munich, with complete footprints and annotations for both the roof type as well as the function. Additionally, 4,790 for Munich and 210 for Beijing building instances have multi-label annotations.

3.4. Multi-label Annotation

In a realistic urban scenario, individual buildings often do not have a single function. For example, for many tall buildings and large structures in the city centers, the lower floors are typically shopping areas, while the upper floors serve as office space or for habitation. It is, therefore, not reasonable to label these buildings with just one function class. To make the annotation of the dataset more accurate, we, thus, introduce multi-label annotation. Yet, in order not to add too much extra complexity to the annotation process, we restrict the multi-label annotation to only two reasonable situations: (1) "Apartment block" as well as "commercial other" and (2) "apartment high", "office building" as well as "retail and mall". For the latter, multiple choice selection of up to three labels at the same time is allowed.

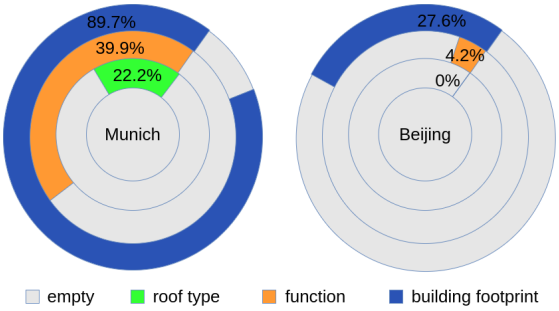


Figure 3. Comparison of the availability of OSM building footprints and attributes for Munich and Beijing. The missing ("empty") data are supplemented by manual annotation.

3.5. Dataset Splits

Our dataset is designed to be used as a whole set as well as two separate subsets: Beijing and Munich. Each subset contains the same amount of data: 400 tiles of satellite images with a size of 600×600 pixels selected from their urban areas. Please note that we also ensured that the data partitions of SuperView (80%) and Gaofen-2 (20%) on each subset is constant, so that this ratio is also kept in the whole set. In the experiments, the dataset is divided (on a random sampling basis) into training, validation as well as test sets with partitions of 7:2:1.

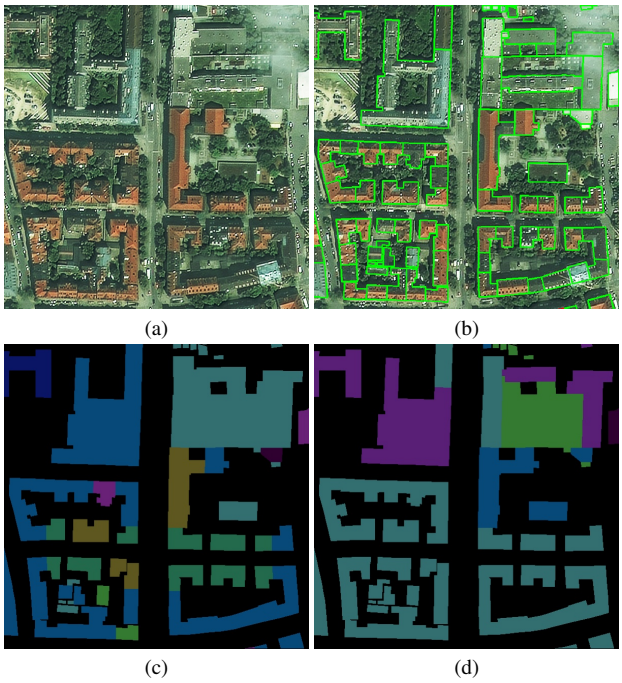


Figure 4. An annotation example: Input image (a), building footprints (b, green polygons), roof types (c) and functions (d, coarse classes)

4. Experiments

High building densities in urban areas, various sizes, fine-grained classes as well as multi-label annotations pose great challenges to instance segmentation methods on the UBC dataset. To quantitatively measure the state-of-the-art under these circumstances, we propose two instance segmentation tasks as well as corresponding evaluation metrics and we evaluate general methods on the dataset. In the first task, we predict roof types for all buildings and use pixel-level masks to localize them. Fine- and coarse-grained roof labels are set as two sub-tasks . The class of each instance in the second task is defined by the building function. Due to the difficulty of reasoning about the function from visual features alone, we only conducted baseline experiments on

coarse function classes.

4.1. Baseline Models

We selected Mask-RCNN [9], Cascade Mask RCNN [1], SOLOv2 [22] and QueryInst [7] as our baseline models. The backbone network for all models is ResNet-50-FPN. The implementation of these methods is based on the MMDetection library [2] [18]. Specifically, Mask-RCNN and Cascade Mask RCNN are classic two-stage models. SOLOv2 is a novel effective single-stage approach. QueryInst is an end-to-end query-based framework that achieves state-of-the-art performance on the COCO dataset.

4.2. Evaluation Metrics

For evaluation, we use the standard COCO metrics [10]: AP_{mask} (averaged over IoU threshold), AP_{50}, AP_{75}, AP_S, AP_M and AP_L. Due to the presence of buildings in high-density areas but also rather small buildings, we adjusted parts of the metrics. The area of objects referred to by S, M, and L is redefined for small as area less than 400 pixels, for medium as area 400-1600 pixels, and for large as area 1600 pixels and above.

To train the CNN-based models, we employ the dataset splits given in Section 3.5 and initialize the network with ResNet50 pre-trained on ImageNet [3]. All models are trained on 4 GPUs for 100 epochs with 4 tiles per GPU. The learning rate is initialized with 0.02 and then reduced by a factor of 0.1 at epochs 60 and 90. Multi-scale training and random flipping are used as data augmentation during training. Instances with multiple function labels are correspondingly employed multiple times for training and testing of the different classes. The rest of the hyper-parameters of the model are set to the same values as in the original MMDetection [2] setup.

4.3. Experimental Setups

4.3.1 Baselines with Fine-grained Roof Type

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN [9]	13.4	23.1	14.7	**3.8**	17.2	17.2
C Mask R-CNN [1]	**15.3**	**25.2**	**17.5**	3.4	19.7	**19.3**
SOLOv2 [22]	13.5	23.1	15.3	3.0	**20.4**	15.8
QueryInst [7]	13.1	21.9	14.9	3.7	18.4	16.4

Table 5. Instance segmentation results using $mask_{AP}$ on UBC test set with fine-grained roof categories. "C Mask R-CNN" denotes Cascade Mask R-CNN [1].

As the geometry type of the roof is reasonably determinable from visual features, we employed the fine-grained classes of roof types to compare the current state-of-the-art models. Table 5 shows the average precision (AP) results also with different intersection over union (IoU) thresholds for the instance segmentation algorithms presented in Section 4.1. The results in Table 5 show that the Cascade Mask R-CNN model outperforms the other models concerning average precision with an IoU of at least 0.5 (AP_{50}) and the mean average precision (mAP) calculated from all different IoUs. The cascade architecture has shown its advantage in the detection of densely distributed buildings with varying size by means of the multi-stage refinement of the IoU threshold. Following the basic concept of "segmenting objects by location" fitting the distribution of building instances, SOLOv2 demonstrated a competitive performance. It, however, also showed its limit for irregularly distributed and objects with varying scale because of the fixed size grid cells. QueryInst performs better for COCO than on the UBC dataset, probably because our dataset mainly covers urban areas with building instances, for which Queryinst doesn't seem to be suitable. The average precision for small buildings (smaller than 10m$\times$10m) AP_s is quite low for all models, showing that the detection of small buildings is a challenge for recent models.

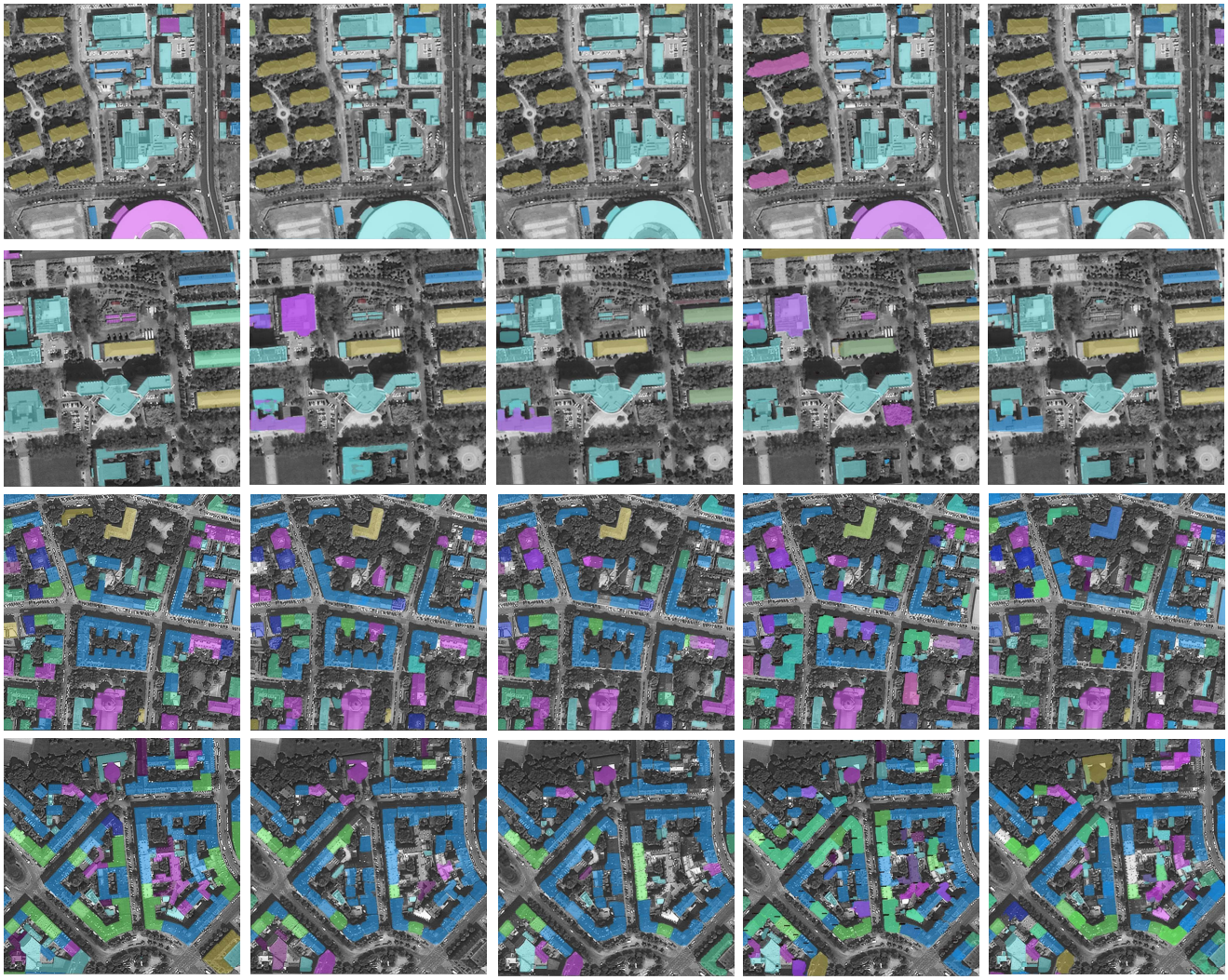
Class-wise results for different roof types are shown in Table 6. Some classes, e.g., flat and hipped roof, contain a large number of building instances and have obvious features for classification. Thus, they could be better detected and classified by the models. Other classes such as arched, flat complex and shed roofs consist of small numbers of instances and also do not have distinctive features, adding to the difficulty and challenge of detection and classification. This implies the necessity to use few-shot learning or feature augmentation to improve the instance segmentation results. Figure 5 shows segmentation results using different models. The Cascade Mask R-CNN model has the best performance for building detection. Precise segmentation of individual buildings and roof type classification are challenging in complex scenes.

4.3.2 Baselines with Coarse-grained Function

Compared with roof type classification, it is difficult to determine the function of buildings from the visual features in just a single image tile, especially in urban areas. In this paper, therefore, only the results with coarse-grained classes of building functions are demonstrated (Table 7). Different from the experimental results for roof type, the Mask R-CNN model achieves the best performance with respect to mAP, while the SOLOv2 model performs best concerning AP_{50}. However, one has to admit that the overall performance of all models is relatively poor. On one hand, it is hard to classify the function based only on the visual geometry features. On the other hand, since the size of each image tile is small, it is often not possible to determine the relationships between nearby buildings with different functions and, thus, discover the complex information necessary to discriminate the various functional areas of a city.

Method	AP	AP_{50}	FL	FC	SH	GA	GM	H1	H2	MA	PI	AR	OT
Mask R-CNN [9]	13.4	23.1	26.0	2.4	**3.9**	21.1	13.5	14.8	40.1	7.8	3.0	**7.3**	**7.1**
Cascade Mask R-CNN [1]	**15.3**	**25.2**	**27.3**	2.3	2.8	**21.5**	13.4	**16.3**	**40.2**	10.9	**20.8**	6.6	6.7
SOLOv2 [22]	13.5	23.1	26.1	3.3	2.9	20.5	14.1	15.2	33.9	10.4	10.4	6.5	5.9
QueryInst [7]	13.1	21.9	21.5	**5.8**	2.1	18.4	**14.6**	15.6	30.4	**17.8**	8.6	5.5	4.7

Table 6. Class-wise instance segmentation results on UBC test set with fine-grained roof classes: FL-flat, FC-flat complex, SH-shed, GA-gable, GM-gambrel, H1-hipped V1, H2-hipped V2, MA-mansard, PI-pinnacle, AR-arched, OT-other



(a) Ground Truth (b) Mask R-CNN (c) Cascade Mask R-CNN (d) SOLOv2 (e) QueryInst

Figure 5. Example results of the selected models in Beijing (rows 1 and 2) and Munich (rows 3 and 4). RGB images converted to gray-scale for better visualization.

4.3.3 Comparisons between Beijing and Munich

Individual cities have unique characteristics influenced by their different culture and history. Particularly, Beijing and Munich have different architectural styles and distributions of buildings' functions in urban areas. To compare the differences between the two cities and the influence on classifications, we produced separate datasets for each city.

We compared these two datasets using the Mask R-CNN model. The results of the experiments are shown in Tables 8

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN [9]	**14.4**	24.9	**15.4**	**5.6**	16.3	**22.0**
C Mask R-CNN [1]	13.6	24.0	14.4	5.1	14.7	21.4
SOLOv2 [22]	14.1	**25.4**	14.4	4.9	**18.0**	21.1
QueryInst [7]	10.4	20.4	10.8	4.3	12.8	16.9

Table 7. Instance segmentation results with coarse-grained function classes using mask AP on the UBC test set. "C Mask R-CNN" denotes Cascade Mask R-CNN.

and 9. The segmentation accuracy in Munich is lower than in Beijing. This is mainly because Munich, on one hand, has a large number of connected closed loops of buildings (apartment building blocks) and, on the other hand, many tiny independent buildings. For the former it is difficult to distinguish the individual building instances inside the building blocks, while the latter are hard to detect. The APs for both roof type and function for Beijing are, therefore, higher than for Munich.

Furthermore, the amount of available data and different characteristics of buildings lead to varying results in different cities. E.g., there are more flat roofs but fewer gable roofs in Beijing than in Munich. Therefore, the flat roof accuracy in Beijing is much higher than in Munich while the gable roof accuracy in Munich is higher than in Beijing. The different characteristics and styles of buildings in different cities also are decisive for the final results of classification for both, roof type and function. There are many high-rise buildings in Beijing but few in Munich. On the other hand, there are many more churches in the city center of Munich than in Beijing. Commercial buildings in Beijing are most likely large shopping malls and look different from residential buildings. Opposed to this, most commercial buildings in Munich look similar to residential ones, therefore, they are difficult to distinguish. Overall, for all the results for Beijing and Munich, the AP of residential buildings is the highest. Industrial buildings have the worst classification results concerning their function, probably because the amount of data is small and some industrial buildings are easily wrongly classified as residential or commercial buildings.

Method	AP	AP_{50}	Roof Type				
			FL	GA	HI	AR	OT
Beijing	22.8	36.6	33.6	18.2	49.0	5.9	7.3
Munich	15.9	28.1	16.3	26.3	30.9	0.0	6.0

Table 8. Results of separate experiments for dividing roof type data in Beijing and Munich, respectively. FL-flat, GA-gable, HI-hipped, AR-arched, OT-other

Method	AP	AP_{50}	Function				
			RE	CO	IN	PU	OT
Beijing	14.0	23.5	40.5	13.2	0.0	9.4	7.0
Munich	13.5	25.3	28.1	10.9	8.2	8.6	11.6

Table 9. Results of separate experiments for dividing function data in Beijing and Munich, respectively. RE-residential, CO-commercial, IN-industrial, PU-public, OT-other

5. Conclusion

We have presented a novel remote sensing dataset with a specific focus on individual buildings and fine-grained classification concerning both, building geometry, i.e., roof type, as well as functionality, i.e., occupation/usage. For classification, predefined classes are given on both a coarse- and a fine-grained level. They can be employed according to different purposes or available numbers of instances. Selected typical urban areas of Beijing and Munich are provided for combined as well as separate investigation. Experiments with multiple state-of-the-art object detection models have been conducted as baseline for further research and possible competition. The experiments demonstrate that the detection and classification of individual buildings in dense urban areas are challenging. The instances show a large diversity concerning size, shape, texture and relationship with neighboring objects, along with influences from history, culture, climate, as well as density of habitation. The function of buildings is a latent feature which can only partially (sometimes not at all) be derived from the appearance. Even for the manual annotation often geo-information from different sources is needed for a reliable decision. We expect it to be a great challenge for classification to find more high-level evidence by integrating, e.g., geometrical features like roof types and the structure of the neighborhood.

In this first version of the dataset selected urban areas of Beijing, China and Munich, Germany, are provided and the input data are solely RGB satellite imagery. In the future, it is planned to extend the dataset in multiple dimensions: (1) Coverage of additional urban areas of interest worldwide, (2) Use of multispectral as well as SAR imagery, and (3) Extension with multi-temporal data. Please note that with an increasing amount of instances, those classes "merged" in the current experiments can be "split" again as individual classes. Additionally, new classes can be added if necessary.

6. Acknowledgements

This work was supported by the National Key R&D Program of China (Grant No. 2021YFB3900504) and National Natural Science Foundation of China (Grant No. 62171436).

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: high quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2019. 6, 7, 8

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open MMLab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 6

[4] edPortal. Gaofen-2. Available at https://directory.eoportal.org/web/eoportal/satellite-missions/g/gaofen-2. 3

[5] Inc. EOS Data Analytics. Superview1. Available at https://eos.com/find-satellite/superview-1/. 3

[6] Hongchao Fan, Alexander Zipf, Qing Fu, and Pascal Neis. Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, 28(4):700–719, 2014. 4

[7] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6910–6919, 2021. 6, 7, 8

[8] Hirsh Goldberg, Myron Brown, and Sean Wang. A benchmark for building footprint classification using orthorectified RGB imagery and digital surface models from commercial satellites. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7. IEEE, 2017. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 6, 7, 8

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 6

[11] 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, 2017. 3

[12] Ribana Roscher, Michele Volpi, Clément Mallet, Lukas Drees, and Jan Dirk Wegner. SemCity Toulouse: A benchmark for building instance segmentation in satellite images. In *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2020*, volume V-5-2020, pages 109–116, Aug. 2020. 3

[13] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012), Nr. 1*, 1(1):293–298, 2012. 2

[14] Michael Schmitt, Seyed Ali Ahmadi, and Ronny Hänsch. There is no data like more data - current status of machine learning datasets in remote sensing. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 1206–1209, 2021. 1

[15] Defence Science and Technology Laboratory. DSTL satellite imagery feature detection. Available at https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection/data. 3

[16] Jacob Shermeyer, Daniel Hogan, Jason Brown, Adam Van Etten, Nicholas Weir, Fabio Pacifici, Ronny Hansch, Alexei Bastidas, Scott Soenen, Todd Bacastow, and Ryan Lewis. Spacenet 6: Multi-sensor all weather mapping dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2, 3

[17] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. 2

[18] Zhi Tian, Hao Chen, Xinlong Wang, Yuliang Liu, and Chunhua Shen. AdelaiDet: A toolbox for instance-level recognition tasks. https://git.io/adelaidet, 2019. 6

[19] Topcoder. Spacenet challenge 7: Multi-temporal urban development challenge. Available at https://go.topcoder.com/spacenet/. 2

[20] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 2, 3

[21] Shenlong Wang, Min Bai, Gellert Mattyus, Hang Chu, Wenjie Luo, Bin Yang, Justin Liang, Joel Cheverie, Sanja Fidler, and Raquel Urtasun. TorontoCity: Seeing the world with a million eyes. *arXiv preprint arXiv:1612.00423*, 2016. 2, 3

[22] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural Information Processing Systems*, 33:17721–17732, 2020. 6, 7, 8

[23] OpenStreetMap Wiki. Openstreetbrowser/category list. Available at https://wiki.openstreetmap.org/wiki/OpenStreetBrowser/Category_list. 4

[24] Junshi Xia, Naoto Yokoya, Bruno Adriano, Lianchong Zhang, Guoqing Li, and Zhigang Wang. A benchmark high-resolution GaoFen-3 SAR dataset for building semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:5950–5963, 2021. 3