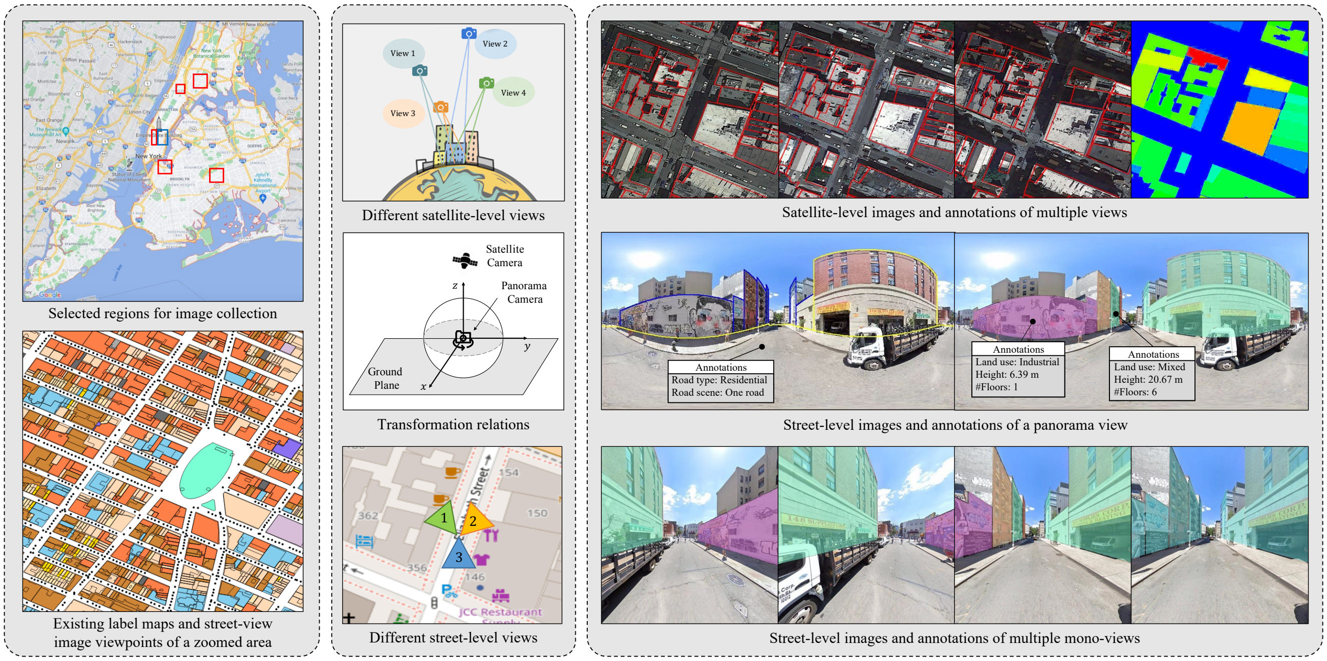
OmniCity: Omnipotent City Understanding with Multi-level and Multi-view Images

Weijia Li[1], Yawen Lai[2], Linning Xu[3], Yuanbo Xiangli[3], Jinhua Yu[1],
Conghui He[2,4*], Gui-Song Xia[5*], Dahua Lin[3,4]
[1]Sun Yat-Sen University, [2]SenseTime Research, [3]The Chinese University of Hong Kong,
[4]Shanghai Artificial Intelligence Laboratory, [5]Wuhan University
liweij29@mail.sysu.edu.cn, alanlyawen@gmail.com, {xl020,xy019,dhlin}@ie.cuhk.edu.hk,
yujh56@mail2.sysu.edu.cn, heconghui@sensetime.com, guisong.xia@whu.edu.cn


Selected regions for image collection

Existing label maps and street-view image viewpoints of a zoomed area

Different satellite-level views

Transformation relations

Different street-level views

Satellite-level images and annotations of multiple views

Street-level images and annotations of a panorama view

Street-level images and annotations of multiple mono-views

Figure 1. An overview of our proposed OmniCity dataset. The satellite and street-level images of our dataset are collected in the six selected regions. The black dots along the streets denote the viewpoint locations and the land-lot colors indicate different fine-grained categories provided by the existing label maps. Note that all the images in the right part correspond to the same geo-location, and the intrinsic transformation relation between the satellite and street-level panorama view is demonstrated in the middle part.

Abstract

This paper presents OmniCity, a new dataset for omnipotent city understanding from multi-level and multi-view images. More precisely, OmniCity contains multi-view satellite images as well as street-level panorama and mono-view images, constituting over 100K pixel-wise annotated images that are well-aligned and collected from 25K geo-locations in New York City. To alleviate the substantial pixel-wise annotation efforts, we propose an efficient street-view image annotation pipeline that leverages the existing label maps of satellite view and the transformation relations between different views (satellite, panorama, and mono-view). With the new OmniCity dataset, we provide benchmarks for a variety of tasks including building footprint extraction, height estimation, and building plane/instance/fine-grained segmentation. Compared with existing multi-level and multi-view benchmarks, OmniCity contains a larger number of images with richer annotation types and more views, provides more benchmark results

[*] Corresponding authors.

of state-of-the-art models, and introduces a new task for fine-grained building instance segmentation on street-level panorama images. Moreover, OmniCity provides new problem settings for existing tasks, such as cross-view image matching, synthesis, segmentation, detection, etc., and facilitates the developing of new methods for large-scale city understanding, reconstruction, and simulation. The OmniCity dataset as well as the benchmarks will be released at https://city-super.github.io/omnicity/.

1. Introduction

Owning over a half global population and contributing the most economic growth, the city areas have been recorded and characterized by various data sources including satellite and aerial imagery, street-level imagery, LiDAR data, public maps, crowd-sourced data, etc. A great number of benchmarks have been proposed towards facilitating different vision tasks in city scene, of which the street-level imagery has been broadly used in multiple driving-related benchmarks [4, 10, 14, 18, 26, 31, 43]. The rich visual information of street-level imagery also enables complicated visual recognition tasks on specific categories, such as person detection [3], vehicle tracking and re-identification [32], and fine-grained land use classification [49].

Nevertheless, constructing pixel-wise annotations for street-level imagery requires substantial human efforts, resulting in the small image quantity and limited view types of existing datasets, especially for the street-view panorama datasets with only hundreds of annotated images [36, 41, 42]. Regarding the annotation categories and levels, existing datasets mostly provide instance-level annotations for dynamic object categories in driving scenes. As a vital component for city understanding, the static objects such as buildings and roads take up a larger proportion of cities and remain a high consistency across the satellite and ground-level images. However, existing street-level datasets either provide pixel-wise building annotations without fine-grained semantic labels [4, 10, 14, 47] or provide fine-grained annotations at only bbox or image level [46, 49].

Compared with street-level images, remote sensing images usually contain less visual information for conducting complicated tasks such as fine-grained land use segmentation and building function recognition. On the other hand, unlike the sparsely-distributed street-level images, remote sensing images have a dense spatial distribution and a worldwide coverage, which are well aligned with the open maps and government datasets at pixel level [22]. These existing maps and datasets contain a variety of satellite-level annotations for buildings (such as the footprint, land use, height, year built), roads (category and line coordinates), and other geographical objects, providing new perspectives for promoting novel city understanding datasets and tasks.

In this work, as illustrated in Figure 1, we construct an omnipotent city dataset unifying data sources from both satellite and street views, linked by geo-locations and urban planning data. Unlike existing city datasets that only support a limited number of tasks, OmniCity dataset incorporate rich geometric annotations and semantic meta data for each image, where multiple tasks can be conducted on. To leverage the existing map labels and the rich visual context from the street-level imagery, we propose an efficient pipeline for producing diverse street-level annotations. Based on this annotation pipeline, we built OmniCity, a dataset that contains over 100K annotated images collected from 25K geo-locations in New York City. We provide benchmark results on OmniCity for a variety of tasks, including building footprint extraction and height estimation on satellite images, as well as fine-grained/instance/plane segmentation of buildings on street-level panorama and mono-view images. To the best of our knowledge, this is the first work that involves fine-grained building instance segmentation on street-level panorama images. We also analyze the potential of OmniCity for promoting new tasks and methods with multi-level imagery.

Our main contributions are summarized as follows:

- We propose a novel pipeline for efficiently producing diverse pixel-wise annotations on street-level panorama and mono-view images.
- We build the OmniCity dataset, which contains well-aligned satellite and street-level images with a larger quantity, richer annotations and more views compared with existing datasets.
- We provide a series of benchmark experimental results for multiple tasks and data sources, and analyze the limitations of the current benchmarks on OmniCity.
- We discuss the potential of OmniCity for facilitating new methods and tasks for large-scale city understanding, reconstruction, and simulation.

2. Related work

2.1. Datasets and methods for street-level tasks

As shown in Table 1, many street-level datasets have been proposed over the past few years. A large proportion of these datasets are designed for visual tasks in driving scene [4, 10, 14, 18, 26, 31, 44], such as 2D/3D object detection, semantic segmentation, object tracking, etc. Several street-level datasets are proposed for a specific object category, such as the EuroCity Persons dataset [3], the CityFlow dataset [32] for vehicle tracking and re-identification, etc. In addition to the above datasets containing only mono-view images, some studies propose new datasets or methods

Table 1. A comparison of OmniCity with existing city-related datasets of street-level, satellite-level and cross-level. The Street and Satellite (Sate.) columns show the available image view types. The last three columns indicate which level of tasks the dataset is designed for (semantic/instance/plane segmentation, object detection (bbox), and image classification), and whether the dataset contains fine-grained building attribute (Attri.) or height labels.

Dataset	#Images	Street	Sate.	Anno.	Attri.	Height
KITTI [14]	15,000	mono	-	semantic	×	×
Cityscapes [10]	25,000	mono	-	semantic	×	×
EuroCity [3]	47,300	mono	-	bbox	×	×
WildPASS [42]	500	multi.	-	semantic	×	×
PASS [41]	400	multi.	-	semantic	×	×
HoliCity [47]	6,300	multi.	-	inst./plane	×	×
SkyScapes [1]	8,820	-	single	semantic	×	×
SpaceNet [38]	60,000	-	multi.	instance	×	×
Christie et al. [9]	11,000	-	single	semantic	×	✓
Li et al. [21]	3,300	-	single	instance	×	✓
TorontoCity [36]	Unknow	multi.	multi.	instance	×	✓
Wojna et al. [39]	49,426	mono	single	image	✓	×
OmniCity	**108,600**	**multi.**	**multi.**	**inst./plane**	✓	✓

for semantic segmentation from panorama images, such as TorontoCity [36], PASS [41], and WildPASS [42]. However, such datasets require expensive annotation efforts and contains only hundreds of pixel-wise annotated panorama images in total. Besides the driving-related datasets, several recent studies target at the recognition tasks for static object categories. Zhu et al. [49] proposed a framework for fine-grained land use classification task using ground-level images. Zhang et al. [46] proposed a building recognition system for detecting their business entity information. HoliCity [47] proposed holistic 3D structure annotations generated from CAD models for panorama images, but it only conducted experiments on mono-view images and lacked the semantic information of buildings and roads.

In summary, existing street-level datasets still have the following limitations. Regarding the image quantity and view types, most existing datasets require substantial annotation efforts and contain only a limited number of annotated images of a mono view. The image quantity of panorama datasets is even several orders of magnitude smaller than mono-view datasets. Regarding the annotation categories and levels, existing datasets mainly focus on dynamic or driving-related categories and lack in fine-grained annotations of an object category. Moreover, the datasets for static objects are lacking in pixel-level or fine-grained annotations. By contrast, our OmniCity contains over 100K satellite and street-level images of multiple views as well as the pixel-wise and fine-grained annotations, demonstrating superiority in annotation quantity, annotation type, and view type compared with existing datasets.

2.2. Datasets and methods for satellite-level tasks

As a data source with a long time series and a large coverage, the satellite imagery has been broadly explored for large-scale city understanding. Unlike the street-level datasets requiring manual annotations, the satellite imagery is already well aligned with existing label maps [20]. The OpenStreetMap (OSM) is one of the most broadly-used map data, which contains publicly available annotations of building footprints, heights, roads, land use, etc., of world wide. In addition, many public datasets provide rich information at a local scale. For example, the PLUTO [1] dataset contains the block and lot information of the whole New York City, and each lot is associated with the land use, year built, number of floors, and other useful information. The Microsoft US building footprint dataset contains over a hundred million computer-generated building footprints. Several challenges provide manually labeled building footprints and the corresponding satellite images [11, 34, 38]. Moreover, some datasets provide fine-grained semantic categories of different objects [1] or building height annotations for 3D reconstruction tasks [9, 21].

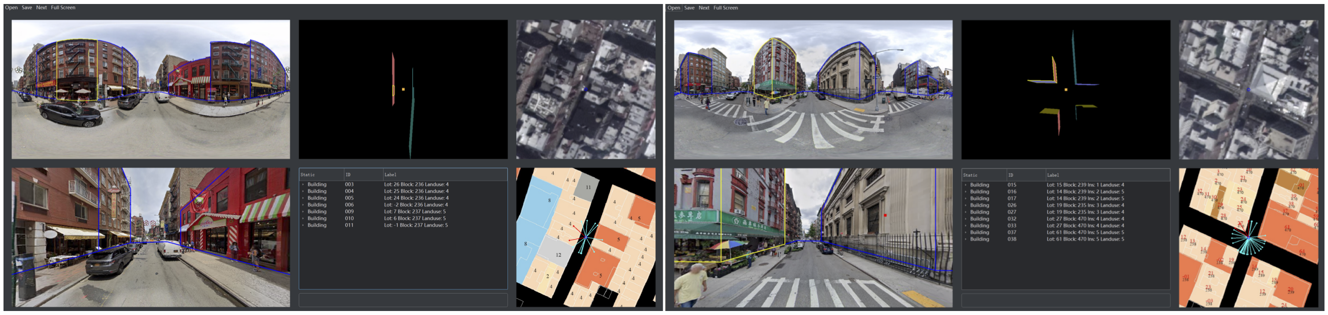
In summary, existing maps and datasets have provided substantial semantic and geometric information that is well aligned with the satellite imagery. In this work, we leverage these rich annotations and the transformation relations between different views to produce auxiliary information for street-level image annotation. Compared with existing work, OmniCity significantly reduces the human labeling efforts and provides more annotation types to enable omnipotent city understanding via multiple tasks and views.

2.3. Datasets and methods for multi-level tasks

The ground-level images usually contain rich visual context that is not visible from the satellite or aerial imagery (e.g. building facade, the side of vegetations, etc.), while the spatial distribution is often sparse and unbalanced in different areas. By contrast, the remote sensing images have a much denser spatial distribution at a global scale, but the visual context is too limited to support complicated fine-grained tasks. Considering the above complementary characteristics, many datasets and methods have been proposed for cross-view scenarios. Cross-view visual recognition tasks integrate the two data types as the model input for predicting building functions, building ages, land use, tree species, etc. [2, 12, 37, 40], while cross-view image matching [17, 28, 33, 48] and image synthesis [25, 27, 30, 45] tasks take only one data type as the model input.

In addition, several datasets are proposed for a variety of satellite and street-level tasks. The TorontoCity [36] contains a wide range of annotations including building height estimation, building instance segmentation, building footprint segmentation, road segmentation, etc., which are conducted on either satellite or street-level images and only adopt FCN and some CNN classification architectures as

[1] https://data.cityofnewyork.us/City-Government/Primary-Land-Use-Tax-Lot-Output-PLUTO-/64uk-42ks.



Static	ID	Label
Building	003	Lot: 26 Block: 236 Landuse: 4
Building	004	Lot: 25 Block: 236 Landuse: 4
Building	005	Lot: 24 Block: 236 Landuse: 4
Building	006	Lot: -2 Block: 236 Landuse: 4
Building	009	Lot: 7 Block: 237 Landuse: 5
Building	010	Lot: 6 Block: 237 Landuse: 5
Building	011	Lot: -1 Block: 237 Landuse: 5

Static	ID	Label
Building	015	Lot: 15 Block: 239 Ins: 1 Landuse: 4
Building	016	Lot: 14 Block: 239 Ins: 2 Landuse: 5
Building	017	Lot: 14 Block: 239 Ins: 2 Landuse: 5
Building	026	Lot: 19 Block: 235 Ins: 3 Landuse: 4
Building	027	Lot: 19 Block: 235 Ins: 3 Landuse: 4
Building	032	Lot: 27 Block: 470 Ins: 4 Landuse: 4
Building	033	Lot: 27 Block: 470 Ins: 4 Landuse: 4
Building	037	Lot: 61 Block: 470 Ins: 5 Landuse: 5
Building	038	Lot: 61 Block: 470 Ins: 5 Landuse: 5

Figure 2. The annotation tool of OmniCity dataset. The left and right images show examples of one road scene with the label map (in the bottom-right window) displayed in land use mode, and crossroads scene with the label map in block-lot mode, respectively.

the baseline models for all experiments. In a recent study [39], a holistic multi-view building analysis dataset is designed for multiple recognition tasks regarding facade material, number of floors, occupancy type, roof geometry, roof pitch, and construction type. The dataset contains only image-wise annotations for street-level classification tasks. The semantic categories of each building are decided by human annotators, resulting in more challenges and subjectivity to the annotation process compared with our work.

In summary, for the existing cross-view studies designed for one specific task, the street-level images are not annotated and only supplement extra features to the satellite-level feature maps. By contrast, our OmniCity provides rich annotation types that are not contained in existing datasets, e.g. the pixel-wise building instances and fine-grained categories on street-level images, which can promote new methods to explore and leverage these annotations to improve the performance. In addition, OmniCity provides the benchmark experimental results of each task using more state-of-the-art models. With our efficient annotation pipeline, it also provides additional fine-grained pixel-wise annotations and benchmark results compared with [36] and [39].

3. Datasets

In this work, we aim at building a dataset for omnipotent city understanding from multi-level and multi-view images. Our proposed OmniCity dataset contains 108,600 images of multiple views, which are collected from 25K geo-locations in New York City. Compared with existing datasets, OmniCity requires much fewer human efforts for street-level image annotation, contains more diverse annotation types for both 2D and 3D tasks, provides richer building semantics at instance segmentation level, and possesses higher scalability for new annotation type supplement and expansion to other cities. The details of data collection, annotation, and statistics are introduced as follows.

3.1. Data collection

As shown in Figure 1, our OmniCity dataset is collected from six selected regions of New York. We download the panorama images in the six selected regions using google street view download 360, with a step distance of 65 meters. The regions for collecting the training and test samples are denoted by red and blue, respectively. We save the geographic coordinates, collection time, panorama id, north rotation, and zoom level for each panorama image. For each panorama site, we collect its corresponding google earth images of three acquisition dates according to the geographic coordinates, constituting three groups of satellite-level datasets with three cases of off-nadir view angles (small/medium/high, denoted by V1/V2/V3). For the annotation data sources, we collect the meta information from PLUTO and OpenStreetMap (OSM). The New York City is hierarchically formed by blocks, lots, and buildings. Each building can be identified by a specific block-lot id. In PLUTO, each lot (building) is associated with rich information, e.g. land use, year built, number of floors, etc. The OSM data contains footprint and height information while lacking in land use information for most buildings. Considering the characteristics of the two data sources, we align the land use attribute (from PLUTO) with the building footprint and height (from OSM) using the geographical coordinates. Overall, each building is assigned with a block-lot id, a land use category, a height value, and the geographical coordinates of a footprint polygon, which will be used as the reference label maps for panorama image annotation.

3.2. Data annotation

Figure 2 shows the panorama image annotation tool proposed in our study. The street-level panorama and satellite images are naturally well-aligned at image-level between the central coordinates from satellite views and the camera pose of panorama views. The instance-level annotations of different views (*e.g.*, building facade and footprint) are further aligned during the annotation process. According to the geo-transformation relation between different views [25], we calculate the angle between the north (upward) direction and each building footprint (in the satellite image), and the angle between the north direction (saved in the panorama file) and each building facade in the horizontal axis. The footprint and facade with the same angle value are aligned to the same building, providing the auxiliary information for the manual annotation process.

The annotation pipeline includes four stages: (1) Image selection, i.e., select the panorama images that are essential to be annotated according to building coverage, occlusion extent, etc; (2) Segmentation annotation, i.e., adjust the floor/top line to fit the bottom/roof of each building, and add the boundary split line considering both auxiliary information and building appearance (e.g. texture discrepancy, doors, etc.); (3) Attribute assignment, i.e., add the attributes (instance ID, block-lot id and land use type) for each building plane; (4) Quality assessment, i.e., check the annotation quality and remove the unqualified images.

In addition, the mono-view images and annotations are automatically generated from those of the panorama images via view transformation. For each panorama image, we select three views using three x-axis angles (-170, 10 and 170, in the range of [-180, 180]) and a fixed y-axis angle of 0. Then we design an image selection rule to filter out the unexpected images of which the buildings are distributed in only one side, building area proportion is smaller than 10%, or the building plane quantity is smaller than 2. The remaining images constitute the final mono-view dataset of our experiments. For satellite-level tasks, the annotations are already well-aligned with the meta information from PLUTO and OSM. Each building footprint on a satellite image is assigned with a block-lot id, a land use category, a height value, and the pixel coordinates of the footprint polygon.

3.3. Statistics of the proposed dataset

The whole dataset contains three sub-datasets (in small, medium and large view angles) for satellite-level tasks and two sub-datasets (in panorama and mono views) for street-level tasks. For satellite-level datasets, the three sub-datasets of images collected from the 25K geo-locations constitute 75K images in total, which are cropped by 512 $\times$ 512 pixels according to the coverage area of the corresponding panorama image. For the street-level datasets, the panorama dataset contains 18K images in 512 $\times$ 1024 pixels, which are selected from the initial 25K panorama images during the image selection and quality assessment phases. Similarly, the mono-view dataset obtained via view transformation and image selection contains 156,00 mono-view images in 512 $\times$ 512 pixels. The ratio of train/test splits is set as 4:1 for all five sub-datasets. In OmniCity dataset, the initial land use categories of PLUTO with similar characteristics and low quantities are merged into one category, resulting in 7 land use categories in total. The categories of 1/2 family building ($\sim$17%), walk-up building ($\sim$23%) and the mixed residential/commercial building categories ($\sim$29%) take up a larger proportion compared with the other four categories, i.e., elevator buildings ($\sim$7%), office buildings ($\sim$10%), industrial/transportation/utility buildings ($\sim$11%) and others ($\sim$1%). For the distribution of building height, most are between 1 and 25 meters while a small percentage ($\sim$1%) reaches over 50 meters.

4. Benchmark results

In this section, we provide a variety of benchmarks for multiple satellite and street-level tasks. The satellite-level tasks in our experiments include building footprint segmentation and height estimation. For both tasks, we conduct experiments on the satellite images with three view angles. For the street-level tasks, we conduct two instance segmentation tasks (i.e., land use and building instance segmentation) on the panorama images, and three instance segmentation tasks (i.e., land use / building instance / plane segmentation) on mono-view images. Please note that these are only preliminary experimental results on OmniCity dataset. More benchmarks of latest models and additional tasks will be continuously updated on OmniCity homepage.

4.1. Experimental setting

The experiments are mainly based on mmdetection [7] with the recommended hyper-parameter settings. We select Mask R-CNN [15] as the baseline method for segmentation tasks, and provide a comparison of different methods in Table 6, including Mask Scoring R-CNN (MS R-CNN) [19], Cascade Mask R-CNN (Cascade) [5], Content-Aware ReAssembly of FEatures (CARAFE) [35], and Hybrid Task Cascade (HTC) [6]. Specifically, we use ResNet-50 [16] with FPN [23] pre-trained on the ImageNet [29] as the backbone for all instance segmentation models. All models are trained on 8 NVIDIA Tesla V100 GPUs for 12 epochs, with a batch size of 16, a learning rate starting from 0.02 and decreasing by a factor of 0.1 from the 8^{th} to 11^{th} epoch, and the stochastic gradient descent (SGD) optimizer with a weight decay of 10^{-4} and a momentum of 0.9. For the height estimation task, we evaluate the performance of two widely-used monocular depth estimation methods on the satellite images of three view angles, i.e. Structure-Aware Residual Pyramid Network (SARPN) [8] and Deep Ordinal Regression Network (DORN) [13], which are trained on 4 NVIDIA Tesla V100 GPUs for 20 epochs. SARPN is trained with a batch size of 8, a learning rate starting from 10^{-4} and reduced by 10% every 5 epochs, and the Adam optimizer with β_1 = 0.9, β_2 = 0.999, and a weight decay of 10^{-4}. DORN is trained with a batch size of 4, a base learning rate of 10^{-4} and the power of 0.9, using SGD optimizer with a weight decay of 0.0005 and a momentum of 0.9.

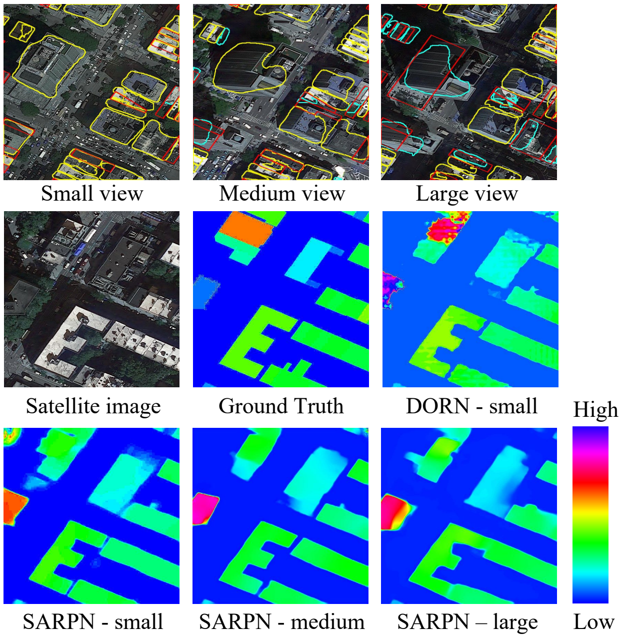

Small view
Medium view
Large view
Satellite image
Ground Truth
DORN - small
SARPN - small
SARPN - medium
SARPN – large

Figure 3. Qualitative results of satellite-level tasks. The yellow/cyan/red polygons denote TP/FP/FN buildings.

4.2. Experimental results of satellite-level tasks

Table 2 shows the baseline results of the satellite-level instance segmentation task, which are evaluated using both COCO [24] and SpaceNet [11] evaluation metrics. The footprint segmentation performance is the best for satellite images with a small view angle, and deteriorates seriously when the view angle gets larger. For all three cases, the prediction score is the highest for buildings with large areas and the lowest for small buildings. Figure 3 provides a qualitative comparison of the footprint segmentation results on satellite images of three types of view angle. The large view angle results in great difficulties for extracting accurate footprint boundaries, due to the partial invisibility of building footprint, the serious shadow effects, etc. The height estimation performance is evaluated in terms of the mean absolute error (denoted by MAE), mean square error (denoted by MSE), and root mean square error (denoted by RMSE), which are commonly used metrics for depth estimation. All metrics are measured in meters at pixel level. Table 3 and Figure 3 show the height estimation results obtained from DORN [13] and SARPN [8] for satellite images with different view angles. Results demonstrate that DORN obtains better results compared with SARPN for all three cases. Both methods achieve the best performance for satellite images with a medium view angle (V2) compared with the other two cases. The footprint and roof have more overlaps and provide less building structure information for satellite images with a small view angle, while the the shadow and parallax effect become more serious with the increase of off-nadir view angle. The above aspects result in challenges for the accurate estimation of building height for satellite images with small and large view angles.

Table 2. Quantitative results of instance segmentation for satellite images with different view angles.

View	Metrics of various thresholds						threshold = 0.5		
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	P	R	F1
V1	**29.7**	**66.0**	**23.5**	**15.9**	**33.9**	**36.7**	**76.9**	**66.3**	**71.2**
V2	23.7	56.6	16.1	11.5	27.2	30.3	73.9	55.0	63.1
V3	18.9	51.4	9.6	9.1	21.5	25.3	70.7	51.7	59.7

Table 3. Quantitative results of height estimation for satellite images with different view angles.

View	SARPN [8]			DORN [13]		
	MAE	MSE	RMSE	MAE	MSE	RMSE
V1	16.18	870.34	29.50	12.71	670.52	25.89
V2	**13.75**	**694.17**	**26.35**	**12.24**	**628.06**	**25.06**
V3	15.32	823.01	28.69	13.40	730.67	27.03

4.3. Experimental results of street-level tasks

We analyze the performance of multiple segmentation tasks on street-level panorama and mono-view images. Table 4 and Figure 4 show the experimental results of panorama-view images on two segmentation tasks, of which the performance on building instance segmentation task (denoted by Instance Seg.) is significantly superior to the fine-grained land use segmentation task (denoted by Landuse Seg.). Table 5 and Figure 5 show the experimental results of mono-view images on three different tasks, i.e., landuse segmentation, instance segmentation, and plane segmentation. Similar to the results of panorama-view images, the baseline method achieves much higher scores for the two binary segmentation tasks (plane and instance segmentation) compared with the fine-grained land use segmentation task. The qualitative results also demonstrate that the baseline method has difficulties in identifying the accurate land use type of some building instances.

Table 6 shows the land use segmentation results obtained from the five methods on street-level panorama images. HTC achieves the best performance for both overall and category metrics followed by MS R-CNN, which indicates that the cascade structure and mask scoring strategy can effectively improves the fine-grained segmentation performance of building instances. From the metrics of each category, we can find that C2, C4 and C5 (i.e., Walkup Buildings, Mixed Residential/Commercial, and Office Buildings) have better performance compared with C1, C3 and C6 (i.e., 1/2 Family Buildings, Elevator Buildings and Industrial/Transportation/Utility) for all methods, with the AP ranging from 20% to 40%. The category of Others has an extremely low AP score due to the small sample quantity. Compared with the residential buildings, the mixed Residential/Commercial and Office buildings have more special

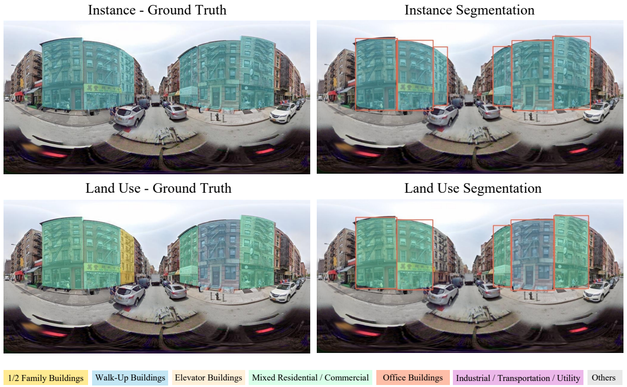


Figure 4. Qualitative results of street-level tasks (i.e., instance segmentation and land use segmentation) on panorama images.

Table 4. Quantitative results on street-level panorama images.

Task	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Landuse Seg.	26.0	34.7	28.5	0.3	12.0	30.4
Instance Seg.	66.7	86.5	72.5	1.7	40.2	74.1

Table 5. Quantitative results on street-level mono-view images.

Task	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Landuse Seg.	23.9	32.1	26.7	0.3	10.6	27.5
Instance Seg.	68.3	88.8	73.8	3.2	33.3	76.1
Plane Seg.	65.1	87.4	71.0	5.0	40.7	73.8

characteristics (i.e. the boundary between first and above floors, facade design, and building structure), contributing to the superior AP scores of these two categories. The performance of the above six categories is also concordant with the ratio of the sample quantity.

4.4. Results analysis and discussions

In this section we summarize the limitations of existing methods for satellite and street-level tasks on our OmniCity dataset. For building footprint segmentation, the performance of existing methods get worse with the increasing of off-nadir view angle, which might due to the serious parallax and shadow effect on the satellite images with a large off-nadir view angle. For the height estimation task, most existing methods directly apply the monocular depth estimation methods to remote sensing scene. These methods usually produce poor results on the invisible side of the footprint boundary especially for high-rise buildings on very off-nadir images. In addition, most existing methods use deep neural networks to regress continuous values but the actual height values are discrete for building and non-building areas, resulting in difficulties for network training and extra efforts for converting the continuous prediction values into discrete height values via post-processing.

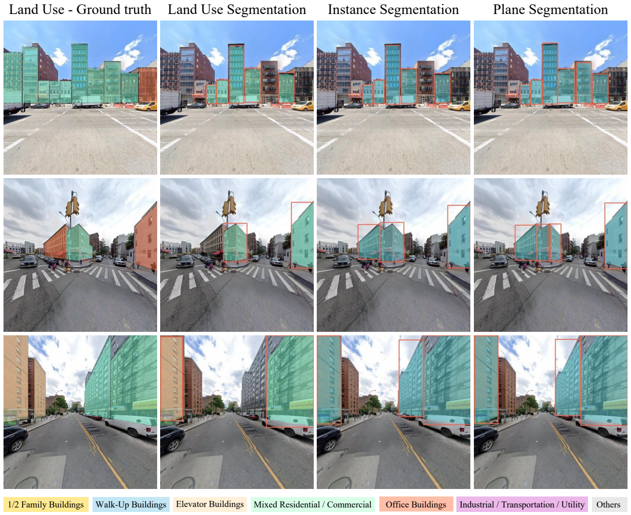


Figure 5. Qualitative results of street-level tasks (i.e., land use /instance/plane segmentation) on mono-view images.

For the street-level tasks based on panorama images, existing methods target at general instance segmentation tasks for commonly-used datasets, e.g. COCO [24], CityScapes [10], BDD100K [44], etc. These datasets often have a single view and a narrow Field of View (FoV). However, for panorama images, the special properties such as the wide FoV covering full 360-degree in the horizontal direction, are not taken into consideration in the design of existing methods. For both mono-view and panorama images, existing methods have difficulties in accurately recognizing the building instance with a small area (e.g. buildings located in the side of the main parts), the land use categories with a small number of building instances, and the categories that are easily confused (e.g. 1/2 Family Buildings and Walk-Up Buildings, Mixed Residential/Commercial and Office Buildings). Figure 6 shows some typical failure cases of the current benchmark methods. New instance segmentation methods should be designed for solving the above limitations considering the characteristics of panorama images, building instances, fine-grained categories, etc.

5. Potential of the OmniCity dataset

Our proposed OmniCity dataset demonstrates great potential for facilitating city understanding, machine perception, and generative modeling researches in many aspects.

First, it can serve as a new dataset for the existing tasks such as image geo-localization/synthesis and segmentation/detection of buildings/trees/land use from cross-view images. OmniCity provides additional annotations that are not contained in existing datasets, e.g. the building instances and fine-grained categories on street-level images, which can promote new methods to explore and leverage

Table 6. Quantitative results of different methods for fine-grained land use segmentation on street-level panorama images.

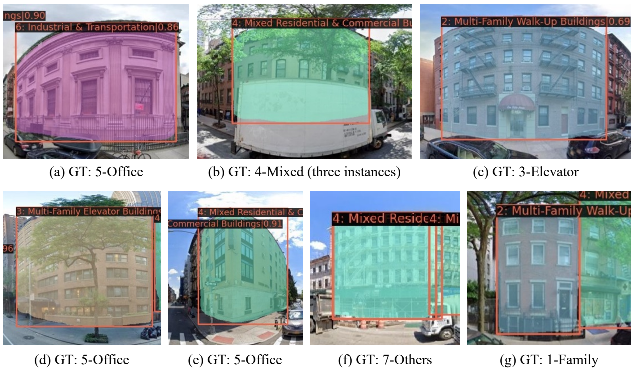
Method	Overall Metrics						Metrics of each category						
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	C1	C2	C3	C4	C5	C6	C7
Mask R-CNN [15]	26.0	34.7	28.5	**0.3**	12.0	30.4	19.6	37.5	25.8	39.2	36.9	22.2	0.8
MS R-CNN [19]	27.1	**35.8**	29.8	0.1	**12.4**	31.5	**22.5**	**39.1**	26.2	**40.8**	38.0	21.7	**1.2**
Cascade [5]	25.9	33.8	28.3	0.2	11.4	30.5	20	38.3	25	38.5	36.7	22.1	0.3
CARAFE [35]	25.9	34.5	28.5	0.1	11.9	30.2	19.6	37.3	24.9	39.9	37.2	21.5	0.8
HTC [6]	**27.2**	35.7	**29.9**	**0.3**	**12.4**	**32.0**	20.8	38.7	**27.2**	39.9	**38.4**	**24.5**	**1.2**

the new annotations to improve the performance of these tasks. The annotations are organized in a unified version, which means multiple tasks can be performed on a single image, and thus can well support the multi-task learning setting. Additionally, since the building instances are directly linked with the urban planning data using block-lot id, it is easy to be enriched with more annotation types from other urban datasets, especially those for social and urban studies.

Second, OmniCity provides a new application scenario or problem setting to existing tasks. For line segment detection and wireframe parsing tasks, existing datasets contains densely distributed line segment and wireframe labels, while our OmniCity focuses on the main line segments and wireframes on the outlines instead of the inner ones, resulting in a much sparser format. The serious shelters from the trees and vehicles also bring challenges to these tasks. New line segment detection and wireframe parsing methods should be designed for the OmniCity scenario.

Moreover, OmniCity facilitates new tasks for city reconstruction and simulation. Treating each panorama as a unique city scene, a complete 3D model representing such a local scene is available, as also shown in the second window panel in our tool (Figure 2). These 3D models are stored in abstract vector formats, with clean and clear vertices indicating the building facades, streets, etc., which are suitable for many shape generation tasks represented with graphs.

Finally, with the well-aligned satellite and street-level images as well as the various annotation types, novel city reconstruction tasks, e.g., 3D building reconstruction from cross-view images, can be derived for producing holistic 3D buildings with both fine-grained semantic category and precise geometry information (vector 3D model). Unlike the existing studies that only target at 3D reconstruction from monocular or multi-view remote sensing imagery, new methods should be designed to leverage the additional information from street-level images to improve 3D reconstruction and semantic prediction for OmniCity scenario.


ngs|0.90
6: Industrial & Transportation|0.86
4: Mixed Residential & Commercial Bu
2: Multi-Family Walk-Up Buildings|0.69
3: Multi-Family Elevator Buildings
96
4: Mixed Residential & C
Commercial Buildings|0.91
4: Mixed Reside
4: Mi
4: Mixed
2: Multi-Family Walk-Up

(a) GT: 5-Office (b) GT: 4-Mixed (three instances) (c) GT: 3-Elevator

(d) GT: 5-Office (e) GT: 5-Office (f) GT: 7-Others (g) GT: 1-Family

Figure 6. Typical failure cases of the current benchmark methods.

6. Conclusion

In this paper, we have proposed OmniCity, a new dataset for omnipotent city understanding from satellite and street-level images of multiple views. The dataset contains over 100K images collected from 25K geo-locations in New York City, of which the annotations are generated from both existing label maps and our proposed annotation pipeline. We provide benchmark experimental results for multiple tasks and data sources based on state-of-the-art methods and analyze their limitations. We believe that OmniCity will not only promote new algorithms and application scenarios for existing tasks, but facilitate novel tasks for 3D city reconstruction and simulation. In our future work, we will keep updating the OmniCity dataset and the benchmarks in the following aspects. Owing to the proposed annotation pipeline and the unified annotations with a vector format and rich meta information (geo-locations, block-lot id, etc.), OmniCity can be efficiently supplemented with more properties of buildings and other geographical object types (roads, sidewalks, trees, green space, etc.), and extended to other cities of different countries. The benchmark results of more state-of-the-art models and new tasks will be provided accordingly. Based on the rich annotation and view types of OmniCity, we also plan to develop new methods for existing and novel tasks, such as object detection, instance segmentation, and 3D reconstruction from cross-view images.

Acknowledgements. This project was funded in part by National Natural Science Foundation of China (No. 42201358, No. U22B2011 and No. 41820104006), Shanghai AI Laboratory, CUHK Interdisciplinary AI Research Institute, and the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK.

References

[1] Seyed Majid Azimi, Corentin Henry, Lars Sommer, Arne Schumann, and Eleonora Vig. Skyscapes fine-grained semantic understanding of aerial scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7393–7403, 2019.

[2] Sara Beery, Guanhang Wu, Trevor Edwards, Filip Pavetic, Bo Majewski, Shreyasee Mukherjee, Stanley Chan, John Morgan, Vivek Rathod, and Jonathan Huang. The auto arborist dataset: A large-scale benchmark for multiview urban forest monitoring under domain shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21294–21307, 2022.

[3] Markus Braun, Sebastian Krebs, Fabian B. Flohr, and Dariu M. Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[5] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2019.

[6] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019.

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[8] Xiaotian Chen, Xuejin Chen, and Zheng-Jun Zha. Structure-aware residual pyramid network for monocular depth estimation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 694–700, 2019.

[9] Gordon Christie, Rodrigo Rene Rai Munoz Abujder, Kevin Foster, Shea Hagstrom, Gregory D Hager, and Myron Z Brown. Learning geocentric object pose in oblique monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14512–14520, 2020.

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[11] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[12] Tian Feng, Quang-Trung Truong, Duc Thanh Nguyen, Jing Yu Koh, Lap-Fai Yu, Alexander Binder, and Sai-Kit Yeung. Urban zoning using higher-order markov random fields on multi-view imagery data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 614–630, 2018.

[13] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018.

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision (CVPR)*, pages 2961–2969, 2017.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018.

[18] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2702–2719, 2019.

[19] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6409–6418, 2019.

[20] Weijia Li, Conghui He, Jiarui Fang, Juepeng Zheng, Haohuan Fu, and Le Yu. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source gis data. *Remote Sensing*, 11(4):403, 2019.

[21] Weijia Li, Lingxuan Meng, Jinwang Wang, Conghui He, Gui-Song Xia, and Dahua Lin. 3d building reconstruction from monocular remote sensing images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12548–12557, 2021.

[22] Weijia Li, Wenqian Zhao, Huaping Zhong, Conghui He, and Dahua Lin. Joint semantic–geometric learning for polygonal building segmentation. In *AAAI*, 2021.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2117–2125, 2017.

[24] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[25] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 859–867, 2020.

[26] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017.

[27] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3501–3510, 2018.

[28] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 470–479, 2019.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision (IJCV)*, 115(3):211–252, 2015.

[30] Yujiao Shi, Dylan John Campbell, Xin Yu, and Hongdong Li. Geometry-guided street-view panorama synthesis from satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[31] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[32] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8797–8806, 2019.

[33] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3616, 2017.

[34] Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development spacenet dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2021.

[35] Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, and Dahua Lin. Carafe: Content-aware reassembly of features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3007–3016, 2019.

[36] 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. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3028–3036. IEEE, 2017.

[37] Jan D Wegner, Steven Branson, David Hall, Konrad Schindler, and Pietro Perona. Cataloging public objects using aerial and street-level images-urban trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6014–6023, 2016.

[38] Nicholas Weir, David Lindenbaum, Alexei Bastidas, Adam Van Etten, Sean McPherson, Jacob Shermeyer, Varun Kumar, and Hanlin Tang. Spacenet mvoi: A multi-view overhead imagery dataset. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 992–1001, 2019.

[39] Zbigniew Wojna, Krzysztof Maziarz, Łukasz Jocz, Robert Pałuba, Robert Kozikowski, and Iason Kokkinos. Holis-

tic multi-view building analysis in the wild with projection pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2870–2878, 2021.

[40] Scott Workman, M Usman Rafique, Hunter Blanton, and Nathan Jacobs. Revisiting near/remote sensing with geospatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2022.

[41] Kailun Yang, Xinxin Hu, Luis M Bergasa, Eduardo Romera, and Kaiwei Wang. Pass: Panoramic annular semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 21(10):4171–4185, 2019.

[42] Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Is context-aware cnn ready for the surroundings? panoramic semantic segmentation in the wild. *IEEE Transactions on Image Processing*, 30:1866–1881, 2021.

[43] Xiaoqing Ye, Mao Shu, Hanyu Li, Yifeng Shi, Yingying Li, Guangjie Wang, Xiao Tan, and Errui Ding. Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21341–21350, 2022.

[44] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[45] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–875, 2017.

[46] Chiqun Zhang, Dragomir Yankov, Chun-Ting Wu, Simon Shapiro, Jason Hong, and Wei Wu. What is that building? an end-to-end system for building recognition from streetside images. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2425–2433, 2020.

[47] Yichao Zhou, Jingwei Huang, Xili Dai, Shichen Liu, Linjie Luo, Zhili Chen, and Yi Ma. Holicity: A city-scale data platform for learning holistic 3d structures. *arXiv preprint arXiv:2008.03286*, 2020.

[48] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021.

[49] Yi Zhu, Xueqing Deng, and Shawn Newsam. Fine-grained land use classification at the city scale using ground-level images. *IEEE Transactions on Multimedia*, 21(7):1825–1838, 2019.