
TMBuD: A DATASET FOR URBAN SCENE BUILDING DETECTION

A PREPRINT

✉ **Ciprian Orhei**

Politehnica University of Timișoara
Timișoara, Romania
ciprian.orhei@cm.upt.ro

✉ **Silviu Vert**

Politehnica University of Timișoara
Timișoara, Romania
silviu.vert@upt.ro

✉ **Muguras Mocofan**

Politehnica University of Timișoara
Timișoara, Romania
muguras.mocofan@upt.ro

✉ **Radu Vasiu**

Politehnica University of Timișoara
Timișoara, Romania
radu.vasiu@upt.ro

October 28, 2021

ABSTRACT

Building recognition and 3D reconstruction of human made structures in urban scenarios has become an interesting and actual topic in the image processing domain. For this research topic the Computer Vision and Augmented Reality areas intersect for creating a better understanding of the urban scenario for various topics. In this paper we aim to introduce a dataset solution, the TMBuD, that is better fitted for image processing on human made structures for urban scene scenarios. The proposed dataset will allow proper evaluation of salient edges and semantic segmentation of images focusing on the street view perspective of buildings. The images that form our dataset offer various street view perspectives of buildings from urban scenarios, which allows for evaluating complex algorithms. The dataset features 160 images of buildings from Timișoara, Romania, with a resolution of 768 x 1024 pixels each.

Keywords Building dataset · facade detection · edge detection · semantic segmentation · edge detection ground-truth · semantic segmentation ground-truth

1 Introduction

Computer Vision (CV) aims to create computational models that can mimic the human visual system. From an engineering point of view, CV aims to build autonomous systems which could perform some of the tasks that the human visual system is able to accomplish [1].

Urban scenarios reconstruction and understanding of it is an area of research with several applications nowadays: entertainment industry, computer gaming, movie making, digital mapping for mobile devices, digital mapping for car navigation, urban planning, driving. Understanding urban scenarios has become much more important with the evolution of Augmented Reality (AR). AR is successfully exploited in many domains nowadays, one of them being culture and tourism, an area in which the authors of this paper carried multiple research projects [2], [3], [4].

Automatic urban scene object recognition describes the process of segmentation and classification of buildings, trees, cars and so on. This job is done using a fixed number of categories on which a model is trained for classifying scene components [5]. Object detection, recognition and estimation in 3D images have gained momentum due to the availability of more complex sensors and an increase in large scale 3D data. Visual recognition of buildings can be a problematic task due to image distortions, image saturation or obstacles that are blocking the line of sight. The assumption that local shape structures are sufficient to recognise objects and scenes is largely invalid in practice since objects may have a similar shape [6].

In the last decades research in this domain has increased; annually, multiple new approaches and algorithms are presented in literature regarding urban building detection. The variety of solutions used to reach the detection goal can be a combination of any of the following: edge detection algorithms [7], [8], line detection [9], [10], line matching features [11], [12], semantic segmentation and so on. In Figure 1 we present snapshots of steps of a building detection algorithm. All proposed algorithms bring to the table a novel approach to solve corner cases of an existing problem.

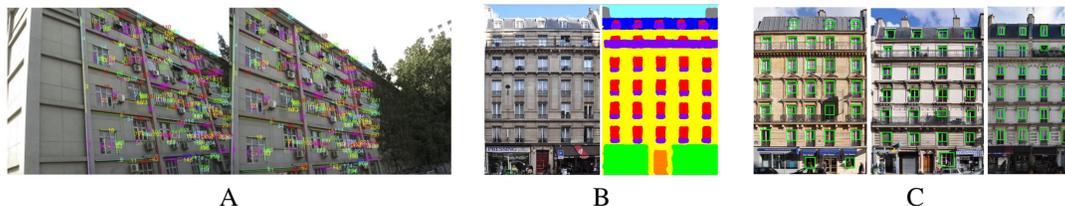


Figure 1: A: Example of line matching algorithm [12]; B: Example of semantic segmentation algorithm [13]; C: Example of window detection algorithm [13]

We believe that this dataset will help enhance the novel algorithms in this domain because of the gap of edges and structural details on buildings images used in other benchmarks. This need occurred when trying to develop algorithms that focus not only on boundaries or contours but on details present in the facade of buildings. For example, in Figure 5 we present an image from the popular BSDS500 [14] dataset, alongside the ground-truth for that image offered by them. In parallel we labeled the same image in our proposed concept of ground-truth. We can observe that the focus of the annotated edges is different: they focus more on boundaries and not on facade details and in our case the other way around. This difference can be an impediment to evaluate correctly the results of an algorithm depending on the scope of it: a general edge detection algorithm will not perform properly if tuned on a building-oriented dataset and of course the other way around.

The Timișoara Building Dataset - TMBuD - [15] is composed of 160 images with the resolution of 768x1024 pixels. Our motivation for this is the belief that this resolution is a good balance between the processing resources needed for manipulating the image and the actual resolution of pictures made with smart devices. Moreover, this is the actual video resolution for filming using a smartphone, the main sensor for building detection systems.

The paper is organized as following: in Section 2 we will present popular existing edge detection datasets with ground-truth and in Section 3 we will present similar semantic annotated datasets. In the end, in Section 4 we will describe our proposed dataset and the issues that we observed that resulted in the need of this new dataset.

2 Edge detection annotated datasets

In this section we present the existing datasets for evaluating edge detection algorithms. Even if edges do not serve as stand alone features in the new CV universe, they still represent a fundamental block for line feature detection.



Figure 2: Examples of images and equivalent ground-truth. Rows: original image, ground-truth; Columns: BSDS500 [16], [14], NYUDV2 [17], MCUE [18], StructED [19]

The Berkeley Segmentation Data Set [16], [14] is one of the most cited paper benchmarks. This benchmark is often used to compare algorithm generated contours or segmentations to human ground-truth data. For the Berkeley database,

1000 representative images of 481x321 RGB images from the Corel image database were chosen. The main criterion for selecting images was that it contains at least one distinctive object [16].

NYU Depth Dataset V2 [17] consists of 1449 RGBD images comprising of commercial and residential buildings in three different cities from US. The image dataset contains 464 different indoor scenes across 26 scene classes. Each image has a dense per-pixel depth labeling using Microsoft Kinect. If a scene contained multiple instances of an object class, each instance received a unique instance label.

The multi-cue boundary detection dataset [18] concerns to study the interaction of several early visual cues (luminance, color, stereo, motion) during boundary detection in challenging natural scenes. They considered a variety of places (from university campuses to street scenes and parks) and seasons to minimize possible biases. The dataset contains 100 scenes, each consisting of a left and right view short (10-frame) color sequence. Each sequence was sampled at a rate of 30 frames per second. Each frame has a resolution of 1280 by 720 pixels.

The Structural Edge dataset [19] propose a new concept of structural edge. Structural edges include occluding contours of objects as well as orientation discontinuities in surfaces are important for understanding the 3D structure of objects and environments. The validity of structural edges was tested using an eye tracking test. The structural edge dataset contains 600 images in natural indoor and outdoor scenes. The structural edges are labeled manually and validated by eye-tracking data from 10 participants with overall 20 trials.

In Figure 2 we present images with the ground-truth from the dataset presented in this section. The mentioned datasets don't focus on certain domain of images of future specific scope to be used. This is a positive point when concerning with a wide range scope algorithm evaluation but is a negative aspect when focusing on a single use case, as we concern ourselves.

3 Semantic Segmentation annotated datasets

In this section we will present existing semantic segmentation datasets that focus on urban scenarios and that would be a good candidate to be used in constructing a building detection algorithm in the end.

The datasets which are selected and used by system designers play a very important role in the quality of the trained model and thereby system performance. So, selecting an appropriate dataset for a task can be one of the most challenging steps at the beginning of the research process [20].

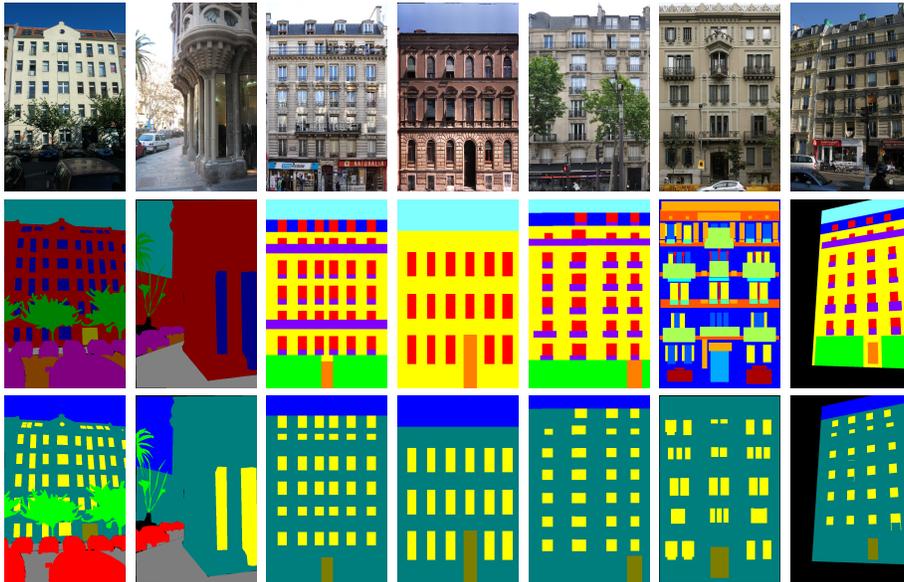


Figure 3: Data sets class correlations. Rows: Original image, Original labels, transition to TMBuD labels; Columns: eTRIMS, LabelMeFacade, ECP, ICG Graz5, INRIA, CPM, VarCity

eTRIMS Image Database [21] is comprised of 60 annotated images and offers two distinct labels: the 4-Class eTRIMS Dataset with 4 annotated object classes and the 8-Class eTRIMS Dataset with 8 annotated object classes. In the 8-class dataset are the following eight object classes: sky, building, window, door, vegetation, car, road, pavement.

LabelMeFacade Database [22] [23] contains 945 images with labeled polygons that describe the different classes. The classes provided are: buildings, windows, sky, and a limited number of unlabeled regions (maximum 20% of the image). The pixelwise labeled images are created by utilizing the eTRIMS categories and a simple depth order heuristic.

Ecole Centrale Paris Facades Database [24] [25] contains 109 images of Paris facades with annotations that have been manually rectified. Classes used for annotation are: window, wall, door, roof, sky, shop.

ICG Graz50 Facade Database [26] is a dataset of rectified facade images and semantic labels that was created with the goal of studying facades. It is comprised of 50 images of various architectural styles (Classicism, Biedermeier, Historicism, Modern and so on).

The Paris Art Deco Facades dataset [27] consists of 80 images of rectified facades of the Art Deco style. The dataset offers 79 RGB images with 6 annotated labels. Occlusions of the facade are ignored but the occlusion reasoning is offered by the dataset.

The CMP Facade Database [28] consists of facade images assembled at the Center for Machine Perception. The dataset includes 606 rectified images of facades from various cities of the world, which have been manually annotated. Annotation is defined as a set of rectangles scope with assigned class labels that can overlap if needed.

VarCity 3D Dataset [29] [30] consists of 700 images along a street annotated with pixel-level labels for facade details. Classes provided are: windows, doors, balconies, roof, etc. The dataset provides images, labels and indexes to the 3D surface together with evaluation source code for comparing different tasks.

In Figure 3 we can observe examples of images and the associated labels that are offered in the dataset presented in this section. As we can see the perspective of each dataset is different. ECP, ICG Graz5, CMP focus more on facade details offering several classes to better understand the facade features. In the VarCity dataset we see that the focus is on the main building discarding the rest of the buildings in the image.

4 Our proposed TMBuD dataset

We intend with our dataset, TMBuD, to unify several ground truth evaluation in the same framework. Of course doing so in a global fashion is close to impossible so we wish to limit the use case to detection, feature extraction and localization of buildings in urban scenarios, based on image understanding.

Building detection is the process of obtaining the approximate position and shape of a building, while building extraction can be defined as the problem of precisely determining the building outlines, which is one of the critical problems in digital photogrammetry [31].

TMBuD is created from images of buildings in Timisoara. Each building is presented from several perspectives, so this dataset can be used for evaluating a building detection algorithm too. The dataset contains ground-truth images for salient edges, for semantic segmentation and the GPS coordinates of the buildings. The dataset contains 160 images grouped in the following sets: 100 consist of the training dataset, 25 consist of the validation data and 35 consist of the test data. We can see examples of images from the dataset in Figure 4.

As we can observe in Figure 4, the database focuses on a view that will be available if the input sensor device is a mobile phone. We consider this to be a very important aspect because the main domain where the building detection algorithm is used is the AR domain. Even if the edge features are focused only in the building area we desire to offer a full understanding of the environment via the semantic segmentation label.

The data was annotated using human subjects that were asked to label (draw) what they perceived as important edges of a building, like the boundaries of the building and differences between facades of the building, different buildings, windows, doors and so on. We asked them not to fill edges or lines that are occluded by other structures even if it's natural that they are present. Secondly they were asked to semantically label the image they created according to the label specification. After this step was finished, we proceeded to unify and correct the edges and labels created by the human subjects into one single ground-truth image. The correction was mainly to eliminate as much as possible the false salient edges that can occur when the data is labeled by a human subject unaware by the inner works of line detection or line matching algorithm.

We believe that a dataset is useful for evaluating algorithms for facade detection or building if it is focused on the building itself present in the image. Firstly, the images should be selected having in mind the street view perspective and uniqueness of features available on them. Secondly, the ground truth images should offer a basis for evaluating boundaries - an important aspect indeed -, but to offer a solution for evaluating facade edges and boundaries.

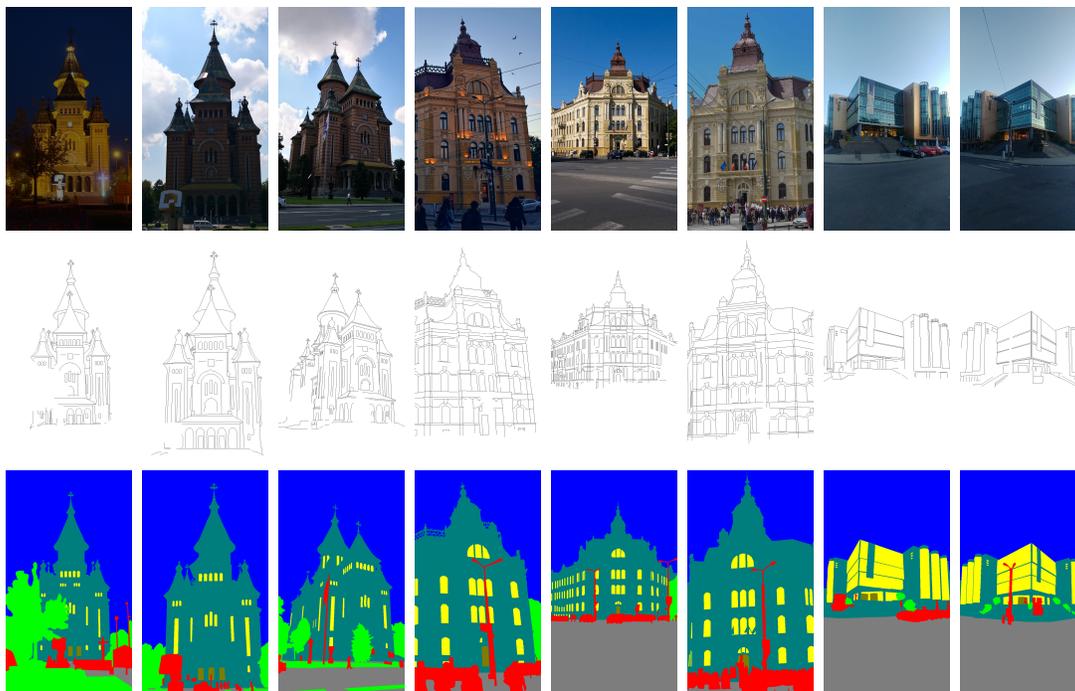


Figure 4: Images from proposed dataset. Rows: original image, edge ground-truth, label ground-truth

Boundary detection and edge detection are similar but not identical. Edges represent discontinuities of brightness which are usually found using low-level CV processes. The process of feature extraction of edges works under the assumption of ideal edge models. Boundary detection is viewed as a mid-level process of finding margins of objects in scenes. This task has close ties both with grouping/segmentation and object shape detection [32].

If we analyse the available datasets and benchmarks for edge detection, as we did in Section 2, we can observe that they focus on edges or boundaries generated from all structures from the image. Of course as we can see in Figure 2 they clearly focus on evaluating and training edge algorithm for natural scenes and do not consider a certain use case of the resulting features.



Figure 5: Images and ground truth

Modern urban building detection techniques like [33], [34] use line segment matching for performing this task. In parallel, the domain of line feature matching [11], [35] for finding relevant features is growing, bringing to the table new solutions for this complex problem. In this scope we consider that our proposed approach for annotating the edges for a dataset becomes more relevant.

As we can observe in Figure 5, we concern ourselves with salient edges produced by the details or shape of the building and ignore the edges produced by adjacent structures in the image, such as persons, cars, sky, ground and so on. We consider that this will help better fine tune the line features extraction algorithms that concern with building detection.

Semantic segmentation is an important aspect in the field of computer vision. The importance of scene understanding is highlighted by the fact that an increasing number of applications emerge from inferring knowledge from imagery. This step in the pipeline has become more popular in object detection applications, even if we talk about building detection.

The proposed dataset focuses more on the scene understanding of the environment rather than on semantically understanding the structures of the building. As we can see in Figure 4, the existing datasets offer solid grounds for training and evaluating semantic segmentation solutions but lack a certain capability to be used to fine tune a semantic segmentation for urban building scenario (as we can see from the last column where we made a transition from there label scheme to ours).

In Table 1 we can observe the existing classes offered by TMBuD, by value and RGB code, and the corresponding classes from the datasets presented in Section 3. Most of the classes are self explanatory but by correlating the classes from TMBuD with other dataset we aim to explain our view for segmenting the environment. We consider it essential to differentiate between BACKGROUND, that we consider unclassified data, and NOISE that we consider elements or objects that appear temporary in the field of view, such as cars, people, terraces, human made temporary structure and so on.

TMBuD	Label	RGB	eTRIMS	LabelMeFacade	ECP	ICG Graz5	INRIA	CPM	VarCity
BACKGROUND	0	(0,0,0)	Not labeled	VARIOUS	OUTLIER	-	VARIOUS	BACKGROUND	BACKGROUND
BUILDING	1	(125,125,0)	BUILDING	BUILDING	WALL BALCONY ROOF CHIMNEY SHOP	WALL	WALL BALCONY ROOF SHOP	FACADE CORNICE SILL BALCONY MOLDING DECO PILLAR SHOP	WALL BALCONY ROOF SHOP
DOOR	2	(0,125,125)	DOOR	DOOR	DOOR	DOOR	DOOR	DOOR	DOOR
WINDOW	3	(0,255,255)	WINDOW	WINDOW	WINDOW	WINDOW BLIND	WINDOW	WINDOW	WINDOW
SKY	4	(255,0,0)	SKY	SKY	SKY	SKY	SKY	BACKGROUND	SKY
VEGETATION	5	(0,255,0)	VEGETATION	VEGETATION	-	-	-	-	-
GROUND	6	(125,125,125)	PAVEMENT ROAD	PAVEMENT ROAD	-	-	-	-	-
NOISE	7	(0,0,255)	CAR	CAR	-	-	-	-	-

Table 1: Dataset’s correlations

The TMBuD does not offer a build-in benchmarking capability of edge detection or semantic segmentation but it is part of EECVF [36] [37], our Python-based End-To-End CV Framework, where a user can evaluate capabilities of algorithms. The dataset offers the possibility of extending or reorganizing the image in the train - validate - test groups by using a Python module that exists in the repository.

5 Conclusion

In this paper we presented a review of existing boundaries and edges dataset and a review of existing semantic segmentation dataset with the scope of highlighting current evaluation solutions. Afterwards we proposed a dataset that is better fitted to serve the tuning and evaluation of urban scenario building detection algorithms.

We believe that the proposed TMBuD dataset can facilitate research in image processing when focusing on urban scenarios. TMBuD has two main benefits: the unified evaluating system for several linked problems from this area and the targeted dataset on human made structures in urban scenarios. Both aspects mentioned can become relevant aspects for future development and research work.

TMBuD has proven to be a useful dataset for evaluation when trying to determine the best fitted edge detection variant or the best fitted semantic segmentation model for urban scenarios [37] [38]. From the experience of our work we consider that the proposed dataset as an useful component in constructing a content based image retrieval urban building systems.

We want to expand in the near future the quantity of the dataset images and ground truth images respecting the same principles that we exposed in the paper: important human made structures in urban areas, from a street perspective and different angles.

Regarding the expansion of the dataset we are thinking of including a series of metadata information to be available for each landmark so we can serve algorithms focused on classifying buildings according to facts like: age of the building, architecture style.

References

- [1] T. Huang, “Computer vision: Evolution and promise,” 1996.
- [2] S. Vert and R. Vasiu, “Relevant aspects for the integration of linked data in mobile augmented reality applications for tourism,” in *International Conference on Information and Software Technologies*, pp. 334–345, Springer, 2014.
- [3] S. Vert and R. Vasiu, “Augmented reality lenses for smart city data: the case of building permits,” in *World Conference on Information Systems and Technologies*, pp. 521–527, Springer, 2017.
- [4] S. Vert, D. Andone, and R. Vasiu, “Augmented and virtual reality for public space art,” in *ITM Web of Conferences*, vol. 29, p. 03006, EDP Sciences, 2019.
- [5] P. Babahajiani, L. Fan, and M. Gabbouj, “Object recognition in 3d point cloud of urban street scene,” in *Asian conference on computer vision*, pp. 177–190, Springer, 2014.
- [6] J. Fu, J.-K. Kämäräinen, A. G. Buch, and N. Krüger, “Indoor objects and outdoor urban scenes recognition by 3d visual primitives,” in *Asian Conference on Computer Vision*, pp. 270–285, Springer, 2014.
- [7] C. Orhei, S. Vert, and R. Vasiu, “A novel edge detection operator for identifying buildings in augmented reality applications,” pp. 208–219, 2020.
- [8] C. Topal and C. Akinlar, “Edge drawing: a combined real-time edge and segment detector,” *Journal of Visual Communication and Image Representation*, vol. 23, no. 6, pp. 862–872, 2012.
- [9] R. G. Von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, “Lsd: A fast line segment detector with a false detection control,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 4, pp. 722–732, 2008.
- [10] C. Akinlar and C. Topal, “Edlines: Real-time line segment detection by edge drawing (ed),” in *2011 18th IEEE International Conference on Image Processing*, pp. 2837–2840, IEEE, 2011.
- [11] L. Zhang and R. Koch, “An efficient and robust line segment matching approach based on lbd descriptor and pairwise geometric consistency,” *Journal of Visual Communication and Image Representation*, vol. 24, no. 7, pp. 794–805, 2013.
- [12] B. Fan, F. Wu, and Z. Hu, “Robust line matching through line–point invariants,” *Pattern Recognition*, vol. 45, no. 2, pp. 794–805, 2012.
- [13] H. Liu, J. Zhang, J. Zhu, and S. C. Hoi, “Deepfacade: A deep learning approach to facade parsing,” 2017.
- [14] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2010.
- [15] “CM Building Dataset Timisoara.” <https://github.com/CipiOrhei/TMBuD>. Accessed: 2021-03-12.
- [16] D. Martin, C. Fowlkes, D. Tal, and J. Malik, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, pp. 416–423, IEEE, 2001.
- [17] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, “Indoor segmentation and support inference from rgb-d images,” in *ECCV*, 2012.
- [18] D. A. Mély, J. Kim, M. McGill, Y. Guo, and T. Serre, “A systematic comparison between visual cues for boundary detection,” *Vision research*, vol. 120, pp. 93–107, 2016.
- [19] W. Sun, S. You, J. Walker, K. Li, and N. Barnes, “Structural edge detection: A dataset and benchmark,” in *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8, 2018.
- [20] A. Zlateski, R. Jaroensri, P. Sharma, and F. Durand, “On the importance of label quality for semantic segmentation,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1479–1487, 2018.
- [21] F. Korc and W. Förstner, “etrimis image database for interpreting images of man-made scenes,” *Dept. of Photogrammetry, University of Bonn, Tech. Rep. TR-IGG-P-2009-01*, 2009.
- [22] B. Fröhlich, E. Rodner, and J. Denzler, “A fast approach for pixelwise labeling of facade images,” in *Proceedings of the International Conference on Pattern Recognition (ICPR 2010)*, 2010.
- [23] C.-A. Brust, S. Sickert, M. Simon, E. Rodner, and J. Denzler, “Efficient convolutional patch networks for scene understanding,” in *CVPR Workshop on Scene Understanding (CVPR-WS)*, 2015.
- [24] O. Teboul, L. Simon, P. Koutsourakis, and N. Paragios, “Segmentation of building facades using procedural shape priors,” *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3105–3112, 2010.

- [25] O. Teboul, I. Kokkinos, L. Simon, P. Koutsourakis, and N. Paragios, “Shape grammar parsing via reinforcement learning,” *CVPR 2011*, pp. 2273–2280, 2011.
- [26] H. Riemenschneider, U. Krispel, W. Thaller, M. Donoser, S. Havemann, D. Fellner, and H. Bischof, “Irregular lattices for complex shape grammar facade parsing,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1640–1647, IEEE, 2012.
- [27] R. Gadde, R. Marlet, and N. Paragios, “Learning grammars for architecture-specific facade parsing,” *International Journal of Computer Vision*, vol. 117, no. 3, pp. 290–316, 2016.
- [28] R. Tyleček and R. Šára, “Spatial pattern templates for recognition of objects with regular structure,” in *Proc. GCPR*, (Saarbrücken, Germany), 2013.
- [29] H. Riemenschneider, A. Bódis-Szomorú, J. Weissenberg, and L. Gool, “Learning where to classify in multi-view semantic segmentation,” in *ECCV*, 2014.
- [30] A. Martinovic, J. Knopp, H. Riemenschneider, and L. Gool, “3d all the way: Semantic segmentation of urban scenes from start to end in 3d,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4456–4465, 2015.
- [31] A. R. Elshehaby and L. G. E.-d. Taha, “A new expert system module for building detection in urban areas using spectral information and lidar data,” *Applied Geomatics*, vol. 1, no. 4, pp. 97–110, 2009.
- [32] X. Ren, “Multi-scale improves boundary detection in natural images,” in *European conference on computer vision*, pp. 533–545, Springer, 2008.
- [33] Y. Yi, Z. Zhang, W. Zhang, C. Zhang, W. Li, and T. Zhao, “Semantic segmentation of urban buildings from vhr remote sensing imagery using a deep convolutional neural network,” *Remote Sensing*, vol. 11, no. 15, p. 1774, 2019.
- [34] W. Wang, W. Gao, H. Cui, and Z. Hu, “Reconstruction of lines and planes of urban buildings with angle regularization,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 165, pp. 54–66, 2020.
- [35] K. Li, J. Yao, M. Xia, and L. Li, “Joint point and line segment matching on wide-baseline stereo images,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, IEEE, 2016.
- [36] C. Orhei, M. Mocofan, S. Vert, and R. Vasiiu, “End-to-end computer vision framework,” in *2020 International Symposium on Electronics and Telecommunications (ISETC)*, pp. 1–4, IEEE, 2020.
- [37] C. Orhei, S. Vert, M. Mocofan, and R. Vasiiu, “End-to-end computer vision framework: An open-source platform for research and education,” *Sensors*, vol. 21, no. 11, p. 3691, 2021.
- [38] C. Orhei, M. Mocofan, S. Vert, and R. Vasiiu, “An analysis of ed line algorithm in urban street-view dataset,” in *International Conference on Information and Software Technologies*, pp. –, Springer, 2021.

Appendix A



Figure 6: Images from proposed dataset. Rows: original image, edge ground-truth, label ground-truth