MULTI-TASK LEARNING FROM FIXED-WING UAV IMAGES FOR 2D/3D CITY MODELLING

M. R. Bayanlou [a, *, α], M. Khoshboresh-Masouleh [b, α]

[a] Aerospace Engineering Department, Sharif University of Technology, Tehran, Iran - mohammadreza.bayanlou@ae.sharif.edu
[b] School of Surveying and Geospatial Eng., College of Eng., University of Tehran, Tehran, Iran - m.khoshboresh@ut.ac.ir

KEYWORDS: Multi-Task Learning, 2D/3D City Modelling, Fixed-Wing UAV, SAMA-VTOL

ABSTRACT:

Single-task learning in artificial neural networks will be able to learn the model very well, and the benefits brought by transferring knowledge thus become limited. In this regard, when the number of tasks increases (e.g., semantic segmentation, panoptic segmentation, monocular depth estimation, and 3D point cloud), duplicate information may exist across tasks, and the improvement becomes less significant. Multi-task learning has emerged as a solution to knowledge-transfer issues and is an approach to scene understanding which involves multiple related tasks each with potentially limited training data. Multi-task learning improves generalization by leveraging the domain-specific information contained in the training data of related tasks. In urban management applications such as infrastructure development, traffic monitoring, smart 3D cities, and change detection, automated multi-task data analysis for scene understanding based on the semantic, instance, and panoptic annotation, as well as monocular depth estimation, is required to generate precise urban models. In this study, a common framework for the performance assessment of multi-task learning methods from fixed-wing UAV images for 2D/3D city modelling is presented.

1. INTRODUCTION

1.1 Motivation

In recent years, the role of traditional methods such as terrestrial mapping and traditional aerial photogrammetry techniques has been dimmed due to the high cost and also the need for a long time to generate a multi-task dataset for scene understanding (Crawshaw, 2020; Khoshboresh Masouleh and Shah-Hosseini, 2020; Masouleh and Sadeghian, 2019; Ruder, 2017; Zhang and Yang, 2018). An affordable and accurate way to generate multi-task data is to use the combination of an Unmanned Aerial Vehicle (UAV) with a high-resolution digital camera (e.g., RGB, Multi-spectral, Thermal, or Hyperspectral) and machine learning methods (Bayanlou and Khoshboresh-Masouleh, 2020; Khoshboresh-Masouleh and Hasanlou, 2020). Although UAV with a high-resolution digital camera is an efficient tool for data generation, there is still a lack of multi-task datasets for scene understanding (Khoshboresh Masouleh and Shah-Hosseini, 2019). With this study, we aim at promoting research on multi-task learning by generating a very high-resolution low-cost dataset of urban and rural objects (e.g., building, parcel boundary, vehicle, building shadow, vegetation, ground, waste object, farmland, and water). In this study, we focus on multi-task learning based on semantic segmentation, building panoptic segmentation, and monocular depth estimation.

A major yet unsolved research topic for accurate 2D/3D city model generation is multi-task learning for scene understanding from high-resolution low-cost photogrammetry and remote sensing data sources (Khoshboresh Masouleh and Saradjian, 2019). In remote sensing and photogrammetry, previous benchmarks include four semantic segmentation datasets designed using satellite, airborne, and UAV platforms for urban scene analysis. We believe all are important datasets for urban scene analysis, but our proposed dataset will be comprised of much larger multi-task data and with more scene complexity regarding the number of objects.

1.2 Related Works

The ISPRS 2D semantic labeling benchmark provides Vaihingen and the Potsdam datasets targeting semantic labeling for the urban scenes. The Vaihingen and the Potsdam datasets are 9 cm and 5 cm resolutions, respectively. There are 6 classes defined for the semantic segmentation task, including impervious surfaces, buildings, low vegetation, tree, car, and background. The ISPRS 2D semantic labeling benchmarks include a set of homogenous scenes from one spatial location, and most deep learning-based methods achieve high accuracy using these kinds of datasets. In recent years, many models, e.g. (Masouleh and Shah-Hosseini, 2018) have achieved high accuracy (mostly above 90%) on these test data (hereafter called Dataset 1 for short).

The ISPRS UAV Semantic Video Segmentation benchmark provides a very high-resolution video dataset targeting semantic labeling for urban scene analysis from an oblique UAV perspective. There are 8 classes defined for this dataset, including building, road, tree, low vegetation, static car, moving car, human, and background (Lyu et al., 2020). This dataset consists of only RGB images and is not suitable for multi-task learning (hereafter called Dataset 2 for short).

The ISPRS Benchmark Challenge on Large Scale Classification of VHR Geospatial Data provides a multispectral high-resolution satellite dataset targeting semantic labeling for urban scene analysis from two Worldview-II satellite images. There are 6 classes defined for this dataset, including impervious surface, building, pervious surface, high vegetation, cars, and water. Moreover, buildings are annotated as single objects for semantic instance segmentation (Roscher et al., 2020). This dataset consists of only remote sensing optical images and is not suitable for multi-task learning (hereafter called Dataset 3 for short).

In contrast to these datasets, our proposed dataset will be comprised of much larger multi-task data (e.g., semantic segmentation, panoptic segmentation, monocular depth

* Corresponding author.
α Authors with same contributions.

estimation, and 3D point cloud) and with more scene complexity in terms of the number of objects (e.g., parcel boundary and building shadow), which makes our dataset more adequate for multi-task learning for scene understanding from UAV images. An overview of existing datasets of annotated imagery can be found in Table 1.

Reference	Dataset 1	Dataset 2	Dataset 3	Ours
Data source	Airborne	UAV	Satellite (WV-2)	UAV
Type	Orthophoto	Video	Nadir view	Orthophoto
DSM/nDSM	High-resolution	No	No	Very high-resolution
Texture distortion	High	-	-	Low (cf. Figure 3)
Semantic annotation	Yes	Yes	Yes	Yes
Semantic classes	6	8	6	8-10
Panoptic annotation	No	No	No (instance)	Yes (just buildings)
Color-based 3D point cloud	No	No	No	Yes
Image size (pix)	6000×6000	4096×2160	3452×3504	2000×2000
GSD	9cm/5cm	-	50cm	2.5cm
Is this a multi-task dataset?	Yes (low potential), including orthophoto, nDSM, and semantic annotation	No	No	Yes (high potential), including orthophoto, DSM, semantic, and building panoptic annotation
Number of labeled multi-task pixels	2014: 1.7 million 2015: 1.4 billion	-	-	2 billion

Table 1. List of the previous datasets and the proposed dataset for 2D/3D urban scene analysis

2. SINGLE-TASK VS. MULTI-TASK LEARNING

Single-task learning in deep neural networks (LeCun et al., 2015) will be able to learn the model very well, and the benefits brought by transferring knowledge thus become limited. In this regard, when the number of tasks increases (e.g., semantic segmentation, panoptic segmentation, and monocular depth estimation), duplicate information may exist across tasks, and the improvement becomes less significant (cf. Figure 1a). Multi-task learning has emerged as a solution to knowledge-transfer issues (Cipolla et al., 2018). Multi-task learning is an approach to scene understanding which involves multiple related tasks each with potentially limited training data (cf. Figure 1b). Multi-task learning improves generalization by leveraging the domain-specific information contained in the training data of related tasks. In urban management applications such as infrastructure development, traffic monitoring, smart 3D cities, and change detection, automated multi-task data analysis for scene understanding based on the semantic, instance, and panoptic annotation, as well as monocular depth estimation, is required to generate precise urban models (Khoshboresh-Masouleh et al., 2020; Khoshboresh-Masouleh and Shah-Hosseini, 2020).

3. DATA COLLECTION

SAMA-VTOL aerial image dataset is a new UAV-based image dataset for a wide range of scientific projects in remote sensing (e.g., 3D object modelling, rural/urban mapping, and digital elevation/surface model processing). High-quality UAV images play an important part in providing and expanding spatial data processing methods (Herwitz et al., 2004; Ishida et al., 2018; Senthilnath et al., 2017). This dataset includes 120 rural/urban scene images with 80% overlap between images (forward overlap) and 60% overlap between flight lines (side overlap) from part of Esfahan province, Iran. The characteristics that make the proposed dataset an excellent scientific dataset are: (i) very high ground sampling distance (GSD) due to suitable fly height selection; (ii) GNSS-PPK (Post Processing Kinematic) system for improving the spatial accuracy without ground control points (GCPs); (iii) various landscape types (e.g., different types of roofs for commercial/residential buildings, and vegetation), and (iv) uses of the new UAV-photogrammetry platform, named SAMA-VTOL (2019) has been developed by TAREQH Corporation (Bayanlou and Khoshboresh-Masouleh, 2020).

3.1 UAV Images

The original UAV RGB images were captured by SAMA-VTOL are provided for the case study. SAMA-VTOL is designed to boost efficiency, safety, and quality for high-resolution low-cost data collection. This dataset consists of 120 rural/urban scene images with 80% forward overlap and 60% side overlap, where the proposed dataset uses the WGS 84 (EPSG::4326) coordinate system, as do most GNSS units. Figure 2 shows the study site in the various landscape types of images collected from Esfahan province. The research site is part of the Esfahan province, Iran. The land cover consists of agricultural land and urban areas.

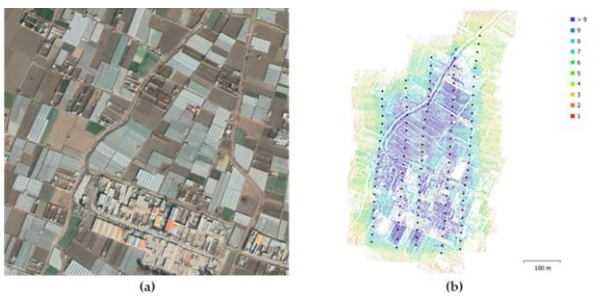

Figure 2. Google Earth imagery of the study area. (a) Research site; (b) Camera locations and image overlap

In this study, SAMA-VTOL was equipped with a Fujifilm X-A3 camera to acquire images. Additionally, the Agisoft Metashape software was used to analysing images and produce dense point clouds, digital surface model (DSM), and orthoimage for evaluating quality and quantity, and QGroundControl software was used for mission planning and flight control.

3.2 Data Processing

The data processing includes automatic aerial triangulation-based bundle block adjustment with camera calibration and model generation by Agisoft Metashape. Agisoft Metashape is a stand-alone software product that performs photogrammetric processing of images and generates 2D/3D spatial data to be used in real-world applications (Agisoft Metashape, 2021).

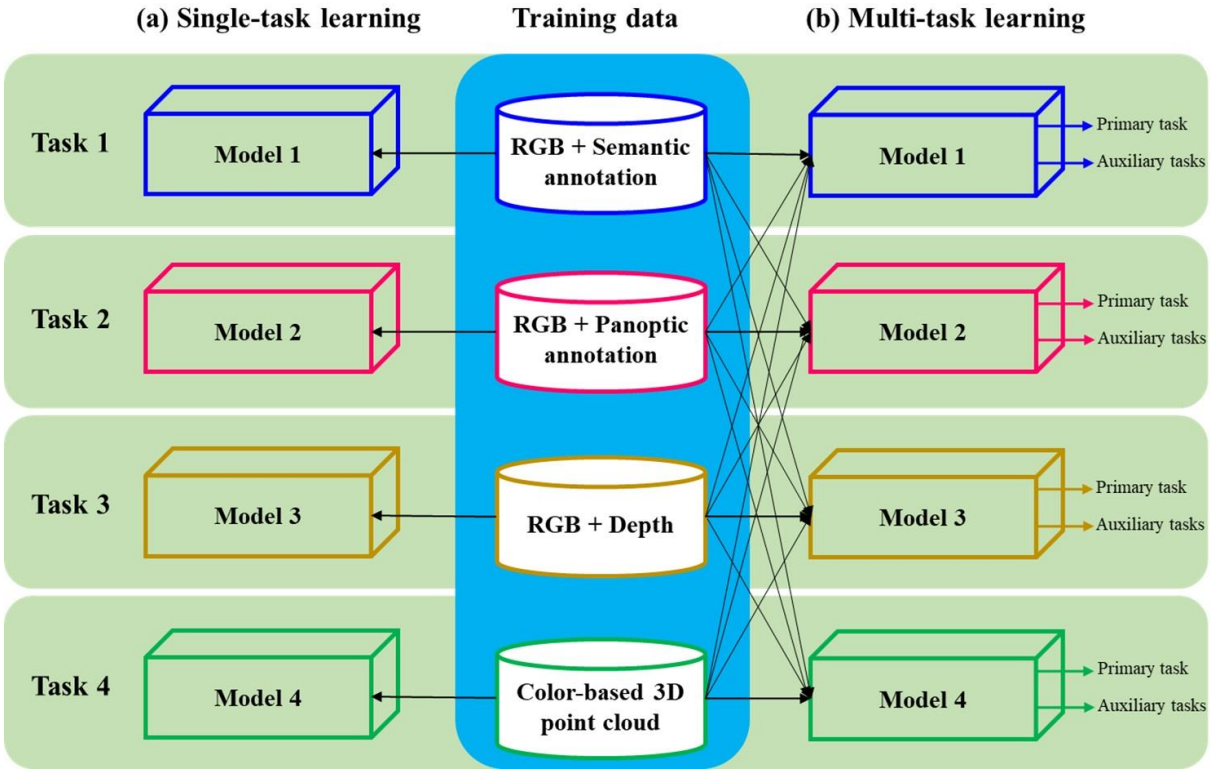

Figure 1. Difference between (a) single-task learning, and (b) multi-task learning

4. RESULTS

This study aims to promote and strengthen the research on multi-task learning for scene understanding from fixed-wing UAV images. The proposed method is composed of two components of UAV-based multi-task data generation and evaluation methods in urban scene analysis. The contributions of this study are as follows.

(1) Development of a multi-task dataset for scene understanding, including very high-resolution RGB orthophoto, digital surface model, semantic annotation, building panoptic annotation (cf. Figure 3).

(2) An evaluation is presented based on the most important challenges for multi-task data analysis for scene understanding.

In this study, we will build a large multi-task dataset for urban scene analysis. The study area is located in Esfahan, Iran. The study region includes buildings with rectangular flat roofs which may also have various tiny structures. Flat roof building is a specific style of urban architecture common in different cities (e.g., New York, Cairo, Tehran, and Esfahan). The planned specifics of the dataset are listed in Table 2. Moreover, Figure 4 shows the study site in the various landscape types with three samples of datasets collected.

GSD	Image size	Tile	Semantic classes		Panoptic class	DSM
2.5cm	2000×2000pix	500	(1) Parcel boundary (2) Vehicle (3) Building shadow (4) Vegetation (5) Building	(6) Ground/Road (7) Waterbodies (8) Farmland (9) Waste object (10) Lane-marking	Building	Yes

Table 2. Detail of the proposed multi-task dataset

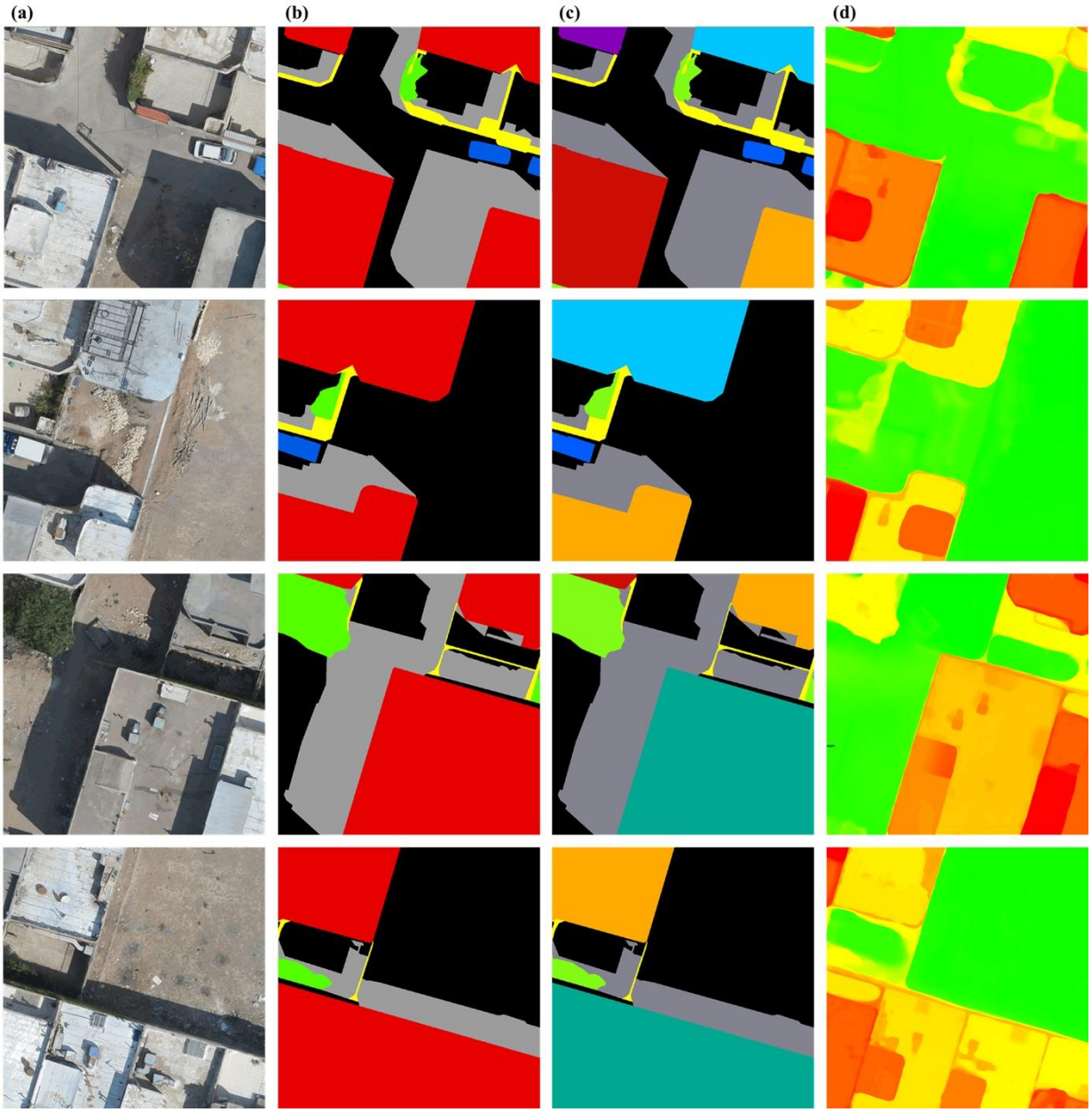

Figure 3. Sample screenshot of the proposed dataset for scene understanding with (a) RGB orthophoto, (b) semantic segmentation annotation, (c) building panoptic annotation, and (d) DSM

Figure 4. Various landscape types in the proposed dataset: (a) water and vegetation covers, (b) waste objects, and (c) different color roofs

5. CONCLUSIONS

In this study, a common framework and a new dataset for the performance assessment of multi-task learning methods from fixed-wing UAV images for 2D/3D city modelling are presented. The UAV RGB images were captured by SAMA-VTOL are provided for the case study. This fixed-wing UAV has an installed very high-resolution Fujifilm sensor that was intended for professional photogrammetry. This UAV can cover 100 ha during one flight and the flight time is 60 min. The results of the experiments in the test area indicate that the SAMA-VTOL is robust to multi-task data generation.

REFERENCES

Agisoft Metashape, 2021. Agisoft Metashape [WWW Document]. URL https://www.agisoft.com/ (accessed 4.3.21).

Bayanlou, M.R.R., Khoshboresh-Masouleh, M., 2020. SAMA-VTOL: A new unmanned aircraft system for remotely sensed data collection, in: SPIE Future Sensing Technologies. Presented at the SPIE Future Sensing Technologies, International Society for Optics and Photonics, p. 115250V. https://doi.org/10.1117/12.2580533

Cipolla, R., Gal, Y., Kendall, A., 2018. Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Salt Lake City, UT, USA, pp. 7482–7491. https://doi.org/10.1109/CVPR.2018.00781

Crawshaw, M., 2020. Multi-Task Learning with Deep Neural Networks: A Survey. arXiv:2009.09796 [cs, stat].

Herwitz, S.R., Johnson, L.F., Dunagan, S.E., Higgins, R.G., Sullivan, D.V., Zheng, J., Lobitz, B.M., Leung, J.G., Gallmeyer, B.A., Aoyagi, M., Slye, R.E., Brass, J.A., 2004. Imaging from an unmanned aerial vehicle: agricultural surveillance and decision support. Computers and Electronics in Agriculture 44, 49–61. https://doi.org/10.1016/j.compag.2004.02.006

Ishida, T., Kurihara, J., Viray, F.A., Namuco, S.B., Paringit, E.C., Perez, G.J., Takahashi, Y., Marciano, J.J., 2018. A novel approach for vegetation classification using UAV-based hyperspectral imaging. Computers and Electronics in Agriculture 144, 80–85. https://doi.org/10.1016/j.compag.2017.11.027

Khoshboresh Masouleh, M., Saradjian, M.R., 2019. ROBUST BUILDING FOOTPRINT EXTRACTION FROM BIG MULTI-SENSOR DATA USING DEEP COMPETITION NETWORK, in: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Presented at the ISPRS WG IV/3, WG I/8 & WG II/4
ISPRS International GeoSpatial Conference 2019, Joint Conferences of 5th Sensors and Models in Photogrammetry and Remote Sensing (SMPR) and 3rd Geospatial Information Research (GI Research) (Volume XLII-4/W18) - 12–14 October 2019, Karaj, Iran, Copernicus GmbH, pp. 615–621. https://doi.org/10.5194/isprs-archives-XLII-4-W18-615-2019

Khoshboresh Masouleh, M., Shah-Hosseini, R., 2020. A hybrid deep learning–based model for automatic car extraction from high-resolution airborne imagery. Appl Geomat 12, 107–119. https://doi.org/10.1007/s12518-019-00285-4

Khoshboresh Masouleh, M., Shah-Hosseini, R., 2019. Development and evaluation of a deep learning model for real-time ground vehicle semantic segmentation from UAV-based thermal infrared imagery. ISPRS Journal of Photogrammetry and Remote Sensing 155, 172–186. https://doi.org/10.1016/j.isprsjprs.2019.07.009

Khoshboresh-Masouleh, M., Alidoost, F., Arefi, H., 2020. Multiscale building segmentation based on deep learning for remote sensing RGB images from different sensors. JARS 14, 034503. https://doi.org/10.1117/1.JRS.14.034503

Khoshboresh-Masouleh, M., Hasanlou, M., 2020. Improving hyperspectral sub-pixel target detection in multiple target signatures using a revised replacement signal model. European Journal of Remote Sensing 53, 316–330. https://doi.org/10.1080/22797254.2020.1850179

Khoshboresh-Masouleh, M., Shah-Hosseini, R., 2020. A Deep Learning Method for Near-Real-Time Cloud and Cloud Shadow Segmentation from Gaofen-1 Images [WWW Document]. Computational Intelligence and Neuroscience. https://doi.org/10.1155/2020/8811630

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Lyu, Y., Vosselman, G., Xia, G.-S., Yilmaz, A., Yang, M.Y., 2020. UAVid: A semantic segmentation dataset for UAV imagery. ISPRS Journal of Photogrammetry and Remote Sensing 165, 108–119. https://doi.org/10.1016/j.isprsjprs.2020.05.009

Masouleh, M.K., Sadeghian, S., 2019. Deep learning-based method for reconstructing three-dimensional building cadastre models from aerial images. JARS 13, 024508. https://doi.org/10.1117/1.JRS.13.024508

Masouleh, M.K., Shah-Hosseini, R., 2018. Fusion of deep learning with adaptive bilateral filter for building outline extraction from remote sensing imagery. JARS 12, 046018. https://doi.org/10.1117/1.JRS.12.046018

Roscher, R., Volpi, M., Mallet, C., Drees, L., Wegner, J.D., 2020. SEMCITY TOULOUSE: A BENCHMARK FOR BUILDING INSTANCE SEGMENTATION IN SATELLITE IMAGES. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. V-5–2020, 109–116. https://doi.org/10.5194/isprs-annals-V-5-2020-109-2020

Ruder, S., 2017. An Overview of Multi-Task Learning in Deep Neural Networks. arXiv:1706.05098 [cs, stat].

Senthilnath, J., Kandukuri, M., Dokania, A., Ramesh, K.N., 2017. Application of UAV imaging platform for vegetation analysis based on spectral-spatial methods. Computers and Electronics in Agriculture 140, 8–24. https://doi.org/10.1016/j.compag.2017.05.027

Zhang, Y., Yang, Q., 2018. A Survey on Multi-Task Learning. arXiv:1707.08114 [cs].