Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Leveraging OpenStreetMap and Multimodal Remote Sensing Data with Joint Deep Learning for Wastewater Treatment Plants Detection

Hao Li\(^{a,}\)*, Johannes Zech\(^{a}\), Danfeng Hong\(^{b}\), Pedram Ghamisi\(^{c,d}\), Michael Schultz\(^{a}\), Alexander Zipf\(^{a,}\)

\(^{a}\) GI Science Chair, Institute of Geography, Heidelberg University, 69120 Heidelberg, Germany
\(^{b}\) Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
\(^{c}\) Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Machine Learning Group, D-09599 Freiberg, Saxony, Germany
\(^{d}\) Institute of Advanced Research in Artificial Intelligence (IARAI), Landstraßer Hauptstraße 5, 1030 Vienna, Austria

\(^{*}\) Corresponding author.
E-mail addresses: hao.li@uni-heidelberg.de (H. Li), johannes.zech@alumni.uni-heidelberg.de (J. Zech), hongdf@aircas.ac.cn (D. Hong), p.ghamisi@gmail.com (P. Ghamisi), michael.schultz@uni-heidelberg.de (M. Schultz), zipf@uni-heidelberg.de (A. Zipf).

A B S T R A C T

Humans rely on clean water for their health, well-being, and various socio-economic activities. During the past few years, the COVID-19 pandemic has been a constant reminder of about the importance of hygiene and sanitation for public health. The most common approach to securing clean water supplies for this purpose is via wastewater treatment. To date, an effective method of detecting wastewater treatment plants (WWTP) accurately and automatically via remote sensing is unavailable. In this paper, we provide a solution to this task by proposing a novel joint deep learning (JDL) method that consists of a fine-tuned object detection network and a multi-task residual attention network (RAN). By leveraging OpenStreetMap (OSM) and multimodal remote sensing (RS) data, our JDL method is able to simultaneously tackle two different tasks: land use land cover (LULC) and WWTP classification. Moreover, JDL exploits the complementary effects between these tasks for a performance gain. We train JDL using 4,187 WWTP features and 4,200 LULC samples and validate the performance of the proposed method over a selected area around Stuttgart with 723 WWTP features and 1,200 LULC samples to generate an LULC classification map and a WWTP detection map. Extensive experiments conducted with different comparative methods demonstrate the effectiveness and efficiency of our JDL method in automatic WWTP detection in comparison with single-modality/single-task or traditional survey methods. Moreover, lessons learned pave the way for future works to simultaneously and effectively address multiple large-scale mapping tasks (e.g., both mapping LULC and detecting WWTP) from multimodal RS data via deep learning.

1. Introduction

Clean water plays a key role in human health, well-being, socio-economic activities, and Sustainable Development Goals (SDGs) (U.S. Environmental Protection Agency, 2016; Persello et al., 2022). Nevertheless, one-third of the global population still cannot access safe drinking water with two out of five people do not have basic hand-washing sanitation (Organization, 2019). In the past two years, the COVID-19 pandemic has demonstrated the substantial importance of hygiene rules, sanitation, and adequate access to clean water for reducing the spread of infectious diseases and preserving the public health of millions. Past studies have shown that the urban poor living in slums can be even more vulnerable to the impacts of COVID-19, given their urgent need for adequate clean water, sanitation, and hygiene facilities (UN news, 2020). To keep sanitized as claimed in SDG 6, wastewater treatment is the most common approach of removing chemical and biological contaminants from human-consumed wastewater and ensuring consistently clean water supply (Tchobanoglous et al., 2003). Therefore, an accurate, up-to-date map of wastewater treatment plants (WWTP) is vital to monitoring the accessibility of sanitation facilities and estimating the potential need for a clean water supply. However, such maps are often either nonexistent or costly to produce.

In recent years, tremendous efforts have been made to map and detect geospatial objects from remote sensing (RS) data (Li et al., 2020),...
where the target objects ranged from common human settlements (Esch et al., 2013) and roads (Mnih and Hinton, 2012) to fine-grained features (e.g., airplane (Zheng et al., 2020) and vehicles (Wu et al., 2020a)). Meanwhile, the rapid development of RS imaging techniques offered an ever-growing list of RS data sources acquired by distinct platforms in various spatial-temporal resolutions, including very high resolution (VHR) images focusing on optical textures and shapes, multispectral images (MSI) and hyperspectral images (HSI) providing the detailed information of object materials and contents (Li et al., 2018; Hong et al., 2021; Hong et al., 2022), as well as light detection and ranging (LiDAR), measuring height information about ground elevation (Itel et al., 2016; Huang et al., 2020; Xu et al., 2020). In this context, the benefit of joint exploitation of multisensor and multimodal RS data sources for more accurate geospatial object detection and LULC mapping (Schmitt and Zhu, 2016; Ghamisi et al., 2019; Salcedo-Sanz et al., 2020) is clear. For example, VHR images are capable of capturing spatial information (i.e., shape and size) of buildings but supplies little information about roof materials, while MSI and HSI can provide more discriminative spectral information in distinguishing buildings of different roof types (e.g., wood or asphalt). Recent work of Hong et al. (2020) presented an inspiring work in this direction and identified the huge potential of multimodal RS data fusion. However, existing works of geospatial object detection were designed either for relatively common objects (e.g., buildings or vehicles) that appear with high frequency in geographical scenes, or from a single modality of RS data. To date, no effective approach has been proposed for the accurate detection of sparse but critical civil infrastructures like WWTP, as geospatial objects from multimodal RS data sources.

Among several factors, the lack of both high-quality training samples and novel joint learning approaches were identified as major challenges in effective deep learning from multimodal RS data at a large scale (Ngiam et al., 2011; Ma et al., 2019; Hong et al., 2020). Regarding the former challenge, fortunately, Volunteered Geographical Information (VGI) (Goodchild, 2007), and specifically the OpenStreetMap (OSM), was recently explored, making use of its rich semantic information (e.g., OSM tag and value) to extract customized geospatial objects, as well as generating geo-referenced training samples, in order to develop effective geospatial object detection models (Chen and Zipf, 2017; Vargas-Munoz et al., 2020). Early attempts in Herfort et al. (2019), Li et al. (2020) reported promising findings of deep learning from VGI, especially OSM data, for building detection, and achieved competitive mapping accuracies compared to the crowdsourcing approach. OSM data has shown great potential, offering a massive and freely available source of human-labeled features as training data for basic geospatial objects (e.g., buildings and roads). However, a more general approach of harvesting OSM data as training data for specific and even sparse geospatial objects (e.g., WWTP) is still needed. As for the latter challenge, given the rapid advances of geospatial artificial intelligence (GeoAI) models and methods (Zhu et al., 2017; Werner et al., 2021), object detection from RS data with deep learning has attracted substantial attention in both academia (Wu et al., 2021) and industrial communities (Sirko et al., 2021). Despite current achievements, however, determining how best to extend the state-of-the-art deep learning models for effective joint learning from multimodal RS data sources in an end-to-end manner remains an open topic.

Inspired by existing works, our research focuses on exploring the potential of leveraging OSM and multimodal RS data (i.e., VHR images and MSI) to detect one of the most critical civil infrastructures, namely WWTP, by proposing a novel joint deep learning (JDL) method (see Fig. 1). In this paper, we delineate this JDL method, which consists of two major networks: a fine-tuned object detection network and, more importantly, a multi-task residual attention network (RAN) to simultaneously tackle land use land cover (LULC) classification and WWTP detection. In addition, the proposed method is able to extract OSM training samples for both deep learning tasks in a fully automatic manner, which ensures its applicability and robustness in real-world map productions. We validate our JDL method by training in multiple federal states within Germany and testing in a selected region around Stuttgart. The LULC and WWTP maps produced are further evaluated against a manually-labeled reference dataset for accuracy assessment. Briefly, the research questions (RQs) we address in this paper are twofold:

- (RQ1): How to jointly explore multimodal RS data with OSM training samples for a multi-task learning purpose?
- (RQ2): How accurately can our JDL method detect large-scale WWTP with the support of LULC information?

The remainder of this paper is structured as follows: Section 2 introduces the relevant work of multimodal RS data fusion, geospatial object detection, socioeconomic relevance of detecting and mapping WWTP, and deep learning from VGI. The JDL method together with the detailed network designs is presented in Section 3, followed by Section 4, which elaborates on the datasets and the experimental set-up, together with the experiment results and potential applications. Section 5 summarizes the lessons learned and provides concluding remarks.

![Fig. 1](image-url) An overview of the proposed method of joint deep learning. There are mainly three input data: (a) VHR image; (b) OpenStreetMap data; (c) Sentinel-2 MSI.
regarding our RQs.

2. Related work

2.1. Socioeconomic relevance of detecting and mapping wastewater treatment plants

Clean drinking water as well as sanitation have been designated a universal human right by the United Nations (Generalversammlung der Vereinten Nationen, 2010) since 2010, however it is still unavailable for hundreds of millions of people in the world (UNICEF/WHO, 2015). For public or private water utility organizations or equipment manufacturers, detecting and mapping WWTP arises of great economic value for their business development and their investment decision-making. They both play an important role in ensuring basic services, expanding them, and closing gaps in the clean water supply. Even public authorities, administrations, or governments may benefit from such insights, incorporating them into the design of funding programs.

The lack of data in sparsely mapped regions makes the use of deep learning from Earth observation data even more beneficial. The socio-economic potential for the use of GeoAI to detect and map WWTP arises from a simple fact that it identifies areas needing attention, which can be designated as undersupplied with the additional insights from demographic data. One of the motivations of this paper is to further explore this potential by proposing a novel JDL method for accurate and automatic WWTP mapping.

2.2. Multimodal remote sensing data fusion

Given the emerging availability of multimodal RS data captured by different satellites, their inherent heterogeneity poses a pressing challenge for more effective and efficient multimodal RS data fusion (Ghamisi et al., 2019). In this regard, multimodal and multisensor data fusion has received intensive research interest across a wide range of applications, ranging from image pansharpening (or resolution enhancement), multisensor fusion and classification, to multimodal/crossmodel feature learning. For instance, Yuan et al. (2018) propose a multiscale and multidepth convolutional neural network (MSDCNN) for MSI image pansharpening. Subsequently, He et al. (2019) develop a novel detail injection based CNN framework for image pansharpening, where MSI details were explicitly formulated in an end-to-end manner. As for multimodal and multisensor data fusion, Rasti et al. (2017) jointly extract spectral and elevation features with a novel sparse and low-rank method for the accurate image classification of HSI and LiDAR data. Similarly, Li et al. (2018) design a three-stream CNN for the deep fusion of spectral-spatial-elevation features derived from HSI and LiDAR to achieve higher classification accuracy than using individual RS data. Moreover, recent work in Hong et al. (2019) proposes a semi-supervised cross-modality learning framework, namely learnable manifold alignment (LeMA), for accurate LULC classification by fusing HSI and MSI data.

Unlike common image-level data fusion methods, multimodal/cross-model feature learning learns directly from feature-level data fusion models. Preliminary works in this direction (Tuia et al., 2014; Hong et al., 2021) investigate either manifold learning based or deep learning based approaches for an effective and robust feature learning purpose, respectively. Besides these existing works, community contributions, for instance Yokaya et al. (2018), have played a key role in promoting multimodal RS data fusion research by providing open-access data benchmarks and organizing regular data fusion contests. However, the majority of existing works and benchmarks of multimodal RS data have focused on image classification tasks; thus the huge potential of multimodal RS data fusion in more sophisticated tasks, such as the geospatial object detection, remains underexplored and deserves more attention.

2.3. Geospatial object detection and mapping

Advances in modern imaging techniques allow Earth observation with a resolution up to a sub-meter level, which enables objects of interest (e.g., buildings or vehicles) to be detected with unprecedented speed and accuracy. Geospatial object detection and mapping have been an essential approach for many real-world applications, such as precision agriculture (Sadgave et al., 2018), wildlife conservation (Kellnerberger et al., 2018), and humanitarian mapping (Li et al., 2020). In one of the most successful examples, Brandt et al. (2020) detected over 1.8 billion individual trees in the West African Sahara, Sahel and sub-humid areas from VHR satellite imagery (i.e., sub-meter resolution) using a deep learning method. Intuitively, such deep learning-based geospatial object detection methods offer an unprecedented ability to monitor and map target objects on a global scale via a fully automatic workflow. Nevertheless, the lack of large-scale training data has become a major obstacle to the development of geospatial object detection (Ding et al., 2021). Fortunately, considerable effort has been dedicated to creating benchmark datasets for multi-class geospatial object detection, such as NWPU VHR-10 (Cheng et al., 2016), DOTA (Ding et al., 2021), DIOR (Li et al., 2020), and FAIR1M (Sun et al., 2022).

Despite the achievements to date, however, it is still an open question how multimodal RS data could be leveraged to boost the performance of current object detection models, especially when relying on single RS data could not yield satisfactory performance. The potential herein lies in complementary perspectives (e.g., spatial or spectral) consisted in multimodal RS data. In addition, the current annotation method used in benchmark development is still costly with respect to labor and time, which can be an even more severe problem in annotating such geographically sparse objects as WWTP.

2.4. Deep learning from VGI

Due to the development of big data and crowdsourcing technology, VGI as a special case of user-generated content continued to harvest big geographic data that was contributed voluntarily by individuals (Goodchild, 2007). More importantly, VGI platforms like OSM provided a promising source of massive, free labels together with rich and detailed semantic information for satellite image analysis (Chen and Zipf, 2017). Deep learning from VGI, therefore, has received increasing research attention from both the VGI and RS communities. An early attempt by Mnih and Hinton (2012) first extracted vector data from OSM for supervised street detection using a deep learning model, where different loss functions were considered in order to minimize the effect of missing error and registration error in OSM labels. In (Chen and Zipf, 2017), a VGI-based active learning workflow called DeepVGI was proposed to classify built-up areas for a humanitarian mapping task; volunteer contributions via the MapSwipe App were used to guide the model training. Moreover, a follow-up work in (Herfort et al., 2019) developed a novel workflow to combine deep learning and crowdsourcing efforts via a task allocation strategy for better and faster humanitarian mapping. The results from Guatemala, Laos, and Malawi showed that the proposed machine-assisted workflow can reduce volunteer efforts by at least 80%. More recently, Li et al. (2020) successfully mapped 13 OSM missing built-up areas by combining clustering of geo-tagged tweets and deep learning building detection models. Although deep learning from VGI provided an inspiring solution for detecting basic geospatial objects (like roads and buildings), few works have investigated mapping more sophisticated objects in OSM with the help of its rich semantic information.

3. Methodology

3.1. Joint deep learning overview

In this paper, we investigate the feasibility of leveraging OSM and
multimodal RS data for accurate detection of WWTP by proposing a novel JDL method (see Fig. 1). The proposed JDL method consists of two major networks: a fine-tuned object detection network and multi-task RAN. To be more specific, the former network is fine-tuned with training samples generated from VHR images and WWTP features (from OSM) so to detect possible WWTP taking VHR images as input. Next, the predicted WWTP boxes together as an output from the former network with their corresponding Sentinel-2 MSI patches are fed into the multi-task RAN to produce a dual-task classification of both LULC and WWTP from Sentinel-2 MSI simultaneously. In short, these two networks are trained consecutively, which will generate the LULC classification as well as WWTP detection as a result. Intuitively, the proposed method jointly learns from VGI and multimodal RS dataset(s) to achieve the fully automatic mapping of WWTP.

3.2. Fine-tuned object detection network

The proposed JDL method starts with a fine-tuned object detection network that learns from OSM data to extract similar representations of WWTP features from VHR images. For this purpose, we fine-tune a Single Shot Multibox Detection (SSD) network (Liu et al., 2016) using a base network of Inception V2 (Szegedy et al., 2016), which was pre-trained on the Microsoft COCO dataset. As a single-stage object detector, the SSD provides a concise and robust solution to detect and map geospatial objects of diverse sizes and shapes (e.g., WWTP). Fig. 2 illustrates the detailed design of the SSD network.

The SSD network has a tiling strategy of default boxes to ensure that specific feature maps learn to predict certain scales of target objects, which differentiates it from other two-stage object detection networks (e.g., Faster R-CNN) with feature maps corresponding to different receptive field sizes. Let \( f \) be the number of feature maps used for prediction; the scale of the \( k \)-th default object box is defined as follows:

\[
S_k = S_{\text{min}} + (k - 1) \frac{S_{\text{max}} - S_{\text{min}}}{f - 1},
\]

where we follow the default setting in (Liu et al., 2016) and set \( S_{\text{min}} \) and \( S_{\text{max}} \) to 0.2 and 0.9, respectively, which forces a regular distribution of all the scales of all layers in between.

Moreover, given the various aspect ratios \( (a_k) \) for the default boxes within different feature maps, for instance \( a_k \in \{1, 2, 3, 4, 5\} \), the SSD network results in six default boxes per feature map location, where the width \( (w_k) \) and height \( (h_k) \) of each default box can be calculated as:

\[
w_k = S_k \sqrt{a_k}, \quad h_k = S_k / \sqrt{a_k}.
\]

As a result, various sizes and shapes of targets are matched with particular feature maps via this tiling strategy (see Fig. 2 (a)-(c)). In other words, the SSD network can capture objects of different sizes and shapes by combining predictions of diverse scales and aspect ratios from all feature maps. This design is of great importance, especially for geospatial object detection, since the sizes and shapes of the target objects (e.g., WWTP) often vary significantly.

The training objective of the SSD network is to optimize the overall training loss. Let \( p_0 = \{1, 0\} \) be an indicator for matching the \( i \)-th SSD default object box to the \( j \)-th ground truth object box. The overall loss function is a sum of the confidence loss (conf) and the localization loss (loc):

\[
L(p, c, l, g) = \frac{1}{N} \left(L_{\text{conf}}(p, c) + L_{\text{loc}}(p, l, g)\right),
\]

where \( N \) is the number of matched default boxes, \( c \) denotes the list of class-wise confidence scores, while \( l \) and \( g \) refer to the predicted box and the ground truth box, respectively.

Specifically, the confidence loss and the localization loss are calculated as follows:

\[
L_{\text{conf}}(p, c) = -\sum_{i \in \text{Pos}} p_i^m \log (\hat{c}_i^m) - \sum_{i \in \text{Neg}} \log (\hat{c}_i^m),
\]

with

\[
\hat{c}_i^m = \frac{\exp(c_i^m)}{\sum_{a} \exp(c_a^m)}
\]

where \( \text{Pos} \) refers to the collection of matched default boxes, \( \text{Neg} \) refers to unmatched ones, and \( m \) is the total number of classes.

\[
L_{\text{loc}}(p, l, g) = \sum_{i \in \text{Pos}} \sum_k \|
\]
As shown in Fig. 2, we adopted an improved Non-Maximum Suppression (NMS), namely Soft-NMS (Bodla et al., 2017), to suppress the number of redundant candidate boxes from all feature maps before classification. After training, inferences with the SSD network need only VHR images as inputs, then generates a list of predicted WWTP features (i.e., $P(x_i)$), which are taken as inputs for the succeeding multi-task classification network.

### 3.3. Multi-task residual attention network

Once the SSD network detects potential WWTP from VHR images, these predictions are used to create inputs for the multi-task RAN. Specifically, MSI data is sampled with the centroid of each detected WWTP features in a sliding window manner, then fed into the RAN for multi-task classification of WWTP and LULC. Our implementation of the RAN (see Fig. 1) consists of three attention modules (AM), where each AM includes a standard residual block (RB) (He et al., 2016) together with an attention block (AB).

As shown in Fig. 3, each AB is made of a trunk branch of major feature processing, and more importantly a soft mask branch to provide attention feature masks. Given an MSI patch $x_{ij}$, where $i$ and $j$ refer to the index of spatial positions and spectral channels, the output of an AB could be recursively formulated as follows:

$$ F_{ij}(x, \phi) = (1 + M_{ij}(x, \mu))^1 \cdot T_{ij}(x, \phi), \quad \text{(7)} $$

where $T_{ij}(x, \phi)$ refers to the trunk branch and its branch-wise parameters $\phi$. Similarly, $M(x, \mu)$ and $F_{ij}(x, \mu, \phi)$ are feature maps and their parameters of the soft mask branch and the complete AB. Besides the benefit of residual learning (He et al., 2016), the soft mask branch serves as an attention mask and provides substantial guidance during discriminative feature selection (from the trunk branch) during network forward inference. Moreover, during network back propagation, the soft mask branch is beneficial in handling noisy labels and making the RAN more robust against overfitting. Specifically, the gradient of masking over trunk branch features is:

$$ \frac{\partial M_{ij}(x, \mu)^1 \cdot T_{ij}(x, \phi)}{\partial \phi} = M_{ij}(x, \mu)^1 \cdot T_{ij}(x, \phi), \quad \text{(8)} $$

where $\phi$ and $\mu$ refer to the parameters of trunk and soft mask branches, respectively. Since $M(x, \mu)$ is irrelevant to $\phi$, it prevents the noisy gradients from updating the trunk branch parameter during gradient back propagation.

Similar to the ResNet, multiple AMs can be easily concatenated in the RAN, where discriminative attention masks provided by each AB also change adaptively when the RAN is going deeper. Intuitively, the RAN combines the advantage of both residual learning and the attention mechanism for better and faster network training. For more details, refer to Wang et al. (2017).

In addition, the RAN allows three activation functions for the normalization of soft mask features: spectral attention, spatial attention, and mixed attention. The definitions of these activation functions are given as follows:

$$ \begin{align*}
    f_{\text{spec}} & = \frac{x_{ij}}{||x_{ij}||} \\
    f_{\text{spa}} & = \frac{1}{1 + \exp(-(x_{ij} - \text{mean}(c))/\text{std}(c))} \\
    f_{\text{mixed}} & = \frac{1}{1 + \exp(-(x_{ij}))}
\end{align*} \quad \text{(9)} $$

where spectral attention ($f_{\text{spec}}$) conducts L2 normalization over all spectral channels for each spatial position of MSI pixels and spatial attention ($f_{\text{spa}}$) normalizes channel-wise feature maps and then use the sigmoid function to get the spatial soft mask. Following the suggestion in Wang et al. (2017), we selected mixed attention ($f_{\text{mixed}}$) to address accuracy concerns; it simply runs a sigmoid function over all spectral channels and spatial positions.

Unlike the original implementation of RAN, our multi-task RAN has two separate fully connected layers besides the aforementioned three AMs. Thus the multi-task loss function is calculated as follows:

$$ L_{\text{MTL}} = L_{\text{WWTP}} + L_{\text{LULC}} \quad \text{(10)} $$

where $L_{\text{WWTP}}$ refers to the sigmoid loss of the binary classification of WWTP, and $L_{\text{LULC}}$ is the softmax loss of multi-classes LULC classification. In this context, the network jointly learns discriminative features of WWTP and LULC objects and continuously complements each task during the optimization of training loss, which aims to boost the accuracy and efficiency of individual task learning. In addition, a similar while single-task RAN is included for a comparison purpose to examine the potential performance gain of multi-task learning.

Correspondingly, distinct evaluation metrics are selected to assess the accuracy of WWTP detection and LULC classification against the reference data. For WWTP detection, we first classify prediction boxes into False Negatives (FN), False Positives (FP), and True Positives (TP), then derive Precision, Recall, and F1 scores for evaluation. Note that True Negatives are not included, as the SSD network did not detect negative samples. As for LULC classification, we evaluate the model performance w.r.t. the overall accuracy (OA), the average accuracy

![Fig. 3. Illustration of the Attention Block used for the multi-task classification of WWTP and LULC.](image-url)
(AA), the Kappa coefficient, and individual class accuracies.

4. Experiment

4.1. Data description

We evaluated the proposed JDL method in Germany, as it has relatively well-developed wastewater treatment facilities and also is a country rich in water resources (Schröter et al., 2005). As shown in Fig. 4, we used three main types of multimodal data sources. Specifically, VHR satellite images were included to fine-tune the SSD object detection network in all federal states of Germany except Baden-Württemberg; MSI data from Sentinel-2 were considered for multi-task learning of RAN to further classify potential WWTP features and LULC classes in a selected area around Stuttgart, which is also the capital and largest city of Baden-Württemberg. Moreover, the geographical and semantic data in OSM were mined to generate massive training labels for both networks (SSD and RAN).

In this paper, we used the OSM tags in Table 1 to extract the most significant structures of WWTP, namely wastewater basins, as target objects, which resulted in 4,187 features for training (see Fig. 4). Specifically, we included all potential wastewater basins within the polygon of WWTP sites (OSM tag: man_made = wastewater_plant) as the training data to fine-tune the pre-trained SSD network. As for evaluation, we manually created a reference layer and mapped all wastewater basins of WWTP (723 features) within our test area around Stuttgart. In addition, we collected the Urban Wastewater Treatment Directive-reported (UWWTD) data (European Environment Agency, 2017) as external reference data; UWWTD recorded all WWTP sites as point features based on a traditional survey method. The VHR satellite imagery tiles were provided by a tile map service (TMS) from Microsoft Bing. As a result, 4,187 OSM wastewater basin geometries were used to label Bing satellite imagery at a zoom level 17 (at a spatial resolution of around 1.2m) via the ohsome2label tool (Wu et al., 2020b). Then WWTP labels together with VHR images were converted to standard training records for the Tensorflow Object Detection API (Huang et al., 2017).

The MSI sensor of Sentinel-2 satellites covers 13 spectral bands at diverse spatial resolutions ranging from 10 meters to 60 meters; these spectral bands capture spectral reflectances from RGB to NIR and SWIR bands (Drusch et al., 2012). In this paper, we used the best pixel Sentinel-2 MSI composite processed by the Food and Agriculture Organization of the United Nations (FAO) via the SEPAL cloud platform data processing system (sepal.io). Specifically, top of atmosphere (TOA) reflectance was converted to surface reflectance then the best pixels were selected from the past three years of April 2020 using a medioid compositing function. For our feature space, we used in total 10 bands at a 10m spatial resolution, where all 20m bands (i.e., B5, B6, B7, B8A, B11, B12) were resampled to 10 meters to concatenate with 10m bands (i.e., B1, B2, B3, and B4). Moreover, we included an open surface water layer (Li et al., 2021) in Germany to filter the raw WWTP predictions as a comparison method to highlight the benefit of multi-task RAN in handling the issues of FP predictions.

Besides the VHR and MSI data, we followed the good practice in (Schultz et al., 2017) of converting OSM LULC-related information into

| Class | Name | OSM tags |
|-------|------|----------|
| 1     | Urban fabric | residential |
| 2     | Industrial, commercial, and transport | retail, industrial, commercial, port, railway |
| 3     | Mine, construction sites | construction, landfill, quarry |
| 4     | Artificial vegetated areas | golf course, park, garden, recreation, stadium, playground |
| 5     | Arable land | farmland, farm, greenhouse, farmyard, horticulture |
| 6     | Permanent crops | vineyard, orchard |
| 7     | Pastures | meadow |
| 8     | Forests | forest, woodland |
| 9     | Shrub | grass, shrub, grassland, greenfield |
| 10    | Open spaces with no vegetation | sand, scree, beach, rock |
| 11    | Inland wetlands | marsh wetland |
| 12    | Water bodies | water, riverbank, basin, dock, reservoir |

Table 1: OSM Tags Used to Identify WWTP in Germany.

| OSM Key | OSM Value | Description |
|---------|-----------|-------------|
| water   | wastewater, basin | A clarifier/settling basin of a wastewater treatment plant.. |
| landuse | basin, pond | An area of land artificially graded to hold water. |
| natural | water, waterway | Any inland body of water, from natural such as a lake or pond to artificial like a basin or pond. |
| man_made | wastewater_plant | A wastewater plant is a facility used to treat wastewater. |

Table 2: The LULC Classification Scheme and the Corresponding OSM Tags.
12 classes, as shown in Table 2, ranging from urban, industrial areas to various areas of vegetation. Since our main objective of including LULC information was to investigate if it can support more accurate detection of WWTP via a multi-tasking learning network, the number of training and testing samples for LULC classification was therefore set to be balanced between different classes. As a result, we used 4,200 LULC samples together with 4,178 OSM wastewater basin samples for training and 1,200 LULC samples for testing, where all testing samples were manually assigned with the reference LULC classes by internal expert volunteers. Similar to the object detection network, we trained the multi-task classification RAN in Germany (except Baden-Württemberg) and evaluated it in our test area around Stuttgart.

4.2. Experimental setup

The pre-trained SSD object detection network (Liu et al., 2016) was built on the Inception V2 backbone network with all parameters pre-trained on the Microsoft COCO dataset (Lin et al., 2014). Specifically, the pre-trained parameters were downloaded from the Tensorflow Detection Model Zoo (Tensorflow, 2020), and the training process for WWTP detection in Germany was run for 50,000 epochs with an initial learning rate of 0.0004. All codes were implemented using Python 3.6, Tensorflow 1.14, and the Tensorflow object detection API (Huang et al., 2017).

Fig. 5 elaborates the detailed network design with individual feature sizes for each layer in the multi-task RAN. Specifically, we followed a basic architecture of RAN-56 (Wang et al., 2017) consisting mainly of three ABs and RBs. Moreover, to facilitate the multi-task learning, we attached two different fully connected layers for the classification of WWTP and LULC. The Nesterov Adam optimizer was selected as the optimization algorithm due to its faster convergence performance, and the default parameters $\beta_1 = 0.9, \beta_2 = 0.999$ were adopted. We set the learning rate, training epoch and batch size of the RAN-56 to 0.0001, 500, and 64, respectively. The window size of the Sentinel-2 MSI patch was empirically set to $28 \times 28$, referring to a 280 meter geographical vicinity. As for the reverse geocoding of WWTP sites, we used the geocoding tool offered by OpenRouteService (ORS), which was built on top of the Pelias stack (ORS, 2021). All experiments ran on a Linux server with 4 GeForce RTX 2080Ti graphical processing units (GPUs), each with 12 GB memory.

4.3. Result and analysis

Table 3 lists the quantitative results of WWTP and LULC classification by comparing raw WWTP predictions, single-task classification methods, and our proposed multi-task method. The raw WWTP predictions were obtained by applying the fine-tuned SSD network across our test area around Stuttgart and collecting all potential WWTP predictions. Single-task RAN (either for LULC or WWTP) was included for comparison purposes. In this context, the key findings are summarized as follows:

- Characterized by a high recall value (over 90%), the fine-tuned SSD network was able to effectively detect potential WWTP while also generating massive FP predictions, which resulted in a poor precision (14.21%) in raw WWTP predictions. Fortunately, the multi-task RAN improved the precision value up to 74% and compromised the recall to an acceptable extent (from 90% to 84%), thus leading to the best F1 score of 0.7922.
- Although most WWTP features are associated with visible water basins, the sophisticated structure of WWTP facilities poses consistent challenges in their classification. It is obvious that a single-task RAN trained on MSI data is not sufficient to handle massive FP predictions, as compared to the multi-task result. This finding confirms the benefits of including LULC context information in the accurate classification of WWTP features via multi-task learning.
- Regarding LULC classification, the benefit of multi-task learning was less significant, as the multi-task RAN (OA: 74% and AA: 71%) slightly outperformed single-task RAN (OA: 73% and AA: 79%) in the LULC classification task. This result was somewhat expected, since one can imagine a limited contribution of WWTP features to the classification of industrial areas and water bodies, as listed in Table 4.

The visual comparison of WWTP detection results in Fig. 6 supports some of our previous findings. Herein, TP together with FP illustrate the predicted WWTP features via different models, while FN refers to manually-created WWTP features that were missed by our model. Based on four selected subareas (i.e., (a), (b), (c), and (d)), it can be seen that multi-task RAN is indeed helpful for filtering out FP predictions generated by our fine-tuned SSD network, especially in subareas (a) and (d). From a multimodal RS data fusion perspective, the difference between real WWTP and those FP urban structures could be better captured by combining VHR and MSI data than relying on a single data source, which again confirms the effectiveness of our JDL method. Next, compared to the single-task RAN, for instance in subareas (b) and (c), the multi-task RAN benefited from the process of simultaneously extracting discriminative features for LULC and WWTP classification, and thereby was able to better distinguish raw WWTP predictions and reduce the cases of misclassification due to the various shapes and sizes of WWTP. In short, the proposed JDL method presents a two-stage solution that is effective at boosting overall performance when detecting geographically sparse objects (e.g., WWTP) via multi-task learning and multimodal RS data fusion.

| Layer                | Feature Size | RAN-56                        |
|---------------------|--------------|-------------------------------|
| Conv1               | 28 x 28      | 5 x 5, 32, padding 2          |
| Max Pooling         | 16 x 16      | 2 x 2, stride 2               |
| Residual Block #1   | 16 x 16      | $\begin{bmatrix} 1 & 1,32 \ 3 x 3,52 \ 1 & 1,64 \end{bmatrix} \times 1$ |
| Attention Block #1  | 16 x 16      | Attention x 1                 |
| Residual Block #2   | 8 x 8        | $\begin{bmatrix} 1 & 1,64 \ 3 x 3,64 \ 1 & 1,129 \end{bmatrix} \times 1$ |
| Attention Block #2  | 8 x 8        | Attention x 1                 |
| Residual Block #3   | 4 x 4        | $\begin{bmatrix} 1 & 1,128 \ 3 x 3,128 \ 1 & 1,256 \end{bmatrix} \times 1$ |
| Attention Block #3  | 4 x 4        | Attention x 1                 |
| Average Pooling     | 1 x 1        | 4 x 4, stride 1               |
| Fully Connected     | WWTP: 1      | LULC: 12                      |
| Total params (10^9) |              | 2.53                          |
| Trainable params (10^9) |            | 2.51                          |

Fig. 5. Detailed architecture of the multi-task RAN-56 and its number of parameters.

| Method               | WWTP detection | LULC classification |
|---------------------|----------------|---------------------|
|                     | Precision (%)  | Recall (%)          |
| Raw WWTP            | 14.21          | 90.73               | 0.2453 | -     | -     |
| Single-task RAN     | -              | -                   | -      | 73.00 | 71.04 | 0.7021 |
| Single-task RAN     | 36.42          | 78.42               | 0.4973 | -     | -     |
| Multi-task RAN      | 74.66          | 84.37               | 0.7922 | 74.83 | 79.03 | 0.7255 |

Table 3. Performance Evaluation of Different Methods.
To elaborate further on the LULC classification task, Table 4 summarizes the classification accuracy per class for single-task RAN (LULC) and multi-task RAN. Although the overall performance gain in LULC classification (Table 3) is insignificant, we notice that multi-task RAN outperformed single-task method in specific classes, for examples in industrial and water bodies, which are commonly associated with our target object of WWTP. Fig. 7 shows the classification maps of the selected subareas as well as the false color Sentinel-2 MSI image. It can be seen that the multi-task RAN was able to clearly delineate the WWTP in contrast to the surroundings LULC context. This visual comparison gives a strong support to the aforementioned findings, and more importantly, confirmed the complementary effect of LULC classification in more accurate WWTP detection. The question of how to further boost such OSM-based LULC classification is beyond the scope of this paper here.

To compare our JDL method with traditional survey data, we included the UWWTD data of all WWTP sites within our test area around Stuttgart. Since the UWWTD data recorded WWTP sites as point features, we thus grouped our JDL-based predictions within 500 meters into spatial clusters using the well-known DBSCAN clustering method (Ester et al., 1996), then took the geometry center of each cluster as a unique WWTP site. The preliminary assumption was that wastewater basins within a certain geographical vicinity belong to the same treatment site. A similar clustering was applied to the reference WWTP layer, which resulted in 139 WWTP clusters (or sites). To validate the JDL-based method, in case the distance between reference and predicted sites was smaller than 500 meters, it was considered as a correct prediction, and vice versa, as an incorrect prediction. In addition, we filtered our raw WWTP predictions with the OSWL provided by Li et al. (2021) for purposes of comparison.

Table 5 lists the validation results of both JDL-based and survey methods regarding the mapping accuracy of WWTP clusters. Although the survey data is still most accurate, the data production for that method can be extremely labor and time intensive. The proposed JDL method, especially with multi-task RAN, offers a promising alternative solution for fully-automatic WWTP mapping and achieves a competitive mapping accuracy compared to the filtered results, based on existing surface water layers (e.g., OSWL). Therefore, we believe there is huge potential to apply our JDL method to large-scale WWTP detection and mapping, especially in rapidly developing countries and regions, since it consumes mainly open access data in a fully automatic manner.

When applied to WWTP, another advantage of our method is that, in addition to single features, other data or information about the corresponding sites can be aggregated. The spatial correlation of features allows further analysis at an object level. The number of similar shaped basins provides an indicator of plant capacity and consequently the level of supply. The number of different basins or facilities can be used as an indicator for the plant’s stages of treatment. The absence of secondary or tertiary treatment again points to the potential to increase the supply of clean water.

Fig. 8 illustrates two major outcomes of our JDL method: the LULC classification maps and the WWTP detection maps (multi-task RAN). Besides the aforementioned findings, we demonstrated how these WWTP maps can be converted into useful information via a reverse-geocoding process, where the address of WWTP sites can be estimated using geographical coordinates. Adding address data increases the entrepreneurial value since it identifies the location precisely, but such a list of addresses plays a vital role in many real-world use cases, as well.

Table 4
The LULC Classification Accuracy for Each Class using a Single-task or Multi-task RAN.

| # | Class                         | Test sample | Single-task RAN | Multi-task RAN |
|---|-------------------------------|-------------|----------------|---------------|
| 1 | Urban fabric                  | 90          | 92.22          | 86.67         |
| 2 | Industrial, commercial, and   | 59          | 66.10          | 96.61         |
|   | transport                     |             |                |               |
| 3 | Mine, construction sites      | 121         | 84.30          | 70.25         |
| 4 | Artificial vegetated areas    | 102         | 71.57          | 73.53         |
| 5 | Arable land                   | 170         | 72.35          | 55.88         |
| 6 | Permanent crops               | 100         | 71.02          | 84.03         |
| 7 | Pastures                      | 81          | 77.78          | 85.19         |
| 8 | Forests                       | 166         | 83.73          | 57.23         |
| 9 | Shrub                         | 91          | 41.76          | 58.24         |
| 10| Open spaces with no           | 68          | 45.59          | 94.12         |
|   | vegetation                    |             |                |               |
| 11| Inland wetlands               | 59          | 64.41          | 89.83         |
| 12| Water bodies                  | 93          | 81.72          | 96.77         |

Fig. 6. Detection Results of WWTP in Four selected Area with Different Methods.
For instance, it supports the assessment of local sanitation levels and the estimation of regional clean water demands. When such data are aggregated with demographic data, areas identified as undersupplied can be highlighted in a more effective way considering the monitoring of SDG 6 (Persello et al., 2022). This could support the development of a scale to measure supply and undersupply on the district or neighborhood level.

From a business perspective, besides the aforementioned knowledge about market opportunities or direct customer acquisition, companies in some sectors may use the data for further investigation on an object level. Wastewater treatment plants have different levels of treatment, each one with a characteristic layout or appearance on aerial imagery. By distinguishing different shapes of basins, the level of treatment can be estimated. The size and number of basins per plant indicates the extent of the pipe system and the dimensions of the pumps. Future works are thus encouraged to further investigate the socioeconomic potential of areas with different methods.

Table 5
Clustering Results and Identification Accuracies of WWTP Sites with Different Methods.

| Sources | Methods            | Clusters | TP   | FP   | FN   | Precision (%) | Recall (%) | F1    |
|---------|--------------------|----------|------|------|------|---------------|------------|-------|
| JDL-based | Multi-task RAN     | 154      | 113  | 41   | 26   | 73.38         | 81.29      | 0.7713|
|         | Filtered Raw WWTP  | 112      | 94   | 18   | 45   | 83.92         | 67.62      | 0.7490|
| Survey  | UWWTD              | 125      | 123  | 2    | 16   | 98.40         | 88.49      | 0.9318|

Fig. 7. LULC classification map of four selected areas with different methods.

Fig. 8. Two major results of the JDL method: Left: the LULC classification map produced by the multi-task RAN; Right: the map of clustered WWTP sites with the WWTP bounding boxes detected by the JDL method and their geocoding addresses.
more efficient processing of big geospatial data in different applications (Werner, 2019).

5. Conclusion

An accurate and up-to-date map of WWTP is vital in monitoring access to clean water and sanitation facilities, however it is difficult to rely on a single modality of RS or VGI data for effective mapping. Inspired by this fact, in this paper we aimed at leveraging OSM and multimodal RS data by proposing a novel JDL method to jointly learn from multimodal RS data (e.g., VHR and MSI) for model training and simultaneously address two tasks, LULC classification and WWTP detection. The proposed JDL method consists of a fine-tuned object detection network, namely the SSD network, and a multi-task classification RAN. Note that our method consumed freely available OSM training samples (e.g., 4,187 WWTP features and 4,200 LULC samples mapped by OSM volunteers within Germany) during the training of both networks, thus offering a general solution for large-scale WWTP and LULC mapping. Experiment results conducted on a selected test area around Stuttgart (with 723 WWTP features and 1,200 LULC samples for testing) demonstrated the effectiveness and superiority of our JDL method compared to single modalities and single tasks. Moreover, the comparison with survey data and existing surface water products showed that the generated WWTP maps are of competitive accuracy and completeness, especially in the multi-task RAN case.

With respect to our RQ1, the integration of a fine-tuned SSD network and multi-task RAN was confirmed to be effective in handling FP predictions generated from the fine-tuned SSD network, and improving the accuracy of WWTP detection with the support of LULC information via multi-task learning. In addition, we presented a flexible approach of harvesting OSM data as training samples, which can be easily scaled up due to the free access of OSM data. The lessons learned encourages future works on automatic mapping and detection of diverse geospatial objects from multimodal RS data. According to RQ2, we compared multi-task RAN with their single-task versions, and further elaborated on the effect of LULC context information in distinguishing WWTP predictions. As a result, the multi-task RAN method outperformed single-task methods with a higher precision (74.66%) and recall (84.37%) value, as well as a highly competitive accuracy compared to single modalities and single tasks. Moreover, the learning framework empirical deep learning models for multiple geospatial data applications to a larger scale and extend current deep learning models for multiple geospatial data applications to a larger scale and extend current deep learning models for multiple geospatial data applications to a larger scale and extend current deep learning models for multiple geospatial data applications to a larger scale and extend current deep learning models for multiple geospatial data applications to a larger scale and extend current deep learning models for multiple geospatial data applications to a larger scale and extend current deep learning models for multiple geospatial data applications.
