TOWARDS LARGE-SCALE BUILDING ATTRIBUTE MAPPING USING CROWDSOURCED IMAGES: SCENE TEXT RECOGNITION ON FLICKR AND PROBLEMS TO BE SOLVED

Y. Sun[1]*, A. Kruspe[2], L. Meng[3], Y. Tian[1], E. J. Hoffmann[1], S. Auer[4], X. X. Zhu[1]

[1] Data Science in Earth Observation, Technical University of Munich,
(yao.sun, yifan.tian, eike.jens.hoffmann, xiaoxiang.zhu)@tum.de
[2] Falculty of Computer Science, Technische Hochschule Nürnberg, anna.kruspe@th-nuernberg.de
[3] Cartography and Visual Analytics, Technical University of Munich, liqiu.meng@tum.de
[4] Remote Sensing Technology Institute, German Aerospace Center (DLR), stefan.auer@dlr.de

KEY WORDS: Building Attributes, Scene Text Recognition (STR), Street-view Images (SVI), Flickr, Crowdsource, OpenStreetMap (OSM), Building Function.

ABSTRACT:

Crowdsourced platforms provide huge amounts of street-view images that contain valuable building information. This work addresses the challenges in applying Scene Text Recognition (STR) in crowdsourced street-view images for building attribute mapping. We use Flickr images, particularly examining texts on building facades. A Berlin Flickr dataset is created, and pre-trained STR models are used for text detection and recognition. Manual checking on a subset of STR-recognized images demonstrates high accuracy. We examined the correlation between STR results and building functions, and analysed instances where texts were recognized on residential buildings but not on commercial ones. Further investigation revealed significant challenges associated with this task, including small text regions in street-view images, the absence of ground truth labels, and mismatches in buildings in Flickr images and building footprints in OpenStreetMap (OSM). To develop city-wide mapping beyond urban hotspot locations, we suggest differentiating the scenarios where STR proves effective while developing appropriate algorithms or bringing in additional data for handling other cases. Furthermore, interdisciplinary collaboration should be undertaken to understand the motivation behind building photography and labeling. The STR-on-Flickr results are publicly available at `https://github.com/ya0-sun/STR-Berlin`.

1. INTRODUCTION

Building information extraction has been a hot topic since three decades in photogrammetry and remote sensing. Numerous studies have been conducted using different types of data, e.g., optical satellite images (Liasis and Stavrou, 2016), airborne LiDAR (Brenner, 2005), Synthetic Aperture Radar (Sportouche et al., 2011; Brunner et al., 2010; Sun et al., 2022). However, most of the research efforts primarily concentrate on building geometries, such as footprints, heights, and 3D models in different levels of details (Li et al., 2020; Sun et al., 2021; Chen et al., 2021; Sun, 2016), and hardly attend to building attributes, such as building type, age, material, ownership, number of households, and more, which are essential for varies urban application such as public facility planning and resource distribution.

Some works employ aerial or satellite images to estimate building attributes, e.g., building functions (Huang et al., 2018; Zhang et al., 2019). However, nadir-looking images are inherently ambiguous as they mainly feature rooftops. In recent years, researchers started to employ street-view images (SVI) featuring building facades. Google Street View images, as a major commercial source, have been used for building age estimation (Li et al., 2018), flood risk of buildings (Chen et al., 2022), building heights (Yan and Huang, 2022), and more. On the other hand, crowdsourced platforms, such as Flickr, Unsplash, and Mapillary, provide huge amounts of street-view images containing valuable information. They are ubiquitous, cheap, easy to collect, and increasingly prevalent in research. Flickr, as an example, has been employed for mapping and understanding landscape aesthetics (Langemeyer et al., 2018), land use classification (Leung and Newsam, 2012), and flood-level estimation (Chaudhary et al., 2019). As for building attributes, Kang et al. (Kang et al., 2018) classified building instances from street-view images using convolutional neural networks. Hoffmann et al. (Hoffmann et al., 2023) employed Flickr images from 42 cities to classify buildings using deep neural networks, demonstrating the mapping potential utilizing crowdsourced data on a large scale.

However, in image classification tasks, texts are often ignored. Texts on building facades contain rich attribute information, such as shop names, building usage, house numbers, and construction years. Scene Text Recognition (STR) is the task of reading texts in natural scenes. Despite the success of Optical Character Recognition (OCR) systems on clean documents, the STR remains a difficult task due to the diverse text appearances in the real world and the imperfect conditions in which these scenes are captured. Sun et al. (Sun et al., 2023) presented STR results and extracted building attributes from a few street-view images, however, large-scale mapping remains challenging.

Aiming at large-scale building attribute mapping using crowdsourced images, we apply STR on Flickr data. In this paper, we report our preliminary results and observations to identify situations in which STR-on-SVI helps map building attributes and address the challenges in this field for future works.

Next, we present the methods in Section 2 and detail the experimental results and analysis in Section 3. Finally, Section 4 concludes the paper and outlines possible future directions.

* Corresponding author

2. METHODOLOGY

Our workflow comprises three main steps: first, build the Flickr dataset; second, associate building functions with the Flickr images; and third, extract texts on the images using STR.

2.1 Flickr image filtering

Flickr covers a diverse range of content and motifs and provides an accessible API encouraging users to share and use photos. For our STR tasks, it is necessary to filter out images unrelated to buildings and images without a valid geotag or compass orientation. Therefore, a filtering pipeline is designed to identify images in the Flickr dataset that meet these criteria. For more details, interested reader is referred to (Hoffmann et al., 2023). Next, we briefly explain it.

2.1.1 Content filtering This step filters out images containing no buildings. It comprises Google Street View similarity filtering and object detection filtering.

First, we filter out non-street-view images from the Flickr dataset. The problem is approached as an image retrieval task, utilizing Google Street View images as the seed dataset and a Flickr dataset. Deep neural network features are utilized for identifying structurally similar images, extracting features from the last hidden layer of a pre-trained VGG16 network on ImageNet (Russakovsky et al., 2015). Cosine similarity is calculated on the resulting feature vectors. For the seed dataset, pre-calculated features are used. Then, pairwise cosine similarity is determined between Flickr images and the seed dataset. Images with similarity parameters below a predefined threshold are discarded.

Next, we apply object detection algorithm to ensure the presence of building facades in Flickr images. The algorithm identifies objects in the previously filtered images, generating a list of objects for each image. If this list contains a house or a building with a size parameter greater than the threshold and a confidence score higher than the threshold, the image qualifies for further processing.

2.1.2 Metadata filtering This step filters out images that cannot be geo-located. It focuses on the image's position and compass direction, that are crucial for calculating a line of sight for matching Flickr images with building footprints.

Geotags can be created automatically by a GPS sensor in the camera or manually by the user, and the latter is often inaccurate since users tend to tag images batch-wise, leading to slight position differences between GPS-tagged images taken without significant movement. To identify images with manually added geotags, we employ a heuristic filter. If two images have the same position, manual tagging without GPS data is suggested. These images are omitted from the subsequent steps.

Next, we check the metadata on the standard EXIF[1], which includes details like the capture date, camera model, settings, and GPS sensor data, such as latitude, longitude, and compass direction. The presence of the GPSImgDirection tag in the EXIF data is checked, and images lacking this tag are rejected.

[1] EXIF is a standard established by the Camera and Imaging Products Association (CIPA) and the Japan Electronics and Information Technology Industries Association (JEITA).

2.2 Mapping OpenStreetMap (OSM) building functions to Flickr images

2.2.1 Building function aggregation in OSM We obtain building footprints and their tags in OSM in the study area. OSM allows users to contribute mapping data in a Wikipedia-like manner. Though OSM provides guidelines for structuring and enriching data, there is no strict enforcement. Consequently, building tags are optional, with only building footprint coordinates being mandatory when added to OSM. OSM guidelines include three tags: building, amenity, and shop, used to indicate building functions.

We implement a classification scheme based on OSM guidelines, assigning each tag value (building, amenity, and shop) to one of three categories: commercial, residential, or other. If multiple tags are present, we ensure they are consistent in their classification. Disagreements among tags result in the building being unmapped to any class. However, if only one tag or all available tags agree on the same class, we assign that class to the building.

2.2.2 Matching Flickr images and OSM buildings In this step, we connect the buildings depicted in an image and their corresponding building footprints in OSM.

We utilize the image's position and compass direction from the EXIF data, which is essential for creating a line of sight. The line of sight identifies possible building candidates by intersecting with their polygons in OSM. From these candidates, we select the building with the closest distance to the image's position as the reference building.

2.3 Information extraction on Flickr images

2.3.1 STR on Flickr STR algorithms are applied on the Flickr images to extract texts on buildings, specifically, TextSnake (Long et al., 2018) for text detection and SAR (Li et al., 2019) for text recognition:

1. Text detection: TextSnake (Long et al., 2018) is a novel approach for text detection in natural scenes. Unlike conventional methods that represent text regions as bounding boxes or polygons, TextSnake models text instances as a sequence of pixels forming snakes. It combines convolutional neural networks and recurrent neural networks to predict text instances' positions and orientations simultaneously. The TextSnake method performs better in handling curved and arbitrarily shaped text instances, making it highly effective for scene text detection tasks.

2. Text recognition: SAR (Li et al., 2019) is based on an attention-based encoder-decoder framework. The model leverages visual attention mechanisms to focus on relevant regions of the input image, allowing it to recognize irregularly shaped and oriented text instances adaptively. The system demonstrates exceptional performance across various challenging scenarios, such as curved and distorted text, making it a robust and efficient baseline for addressing irregular text recognition tasks.

2.3.2 STR results filtering As no ground truth labels are available for the texts in the Flickr images, we apply the following criteria to filter the results obtained from the STR process. The aim is to eliminate STR results with lower confidence levels.

1. Text score and box score: Each recognized text in STR is associated with a box score and a text score, indicating the confidence of the detection and the recognition, respectively. Both the scores range from 0 to 1, and the larger number indicates higher confidence. Therefore, we filter the results using pre-defined thresholds on both scores.

2. Stopwords: Stopwords are common words that do not carry significant meaning, such as "the," "is," and "and,". While essential for sentence structure, they can add noise during text analysis. Filtering them out helps focus on meaningful words, improve efficiency, enhance relevance, and boost search accuracy. Since stopwords are unlikely frequently appear on building facades, we filter them out in STR results as misrecognition.

3. Repetitive letters: Lastly, we filter out text strings containing repetitive letters that are not words and are misrecognized from building structures, such as windows and balcony railings.

3. EXPERIMENTS AND DISCUSSION

3.1 Data and study area

The Flickr dataset in Berlin was used in our experiments, containing 3,431 street-view building images filtered from 929,508 Flickr images queried using the Flickr API. After matching the images with OSM building footprints, 1,833 (53.42%) are labeled as residential, 605 (17.63%) as commercial, and 993 (28.94%) as other.

3.2 Experiments and results: STR on Flickr

First, we applied STR on the Flickr images to extract texts on buildings, using pre-trained models, TextSnake (Long et al., 2018) and SAR (Li et al., 2019) for text detection and recognition, respectively, as introduced in Section 2. The results with both text score and box score > 0.8 were kept. Second, we filtered out results with stopwords in both English and German, and filtered out text strings with repetitive letters that are not words and were misrecognized from building structures.

After filtering, valid texts were recognized on 1,558 images (45.4% of the dataset). The number of recognized texts per image ranges between 1 and 32. Figure 1 shows examples of STR results. Table 1 summarizes the number of images of the Flickr Berlin dataset and the number of images with STR results for each building type. More STR-on-Flickr results can be found at `https://github.com/ya0-sun/STR-Berlin`.

3.3 Correctness of numbers recognized by STR

In 32 images, only numbers were recognized. We manually checked the STR results of these images to verify the results.

building types	*Flickr images*	*with texts in STR results*
residential	1,833	892
commercial	605	330
other	993	336
total	3,431	1,558

Table 1. Number of images of Flickr Berlin dataset and number of images with STR results for each building type.

3.3.1 Correctness of numbers in STR results Manual comparison confirmed that STR results on 29 images are correct. Although we could not manually label the whole dataset, the correctness of the recognized numbers, i.e., 90.62%, for the small subset, indicates an overall high accuracy of the STR results.

3.3.2 Objects containing recognized numbers Among the 29 correctly recognized images, 19 are on buildings, i.e., 65.52%, most of which are house numbers and construction years, e.g., Fig 2(a)(b), and a few are on building walls, e.g., Fig 2(c). Some numbers are on other objects, including static objects, such as speed limit signs and roads (Fig 2(d)(e)), and moving objects, such as vehicles and race runners' bibs. Table 2 summarizes the objects on which numbers are recognized.

The comparison suggests that removing non-building objects in a pre-processing step, e.g., the filtering process in Section 2, is necessary.

3.4 STR results and OSM building functions

We compared STR results with building function labels from OSM.

Intuitively, residential buildings are expected to have fewer texts on their facades, and commercial buildings should have more; however, no correlation is observed in our experiments.

3.4.1 Flickr images with texts recognized by STR STR recognized texts on 892 residential, 330 commercial, and 336 other images, respectively, i.e., about $1/2$ residential and commercial buildings have texts recognized, and the number for other buildings is about $1/3$.

Residential buildings are expected to have fewer texts. Thus, we investigated Flickr images of residential buildings with STR-recognized texts. Some examples are shown in Figure 3. We found two main contributors:

1. Shops on the ground floor or lower floors of residential buildings, as an example given in Figure 3(d). The most common businesses include café, restaurant, pharmacy, bakery, butcher, kiosk, and bank.

2. Texts on non-building objects that occlude the buildings. The most common objects are road signs, vehicle registration plates, and reserved parking signs, as examples shown in Figure 3(e) and (f).

The first case calls for a better definition of building types in OSM, e.g., adding mixed or secondary usage. These buildings have other usages besides their primary usage. Object detection can help solve the second case by removing non-building objects.

Object		*#image*
building	house number	15
	construction year	1
	other numbers on walls	3
none-building & static	street furniture	2
	road surface marking	1
none-building & moving	race bib	3
	vehicle	2
watermark	timestamp of the photo	1

Table 2. Objects and the corresponding number of images on which numbers are recognized in STR.

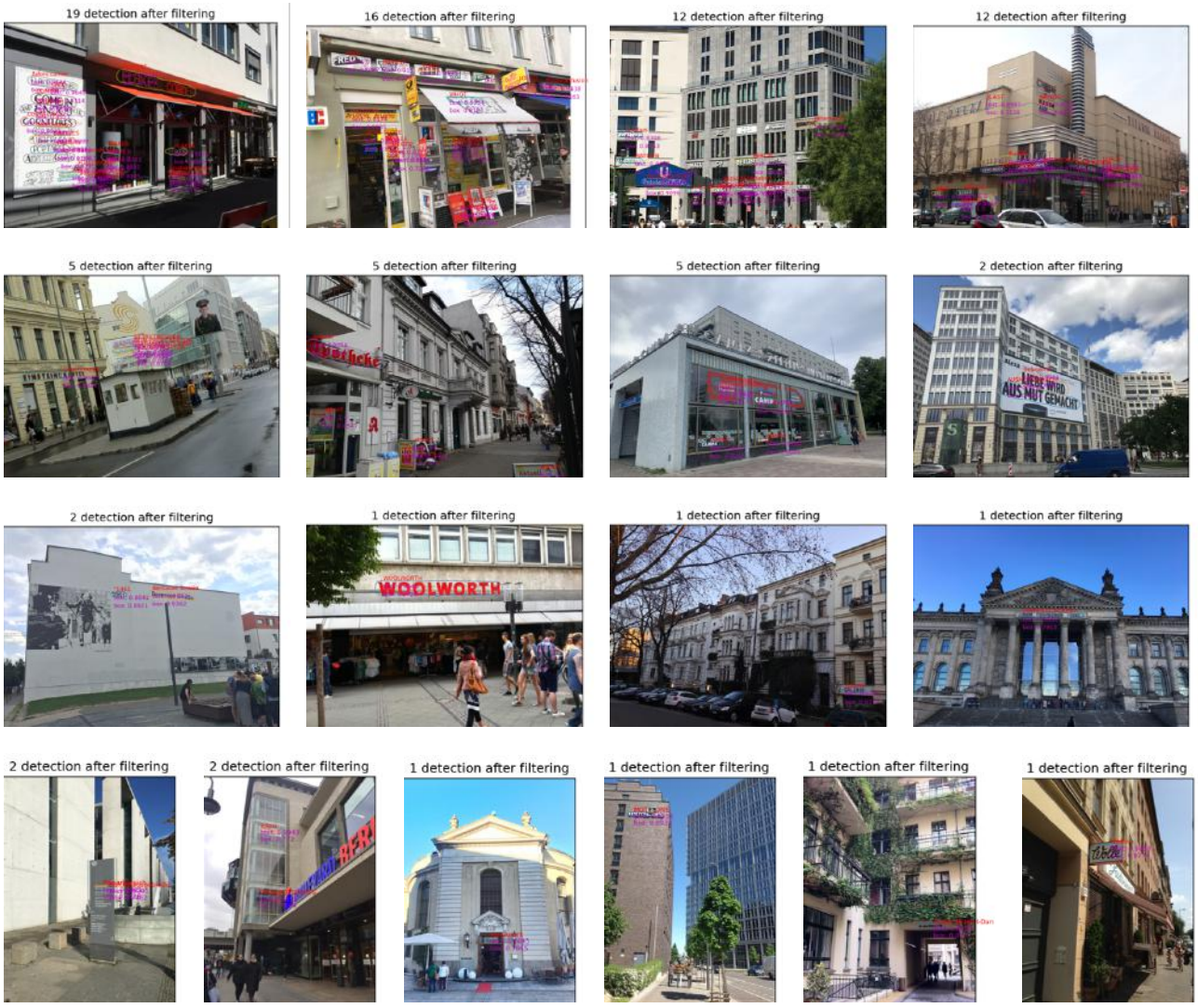


Figure 1. Examples of STR results on Flickr building images. The number of detection after filtering is written above each image and the STR results, i.e., texts/text scores and box/box scores, are displayed on the images at the corresponding location.

(a) '19' (b) '1952' (c) '1990' (d) '30' (e) '10'

Figure 2. Examples of number strings in STR results. The STR results are listed below the corresponding image as well as displayed on the images.

		Does the image contain:			*Is the building function recognizable by:*	
	Flickr example	*texts?*	*non-text signals?*	*target building occluded?*	*STR?*	*human?*
1	Figure 4 (a)(b)	yes	-	-	no	yes
2	Figure 4 (c)(d)(e)(f)	no	yes	-	no	yes
3	Figure 4 (g)(h)	no	no	-	no	no
4	Figure 4 (i)	-	-	yes	no	no

Table 3. Four cases without STR results on Flickr images featuring commercial buildings.

(a) (b) (c)

(d) (e) (f)

Figure 3. Examples of residential buildings: (a)(b)(c) have no text recognized in STR results; (d) shows a restaurant on the ground floor of a residential building; (e) and (f) show street name signs and car plates in front of the residential buildings leading to text recognized in STR results.

3.4.2 Flickr images with no text recognized by STR Texts were not recognized on $1/2$ residential and commercial building images and $2/3$ other building images.

Commercial buildings are expected to have more text on their facades. Thus we further investigated Flickr images of commercial buildings with no text recognized by STR. Some examples are shown in Figure 4. Further investigation of the dataset led us to four cases, listed below and summarized in Table 3:

1. Image contains texts that STR does not recognize because of insufficient image quality, e.g., resolution, light conditions, and blurs, but are easy for humans to recognize building functions. Examples are shown in Figure 4(a) and (b).

2. Image contains no text, but other human-readable signals, such as brand symbols, display windows, outdoor restaurant tables, and outside shop statues, can help classify building functions. Some examples shown in Figure 4(c)(d)(e)(f).

3. Image contains neither texts nor other signals, as shown in Figure 4(g)(h), and it is challenging for humans to classify building functions.

4. Buildings in the image are occluded or partially occluded, as shown in Figure 4(i), and it is difficult for humans to tell building functions.

In all the above situations, STR on Flickr, or STR on SVI in more general cases, is unsuitable for building function classification. The task's difficulty is especially addressed in Cases 3 and 4, where discerning building types is difficult for humans.

3.5 STR failure case analysis

We observed seven common failure cases in the results that constitute a common challenge for STR on street-view images, and we grouped them below according to the algorithm, the data, and the task, and show some examples in Figure 5:

1. Related to STR algorithms

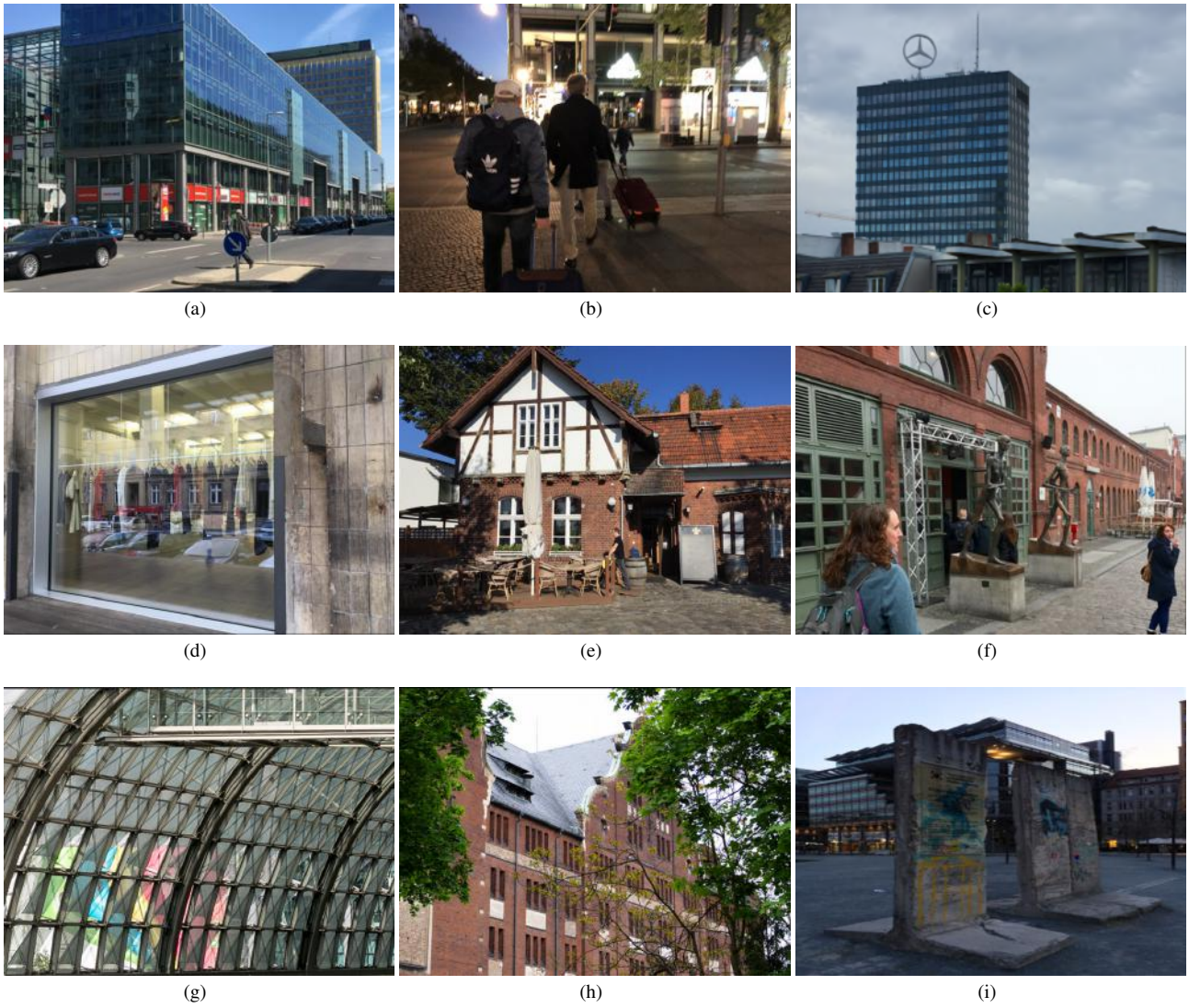
   (a) *Special characters:* existing models are mostly trained on datasets of English alphanumeric characters, which fail to predict special characters. Figure 5(a) shows an example that on the shop sign, 'ä' and 'ü' in German are recognized as 'a' and 'u.'

   (b) *Font styles*: texts and graffitis are often misrecognized because of font styles. The diverse character expression requires the models to recognize generalized visual features. Such as an example is shown in Figure 5(b): texts in regular fonts are recognized, but not the store name between 'record' 'store' and 'CAFE' 'BAR.'

   (c) *Low resolution or lighting condition:* The models can not handle low-resolution images or photos taken in insufficient lighting conditions. In our test, we accessed the high-resolution Flickr data, but it remains problematic if one uses lower-resolution crowdsourced images, such as Mapillary. Super-resolution modules may improve performance.

2. Related to the images

   (a) *Occlusion:* partially occluded texts on buildings are not well recognized. Other objects occluding buildings lead to wrong results for building information extraction. Figure 5(c) shows an example.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 4. Examples of commercial buildings with no STR results.

(a) 'Vollkornbackerei', 'Steinmuhle' (b) 'record', 'store', 'CAFE', 'BAR', ... (c) 'hopopol', 'BERLINER'

(d) 'Message-ID:' (e) '1993' (f) 'HOTEL'

Figure 5. Examples of failure cases in STR results. The STR results are listed below the corresponding image as well as displayed on the images.

(b) *Multiple buildings in one image*: in the current workflow, one Flickr image is associated with one building footprint in OSM data. However, it is common for multiple buildings to be contained in one image, such that errors occur as texts on one building are mapped in another. Future work should improve the matching between Flickr images and OSM data to solve this problem.

(c) *Watermark:* Watermarks on images added by users before uploading or time stamps added by digital cameras often consist of texts or numbers that need to be removed in the STR results.

3. Specifically related to the task of STR on SVI

Non-text repetitive patterns: we observed that the models misrecognize repetitive patterns on buildings, such as windows and balcony railings, as texts, as shown in Figure 5(d)(e)(f).

The STR-recognized text 'HOTEL' in Figure 5(f) is particularly misleading and challenging to identify. We suspect that the domain gap between the STR datasets used for training and the street-view images obtained from Flickr be the reason for the observed issues.

4. CONCLUSION AND FUTURE WORK

This work explores building attribute mapping with crowdsourced street-view images, specifically focusing on texts on building facades. We create a Berlin Flickr dataset and employ pre-trained STR models for text detection and recognition. Since ground truth labels for texts are unavailable, we manually checked a subset of images recognized by STR, indicating high accuracy. We further analyzed the relationship between building functions and text recognition, finding no clear correlation.

We identified three main reasons that impose challenges to the task of building attribute mapping using STR: *First, the discrepancy between street-view images and popular datasets for STR tasks.* Images in STR datasets, e.g., COCO Text (Veit et al., 2016), are often text-centric, but street-view images often feature more objects in larger areas and have much smaller text regions, making text recognition challenging, especially in complex and cluttered scenes. *Second, lack of ground truth labels.* Without ground truth, evaluating STR methods on numerous images is unrealistic. Although manual checking on a subset of images with numerical texts shows promising results, further assessment is required to validate performance and subsequently map texts as building attributes. *Third, inaccurate mapping between Flickr images and building footprints.* Currently, each Flickr image is associated with only one building footprint, causing a mismatch of texts from other buildings or objects in the image to the building footprint. Future improvements are needed for matching between Flickr images and OSM building footprints. These findings underline the constraints of STR algorithms on crowdsourced street-view images and the importance of labels in scenes, which will be addressed in future works toward large-scale building attribute mapping.

It is worth noting that the STR approach is suited for buildings with visible texts on their facades but not other cases. Considering alternative approaches or data sources is necessary for large area mapping. In addition, this study views Flickr data as a valuable resource to expand street view data through crowdsourcing and did not investigate its distribution within the city. However, considering the relationship of Flickr image locations with hotspots in cities, interdisciplinary collaboration may be essential to understand the reasons behind building photo-

graphy and labeling in the dataset and the needed approaches to map building attributes not only for hot spots but for the entire city.

ACKNOWLEDGEMENT

The work is jointly supported by the German Research Foundation (DFG) under the grant ZH 498/14-1 and ME 1846/16-1 for the project OpenStreetMap Boosting using Simulation-Based Remote Sensing Data Fusion (Acronym: OSMSim) and by the Technical University of Munich (TUM) Georg Nemetschek Institute under the project Artificial Intelligence for the automated creation of multi-scale digital twins of the built world (Acronym:AI4TWINNING).

References

Brenner, C., 2005. Building reconstruction from images and laser scanning. *International Journal of Applied Earth Observation and Geoinformation*, 6(3), 187–198.

Brunner, D., Lemoine, G., Bruzzone, L., Greidanus, H., 2010. Building Height Retrieval From VHR SAR Imagery Based on an Iterative Simulation and Matching Technique. *IEEE Transactions on Geoscience and Remote Sensing*, 48(3), 1487–1504.

Chaudhary, P., D'Aronco, S., Moy de Vitry, M., Leitão, J. P., Wegner, J. D., 2019. Flood-water level estimation from social media images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4(2/W5), 5–12.

Chen, F.-C., Subedi, A., Jahanshahi, M. R., Johnson, D. R., Delp, E. J., 2022. Deep learning–based building attribute estimation from Google Street View images for flood risk assessment using feature fusion and task relation encoding. *Journal of Computing in Civil Engineering*, 36(6), 04022031.

Chen, S., Mou, L., Li, Q., Sun, Y., Zhu, X. X., 2021. Mask-height R-CNN: An end-to-end network for 3D building reconstruction from monocular remote sensing imagery. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*.

Hoffmann, E. J., Abdulahhad, K., Zhu, X. X., 2023. Using social media images for building function classification. *Cities*, 133, 104-107.

Huang, B., Zhao, B., Song, Y., 2018. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sensing of Environment*, 214, 73–86.

Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X. X., 2018. Building instance classification using street view images. *ISPRS journal of photogrammetry and remote sensing*, 145, 44–59.

Langemeyer, J., Calcagni, F., Baró, F., 2018. Mapping the intangible: Using geolocated social media data to examine landscape aesthetics. *Land use policy*, 77, 542–552.

Leung, D., Newsam, S., 2012. Exploring geotagged images for land-use classification. *Proceedings of the ACM multimedia 2012 workshop on Geotagging and its applications in multimedia*, 3–8.

Li, H., Wang, P., Shen, C., Zhang, G., 2019. Show, attend and read: A simple and strong baseline for irregular text recognition. *Proceedings of the AAAI conference on artificial intelligence*, 33(01), 8610–8617.

Li, Q., Mou, L., Hua, Y., Sun, Y., Jin, P., Shi, Y., Zhu, X. X., 2020. Instance segmentation of buildings using keypoints. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*.

Li, Y., Chen, Y., Rajabifard, A., Khoshelham, K., Aleksandrov, M., 2018. Estimating building age from google street view images using deep learning (short paper). *10th international conference on geographic information science (GIScience 2018)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Liasis, G., Stavrou, S., 2016. Satellite images analysis for shadow detection and building height estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 119, 437–450.

Long, S., Ruan, J., Zhang, W., He, X., Wu, W., Yao, C., 2018. Textsnake: A flexible representation for detecting text of arbitrary shapes. *Proceedings of the European conference on computer vision (ECCV)*, 20–36.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al., 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115, 211–252.

Sportouche, H., Tupin, F., Denise, L., 2011. Extraction and Three-Dimensional Reconstruction of Isolated Buildings in Urban Scenes From High-Resolution Optical and SAR Spaceborne Images. *IEEE Transactions on Geoscience and Remote Sensing*, 49(10), 3932–3946.

Sun, Y., 2016. 3D building reconstruction from spaceborne TomoSAR point cloud. Master's thesis, Technical University of Munich.

Sun, Y., Auer, S., Meng, L., Zhu, X. X., 2023. Artificial Intelligence based Building Attributes Enrichment in OpenStreetMap using Street-view Images. *Abstracts of the ICA*, 6, 250.

Sun, Y., Hua, Y., Mou, L., Zhu, X. X., 2021. CG-Net: Conditional GIS-Aware Network for Individual Building Segmentation in VHR SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, 1-15.

Sun, Y., Mou, L., Wang, Y., Montazeri, S., Zhu, X. X., 2022. Large-scale building height retrieval from single SAR imagery based on bounding box regression networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184, 79-95.

Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S., 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*.

Yan, Y., Huang, B., 2022. Estimation of building height using a single street view image via deep neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 192, 83–98.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P. M., 2019. Joint Deep Learning for land cover and land use classification. *Remote sensing of environment*, 221, 173–187.