Localization and Change Detection Through Aerial Environment Perception

Vivian Chiciudean, Horatiu Florea, Florin Oniga and Sergiu Nedevschi
Computer Science Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
FirstName.LastName@cs.utcluj.ro

***Abstract*—Urban development is advancing at an exceedingly rapid pace, significantly complicating the task of updating reference maps and requiring considerable effort in terms of actualization and monitoring. To address this challenge, we propose an automated mechanism for detecting changes in maps, specifically at the building level. Our method integrates cadastral maps with low-altitude aerial imagery, identifying discrepancies between the reference maps and the perceived images. We use captured images from standard Unmanned Aerial Vehicles and derive the cadastral maps from OpenStreetMap. The change detection process employs a mechanism that localizes the drone's position by aligning the perceived scene with the reference map. We utilize established foundation models, such as Grounded SAM and Depth Anything, for image-level perception based on semantic segmentation and depth estimation, capitalizing on their robustness and generalization capabilities. Subsequently, we construct a bird's-eye view representation that mirrors the reference map and generate a set of discriminative features. The obtained features are utilized to formulate hypotheses as rigid transformations, which are tested by projecting the image onto the reference map and validated using the Intersection over Union (IoU) metric. Once the perceived image is aligned with the reference map, we obtain the global GPS position and camera orientation, implicitly determining the global location of the buildings. After achieving accurate localization, the focus shifts to identifying differences between the perceived image and the reference map. In terms of experimental results, we tested the method in an aerial environment using a subset of the UAVid dataset, considering significant GPS disturbances, and achieving high precision in localization regarding both the position and the camera orientation.**

Keywords*—Single Image Localization, Aerial Environment Perception, Change Detection

I. Introduction

The swift and extensive expansion of urban areas poses a substantial challenge to the effective monitoring of urban development. Conventional methods, which rely on manual annotations and community efforts [1] to update reference maps, are both labor-intensive and time-consuming. As urban landscapes evolve at an unprecedented pace, relying on manual updates for maps becomes increasingly unfeasible, thereby highlighting the need for more efficient and automated solutions.

Recent advancements in foundation models [2] have revolutionized computer vision, significantly enhancing our ability to analyze and interpret visual information. Notable improvements have been observed in tasks such as object detection [3], instance segmentation [4], semantic segmentation [5], video panoptic segmentation [6], visual question answering [7], and monocular depth estimation [8].

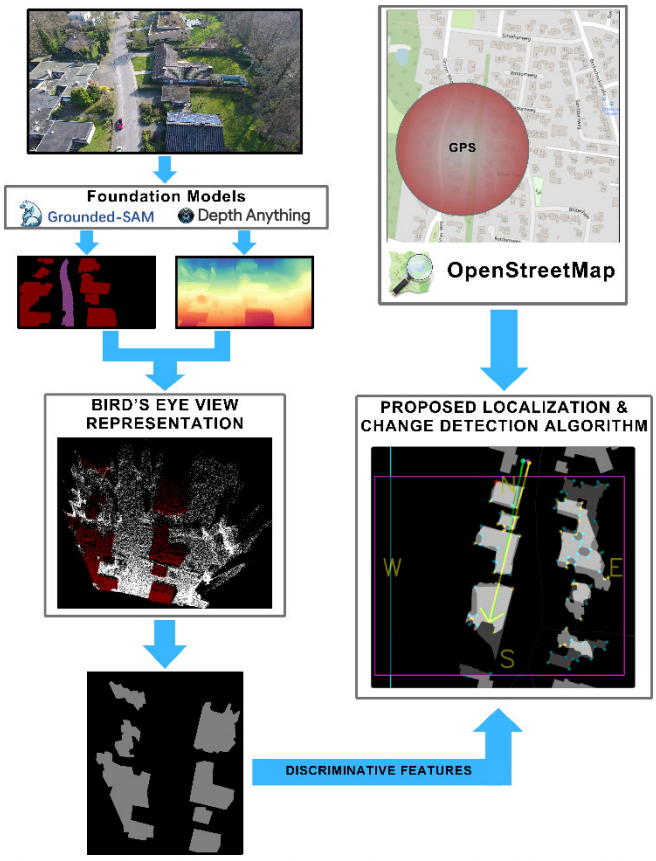


Fig. 1. Overview of the proposed approach. We use foundation models to obtain the aerial environment perception, fuse them into a bird's-eye view representation, extract relevant data (building's footprint) and localize the perceived image into a reference map.

Models such as CLIP [9], DINO [10] [11], Grounded-DINO [12], Segment Anything [13], Grounded SAM [14], and Depth Anything [15] have demonstrated exceptional performance and adaptability across various applications. Their increased generalization capabilities, resulting from training on diverse and large datasets, enable a deep understanding of visual characteristics and a more comprehensive perception of their surroundings.

Unmanned Aerial Vehicles (UAVs), commonly known as drones, are increasingly utilized for various surveillance and monitoring applications. However, the onboard GPS sensors in standard systems can exhibit errors of up to tens of meters [16] and are especially vulnerable to transmission jamming. While RTK GPS [17] systems can address these issues by offering precise and reliable positioning, they are expensive and require a ground-based reference station.

To address the challenges posed by the rapid development of new buildings and the necessary manual updates of existing

maps, we propose a processing pipeline (Fig. 1) that performs a precise localization and detects building changes in an urban environment. This approach leverages the integration of cadastral maps with low-altitude aerial imagery to identify discrepancies between reference maps and current observations. By employing standard drones to capture aerial views and utilizing OpenStreetMap [18] [19] (OSM) for cadastral map emulation, our method addresses the challenge of maintaining up-to-date maps in dynamic urban environments. The proposed localization and change detection pipeline addresses the limitations of unreliable GPS signals by aligning the perceived scene with a reference map. This alignment utilizes pre-trained models, such as Depth Anything [15] and Grounded SAM [14], for depth estimation and semantic segmentation. These models provide robust and generalized image-level perception, which is essential for accurate localization and change detection. The proposed pipeline also involves constructing a bird's-eye view (BEV) that mirrors the reference map representation. This representation is then used to extract discriminative features (90º angles and length of building sides). The obtained features are utilized to formulate hypotheses, which are tested by projecting the image onto the reference map and evaluated using the Intersection over Union (IoU) metric. Subsequently, to address perception errors, we search for pairs of valid hypotheses that result in identical rigid transformations. This process ensures accurate localization and enables a detailed comparison between the perceived image and the reference map, revealing any changes that have occurred.

To validate the effectiveness of our proposed pipeline for localization and change detection, we conducted experiments using the UAVid [20] dataset and the corresponding OSM [18] [19] areas. We employed traditional photogrammetry techniques [21] [22] to obtain precise camera poses and parameters, which were then aligned with the environment to achieve accurate global GPS measurements. We focused on a specific area within the real-world neighborhood where the images were captured and compared the outputs of our method with the GPS coordinates. For each image, we considered a search area of 120 x 120 meters without information on camera orientation. The extensive testing revealed an average heading error of 1 degree and a translation error of 1.5 meters, demonstrating superior accuracy compared to standard GPS systems typically used in UAVs. Additionally, we showcased the change detection algorithm across different experimental setups, highlighting the pipeline's capabilities in urban scenarios.

We summarize the main contributions of our work as follows:

- The development of a pipeline for localization and change detection of buildings in reference maps
- The development of an algorithm for low-altitude aerial imagery localization in urban scenarios, utilizing the hypothesis generation and validation method
- The development of a procedure for converting the image-level perception to the reference map representation level
- The combination of photogrammetry-based approaches with pretrained foundation models to obtain a BEV representation

II. Related Work

A. Image-based Localization

Traditional image-based localization typically relies on texture (RGB) images and feature matching techniques. Common approaches utilize SIFT-based (Scale-Invariant Feature Transform) algorithms [23] to detect, describe, and match local features. These features are then used to compare the perceived image with a georeferenced image. Thus, identifying the exact correspondence and the necessary transformation to localize the initial image. A better approach employs the FAST key point detector [24] and the BRIEF descriptor [25], specifically ORB-based (Oriented FAST and Rotated BRIEF) algorithms [26]. However, these methods have limited functionality due to their dependence on georeferenced images and the presence of similar texture-level features, making them significantly susceptible to environmental factors.

Recently, various approaches have addressed the problem of image-based localization using end-to-end deep learning techniques. PlaNet [27] formulates localization as a classification problem, where the current position is matched to the most appropriate location in the training set. PoseNet [28] [29], on the other hand, approaches 6DoF pose estimation as a regression problem. The authors of [30] combine a convolutional neural network (CNN) with long short-term memory (LSTM) units to create a direct mapping from the input image to the camera pose. AtLoc [31] demonstrates that attention mechanisms can be employed to direct the network's focus towards more geometrically robust objects and features. Additionally, MapNet [32] facilitates the learning of data-driven map representations by integrating various sensory inputs, such as visual odometry and GPS, alongside images, to enhance camera localization.

Despite the potential benefits of specialized neural networks for localization in specific scenarios, the scarcity of publicly available datasets presents a significant challenge to their generalization capabilities. Consequently, models are often trained to operate effectively only in particular environments. Furthermore, the aforementioned approaches typically utilize street-level views captured from main roads, which restricts the amount of information gathered compared to aerial views. To overcome these limitations, we leverage the generalization capabilities of foundation models to develop a robust localization system for aerial imagery.

B. Change Detection

Change detection based on aerial data is a crucial method for urban planning, environmental monitoring, disaster assessment, and map revision on the Earth's surface [33]. Traditionally, change detection has been addressed by integrating various types of data from multiple sources. Commonly used data include 3D LiDAR scans of real-world environments or images acquired through remote sensing and satellite views [34]. Nevertheless, these kinds of data are very expensive to gather or suffer from poor spatial resolution, limiting the ability to capture low-level details. Another approach [35] involves scene understanding and captioning to detect changes between two different views of the environment, but these methods tend to perform well only in simplified indoor scenarios. To support and drive progress, several datasets [36] [37] [38] have been released to facilitate end-to-end training and the development of change detection systems. However, these datasets primarily focus on disaster

assessment, such as post-flood, hurricane, or earthquake scene understanding.

To address these limitations, we propose a novel approach for low-altitude scene understanding that converts aerial perceptions into representations compatible with reference maps. Our method uses high-resolution UAV imagery and the strong generalization capabilities of pretrained foundation models to localize and detect building-level changes, allowing change detection in complex urban environments.

III. Detailed Methodology

The inputs of the proposed pipeline are a reference map, an RGB texture image acquired from a drone, and a rough map area. The primary steps to achieve change detection regarding buildings are as follows:

A. Depth and semantic segmentation inference based on pretrained foundation models.

B. BEV representation.

C. Conversion to map representation.

D. Hypotheses generation.

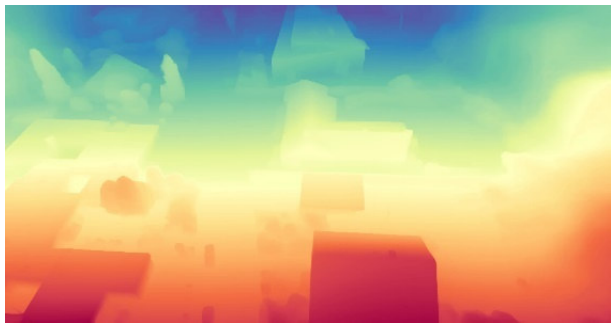
E. Hypotheses validation and pair searching.

To visually support the proposed approach, we utilize a subset of the UAVid [20] dataset. This UAV-acquired semantic dataset comprises 42 video sequences collected from two countries, China and Germany. We used approximate locations of two video sequences (Gronau, Germany) in conjunction with camera poses and intrinsics obtained using COLMAP [21] [22] through a Structure-from-Motion approach. Registering the 3D reconstructions with aerial LiDAR data of the locations [39] yields an accurate global localization that can be used for evaluation.

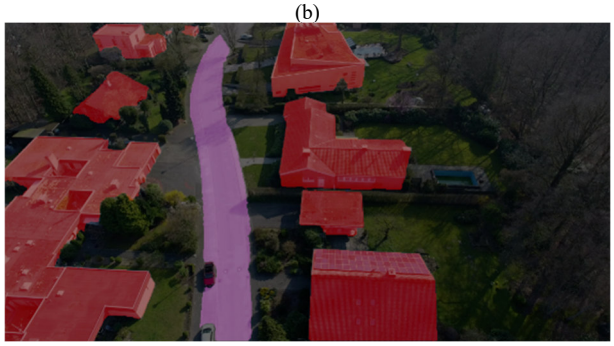
A. Depth and Semantic Segmentation Perception Through Foundation Models

The first step in the proposed pipeline is to understand the perceived image (Fig. 2 (a)). For this, we leverage two pretrained foundation models to obtain depth and semantic segmentation. Specifically, we use Depth Anything (v2) [15] to generate a metric monocular depth estimation (Fig. 2 (b)) of the perceived image. For the semantic segmentation (Fig. 2 (c)), we employ Grounded SAM [14] and use prompts such as “buildings” and “main road” to extract the most relevant information for our system.

(a)



(b)

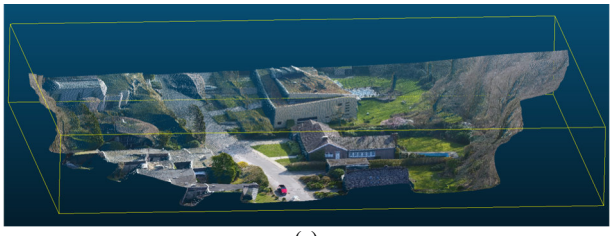


(c)

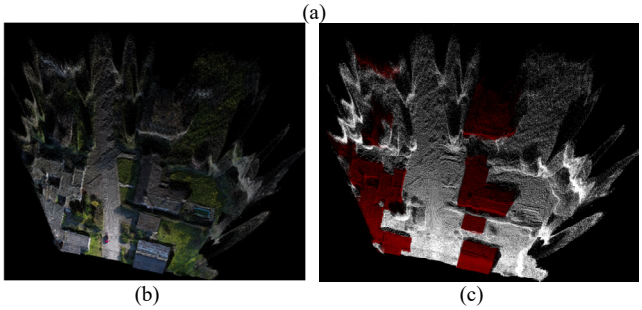
Fig. 2. Foundation model perception results: (a) initial RGB frame; (b) monocular depth estimation; (c) semantic segmentation (buildings and road) overlayed on top of the RGB frame.

B. BEV Representation

To further understand the scene, we use the intrinsic camera parameters and the scale information obtained from the 3D reconstruction process. We back-project each scaled depth point from the estimation (Fig. 2 (b)) to form a 3D point cloud. Simultaneously, we propagate the texture (Fig. 2 (a)) and semantic (Fig. 2 (c)) information into the 3D geometry, creating an enhanced point cloud with various information (Fig. 3). To obtain the BEV representation, we leverage the "road" segmentation and detect the plane that approximates the ground-level area. Next, we position our camera perpendicularly to that plane, center the point cloud, and project each visible 3D point, resulting in the RGB (Fig. 3 (b)) and semantic (Fig. 3 (c)) BEV representation. In this step, each pixel in the BEV representation corresponds to a 30 x 30-centimeter area of the real-world environment.



(a)



(b) (c)

Fig. 3. Obtaining a bird’s-eye view (BEV) representation: (a) obtained 3D point cloud with texture information; (b) texture BEV; (c) semantic BEV (buildings).

C. Conversion to map representation

Now, we shift our focus on converting the BEV to a representation similar to the reference map. From the BEV projection, we extract the top view of each building present in the perceived image (Fig. 4 (a)) and apply morphological operations to obtain a preliminary contour of them (Fig. 4 (b)). Next, we compute the polygons that approximate each building and discard those with an area less than a fixed threshold (Fig. 4 (c)).

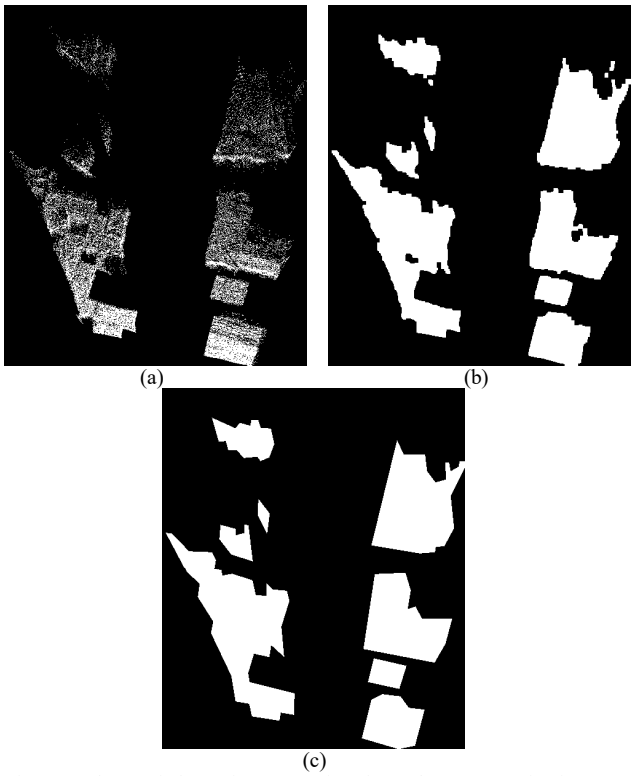


Fig. 4. Relevant information extraction from the BEV projection: (a) building's footprint, initial mask; (b) initial contour; (c) polygonal approximation of the building's footprint.

However, the current contours of the buildings are quite coarse and noisy compared to the OSM representation that we will use as a reference map. Therefore, a postprocessing step is required to close the gap between the two representations. For that, we first simplified the contours of the polygons by identifying the most representative points using the Shi-Tomasi Corner Detector [40]. The new contour is highlighted in red in Fig. 5 (a). Next, we iterated over each polygon and discarded every vertex forming an angle in the range of 170 to 190 degrees, as they did not significantly enhance the polygonal representation of the buildings and introduced unnecessary complexity. The vertices that were discarded are highlighted in red in Fig. 5 (b), while the final result is presented in Fig. 5 (c).

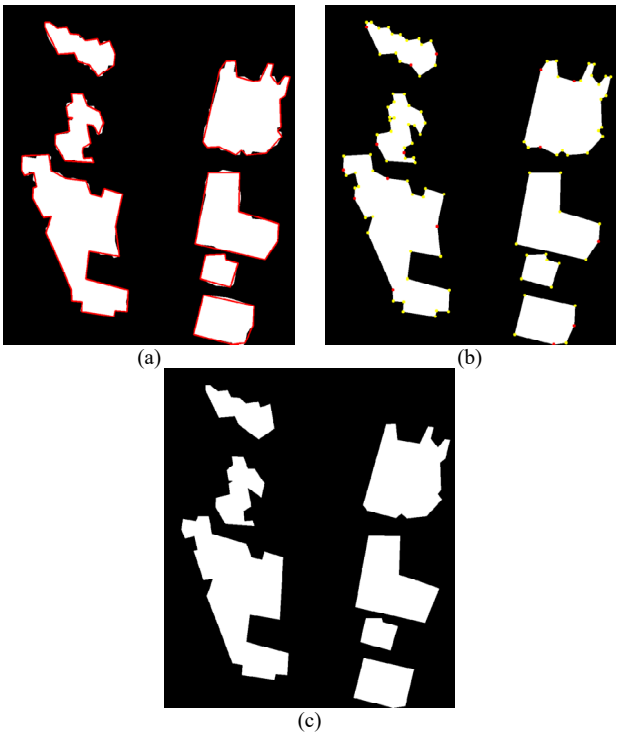


Fig. 5. Postprocessing operations for building's footprint extraction: (a) footprint obtained after Shi-Tomasi corner detection; (b) discarded vertices by angle of the sides; (c) final building's footprint, converted from perceived image to reference map representation.

D. Hypotheses generation

Once we obtain the best possible contour approximation from the bird's-eye view and convert our understanding of the scene into the map representation, we shift our focus to identifying discriminative features for hypothesis generation.

OpenStreetMap, which consists of lines, points, and polygons, serves as our reference map. Therefore, we need to identify features that are suitable for comparison. In our scenario, where buildings of various shapes may be present, the most suitable features for hypothesis generation are the vertices of polygons forming angles between 80 and 100 degrees. To further refine the method, we also consider the length of the edges at each vertex. To clarify the underlying idea, vertices with longer edges forming angles close to 90 degrees are likely the most reliable features.

Based on the identified features, we generate hypotheses and determine rigid transformations (i.e., one-dimensional rotations and 2D translations) between a vertex in the converted map representation of the perceived scene and each potential vertex within a specified area of the reference map. Each rigid transformation is verified for both rotation and translation within the reference map. The rotation is verified by comparing the angle between the right edge of a perceived building vertex and the corresponding right edge in the reference map, as well as the angle between the left edge of the perceived building and the corresponding left edge in the reference map. If the angle difference exceeds 10 degrees, the rotation is deemed invalid. Translation is validated using the coordinates of the area of interest. If the metric difference between the reference point and the perceived point exceeds the map size after accounting for rotation, the translation is considered invalid. Consequently, such transformations are excluded from hypothesis generation. Fig. 6 illustrates the process of generating valid rigid transformations from a set of discriminative features. These features are highlighted in blue

in both the cropped perceived image and the OSM reference map.

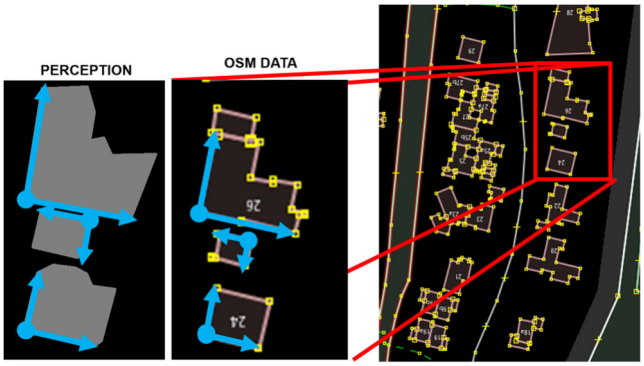


Fig. 6 Hypotheses generation and rigid transformation verification based on discriminative features: the discriminative features (highlighted in blue) present in the OSM-like perception image (left) compared to the reference data (right).

E. Hypotheses validation and pair searching

After generating valid rigid transformations, we focus on the hypotheses that were obtained. For each hypothesis, we apply the transformation to the perceived image, considering the search space of the reference map. We then compute the Intersection over Union (IoU) score between the buildings in the transformed perception image and those in the reference map. We retain only those hypotheses that achieve an IoU score greater than a specified threshold; our experiments indicate that 50% is an effective threshold. Subsequently, we search for pairs of valid hypotheses that yield nearly identical rigid transformations in terms of rotation and translation. This pair-searching strategy proves more robust, offering greater tolerance to perception errors and improved performance in scenarios where the perception image closely resembles different regions of the reference map. This process is illustrated in Fig. 7, where we highlight the best three hypotheses in red.

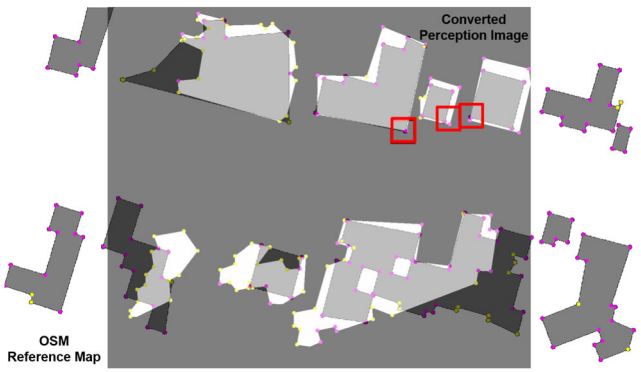
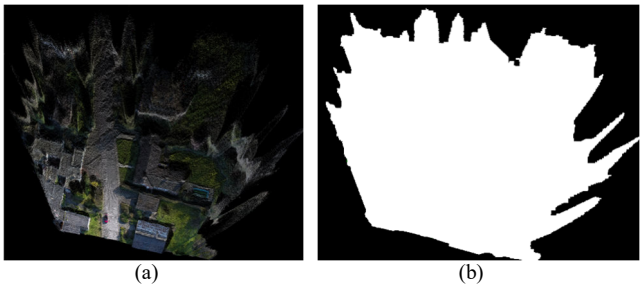


Fig. 7. Selecting the best valid hypothesis: converted perception image overlayed on top of the reference map based on the discriminative features with valid rigid transformations (highlighted in red).

Our experiments revealed that the most effective way to compute the IoU score and validate the hypotheses is by considering the 3D observable area of the perceived image. This area encompasses the entire real-world region captured by the image, as depicted in Fig. 8 (b). This approach enhances the robustness of our method by effectively managing scenarios with a limited number of buildings in the perceived image.



(a) (b)

Fig. 8. The observable area mask of the initially perceived image (Fig. 2 (a)) represented from a bird's-eye view: (a) texture BEV; (b) observable area mask.

IV. Experimental Results

We conducted experiments using a subset of the UAVid [20] dataset combined with OpenStreetMap [18] [19] data from various timestamps as reference maps to evaluate the performance of our proposed method.

For the UAVid dataset, which provides aerial views of the real-world environment, we selected two video sequences from Gronau, Germany: seq13 and seq31. Each sequence contains 901 frames, with images acquired at an altitude of approximately 50 meters and a camera pitch of 45 degrees. Each video covers an area of roughly 200 x 50 meters. Portions of the areas captured in these sequences are also visible on Google Earth (Fig. 9 left) and OpenStreetMap (Fig. 9 right). These sequences were selected due to the increased number of buildings visible in each video frame. Additionally, significant changes, including building modifications and demolitions, have occurred in the real-world environment since the data was acquired.

For the OpenStreetMap reference map, we assumed severe GPS errors with no information on camera orientation and a possible search area of around 120 x 120 meters relative to the last GPS position.

(a)

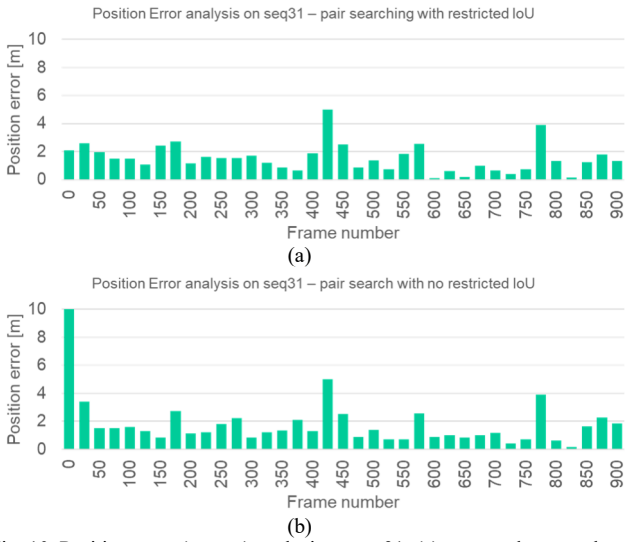
(b)

Fig. 9. Different satellite views of Gronau, Germany: (a) seq13; (b) seq31.

We used approximate locations of the two video sequences along with 3D reconstructions of the scenes generated using COLMAP [21] [22]. Ground truth global GPS localization needed for evaluation was recovered through registration of the 3D reconstruction [39] with aerial LiDAR data of the flight areas.

The quantitative evaluation results for seq31, using reference data from 2024, are presented in terms of position and orientation errors, as well as Intersection over Union (IoU) with the reference map. These results illustrate the effectiveness of localization across different frames.

For translation errors (Fig. 10 (a)), the average error is 1.52 meters, with a standard deviation of 0.99 meters, a maximum of 5 meters, and a minimum of 0.1 meters. Fig. 10 compares these results with those obtained without using the observable area mask in the IoU calculation (Fig. 10 (b)). The ablation study reveals that omitting the restricted IoU computation leads to a significant increase in the position error. We obtained an average of 6.18 meters, a standard deviation of 27.76 meters, a minimum error of 0.14 meters, and a maximum error of 172.62 meters.

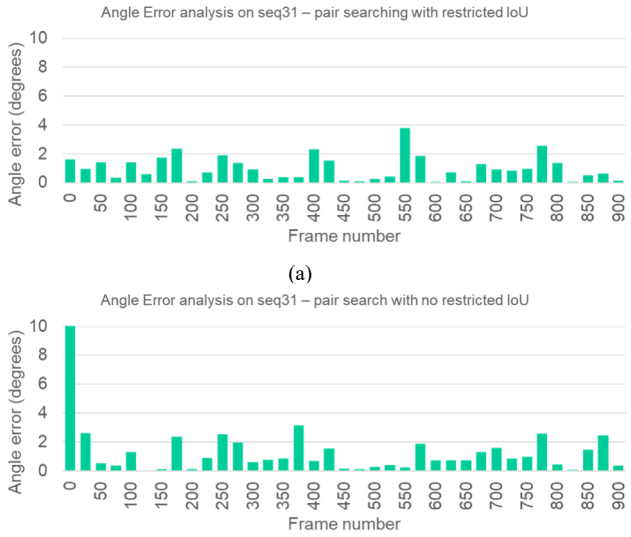


(a)

(b)

Fig. 10. Position error (meters) analysis on seq31: (a) proposed approach; (b) ablation study without using the observable area mask.

Regarding heading errors, the results are shown in Fig. 11. The proposed approach (Fig. 11 (a)) achieves an average rotation error of 1 degree, with a standard deviation of 0.84 degrees, a maximum of 3.77 degrees, and a minimum of 0.07 degrees. If the restricted IoU computation is not used, as depicted in Fig. 11 (b), the results are notably worse. The average error increases to 2.62 degrees, with a standard deviation of 9.52 degrees, a minimum error of 0.02 degrees, and a maximum error of 59.5 degrees.

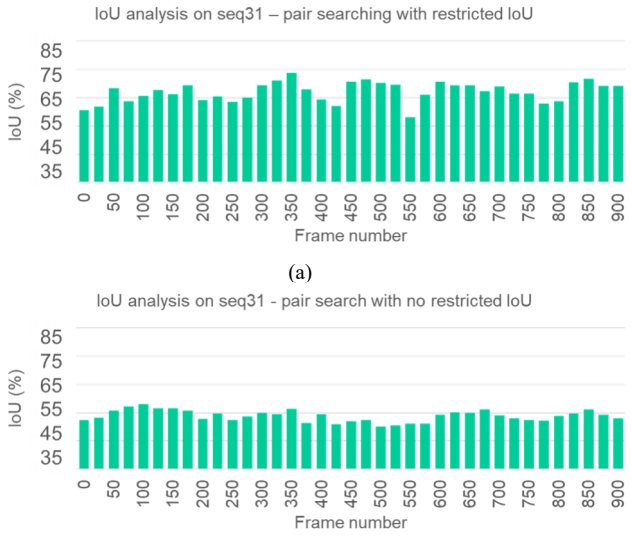


(a)

(b)

Fig. 11. Angle error (degrees) analysis on seq31: (a) proposed approach; (b) ablation study without using the observable area mask.

Additionally, for mean IoU (Fig. 12), the proposed approach (Fig. 12 (a)) shows substantial improvements over the ablation study (Fig. 12 (b)). We achieve a mean IoU of 67%, with a standard deviation of 3%, a maximum of 74%, and a minimum of 58%. In contrast, the ablation study reports a maximum IoU of only 58%, a minimum of 50%, an average of 54%, and a standard deviation of 2%.



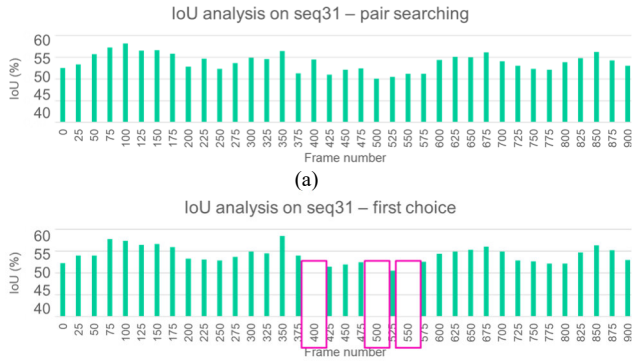
(a)

(b)

Fig. 12. Intersection over Union (%) analysis on seq31: (a) proposed approach; (b) ablation study without using the observable area mask.

We also conducted an ablation study to compare two strategies: selecting the hypothesis with the highest IoU score below a specified threshold (Fig. 13 (b)), versus searching for a pair of potential hypotheses to derive a final transformation (Fig. 13 (a)). We found that the first approach fails to correctly localize the perceived image in some corner cases, whereas

the proposed approach is more robust and consistently performs better.



(b)

Fig. 13. Intersection over Union (%) analysis on the searching strategy: (a) pair searching, proposed approach; (b) first choice, unlocalized frames are highlighted with pink rectangles.

As a qualitative evaluation, we observed that the proposed method performs well even with sudden camera movements and rotations (Fig. 14). Our approach provides better results compared to GPS interpolation based on two previous accurate GPS positions, considering an update frequency of around one second, which is typical for most UAVs.

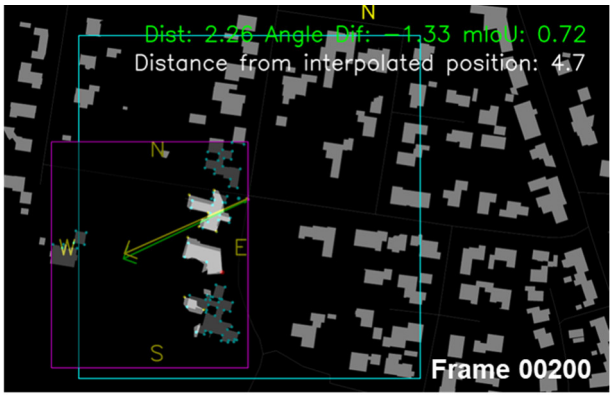


Fig. 14. Tolerance to sudden camera rotations (UAVid, seq13, frame 200). We represented the localized area with a magenta rectangle and the search space with a cyan rectangle.

Subsequently, we evaluated the change detection mechanism to demonstrate its effectiveness using data from different time periods. The UAVid video sequences were collected in 2018, which allowed us to use these sequences to qualitatively assess the proposed approach. We selected video sequence 13 (seq13) from UAVid and alternated the OpenStreetMap official data between 2018 and 2024 for the reference map. Localization was performed in both scenarios, revealing that the proposed approach effectively accommodates scene changes over time (Fig. 15).

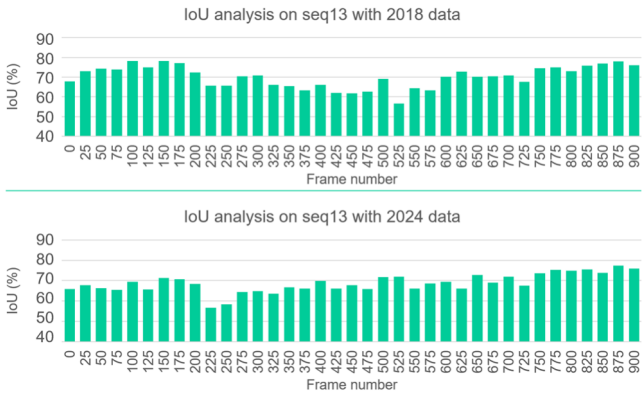
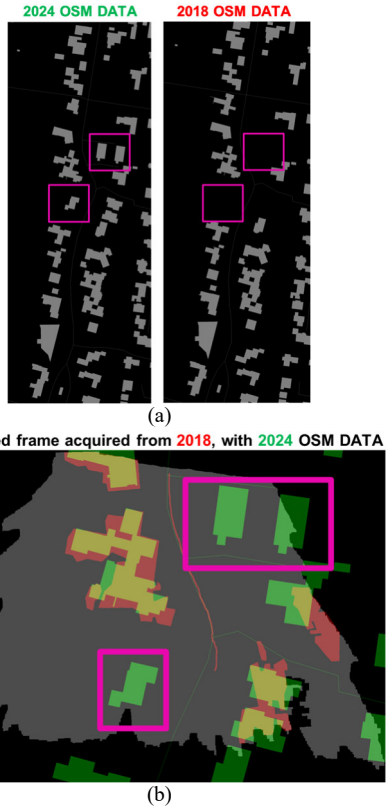


Fig. 15. Toleration to scene changes using acquired images from 2018: top – results on reference data from 2018; bottom – results on reference data from 2024.

For a qualitative evaluation, we examined areas from the reference map that had different buildings at the time the images were acquired. We aimed to detect changes in building shapes or the construction of new structures. Newly constructed buildings are marked in green (Fig. 16), while modified buildings are marked in red (Fig. 17). The yellow area represents buildings that are present both in the perceived image and in the OSM official data. To minimize detection noise, we considered the observable area and applied an intersection threshold at the building level.



(a)

(b)

Fig. 16. Change detection in a reference map with UAV-acquired images – new buildings: (a) reference map from different times; (b) converted perception image (2018) overlayed on top of the reference map (2024).

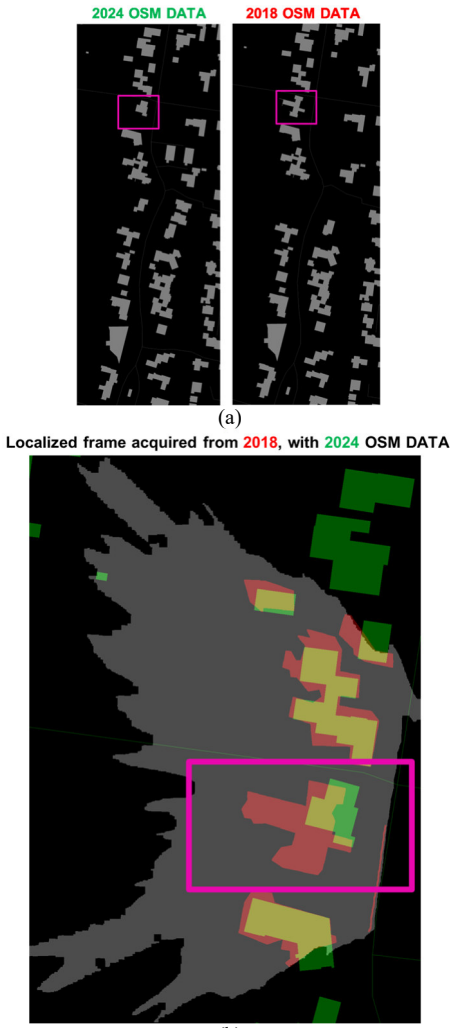


Fig. 17. Change detection in a reference map with UAV-acquired images – modified building: (a) reference map from different times; (b) converted perception image (2018) overlayed on top of the reference map (2024).

V. CONCLUSIONS

In this paper, we proposed a processing pipeline for image-based localization and change detection of buildings. The pipeline generates a bird's-eye view representation of the scene and extracts discriminative features from a single image. These features are then used in a hypothesis generation and validation step to determine valid rigid transformations. The resulting rigid transformations localize the initial image within the reference map, providing the corresponding global GPS position of the camera and implicitly localizing the buildings. Subsequently, change detection is performed as an additional step. Our localization evaluation revealed an average heading error of 1 degree and a translation error of 1.5 meters over a search area of 120 x 120 meters, with no prior orientation information. Additionally, we demonstrated that the change detection component shows promising results. In terms of future work and improvements, the temporal fusion of multiple frames may enhance building contour accuracy and help mitigate depth bleeding and occlusion issues. We also plan to explore the potential of updating cadastral maps through the proposed localization and change detection algorithm.

ACKNOWLEDGMENT

The research reported in this paper was supported by the Lockheed Martin Corporation and the "DeepPerception - Deep Learning Based 3D Perception for Autonomous Driving" grant funded by the Romanian Ministry of Education and Research, code PN-III-P4-PCE-2021-1134.

REFERENCES

[1] P. Mooney, M. Minghini and others, "A review of OpenStreetMap data," *Mapping and the citizen sensor,* p. 37–59, 2017.

[2] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang and F. S. Khan, "Foundational models defining a new era in vision: A survey and outlook," *arXiv preprint arXiv:2307.13721,* 2023.

[3] Z. Zou, K. Chen, Z. Shi, Y. Guo and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE,* vol. 111, p. 257–276, 2023.

[4] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: state of the art," *International journal of multimedia information retrieval,* vol. 9, p. 171–189, 2020.

[5] S. Hao, Y. Zhou and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing,* vol. 406, p. 302–321, 2020.

[6] D. Kim, S. Woo, J.-Y. Lee and I. S. Kweon, "Video panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[7] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015.

[8] C. Godard, O. Mac Aodha, M. Firman and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark and others, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021.

[10] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[11] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby and others, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193,* 2023.

[12] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu and others, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499,* 2023.

[13] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo and others, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[14] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan and others, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159,* 2024.

[15] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[16] K. N. Tahar and S. S. Kamarudin, "UAV onboard GPS in positioning determination," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* no. 41, pp. 1037-1042, 2016.

[17] T. Takasu and A. Yasuda, "Evaluation of RTK-GPS performance with low-cost single-frequency GPS receivers," in *Proceedings of international symposium on GPS/GNSS*, 2008.

[18] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive computing,* vol. 7, p. 12–18, 2008.

[19] J. E. Vargas-Munoz, S. Srivastava, D. Tuia and A. X. Falcao, "OpenStreetMap: Challenges and opportunities in machine learning and remote sensing," *IEEE Geoscience and Remote Sensing Magazine,* vol. 9, p. 184–199, 2020.

[20] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS journal of photogrammetry and remote sensing,* vol. 165, p. 108–119, 2020.

[21] J. L. Schönberger, E. Zheng, M. Pollefeys and J.-M. Frahm, "Pixelwise View Selection for Unstructured Multi-View Stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[22] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision,* vol. 60, p. 91–110, 2004.

[24] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, 2006.

[25] M. Calonder, V. Lepetit, C. Strecha and P. Fua, "Brief: Binary robust independent elementary features," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, 2010.

[26] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International conference on computer vision*, 2011.

[27] T. Weyand, I. Kostrikov and J. Philbin, "Planet-photo geolocation with convolutional neural networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, 2016.

[28] A. Kendall, M. Grimes and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015.

[29] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *2016 IEEE international conference on Robotics and Automation (ICRA)*, 2016.

[30] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck and D. Cremers, "Image-based localization using lstms for structured feature correlation," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[31] B. Wang, C. Chen, C. X. Lu, P. Zhao, N. Trigoni and A. Markham, "Atloc: Attention guided camera localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[32] S. Brahmbhatt, J. Gu, K. Kim, J. Hays and J. Kautz, "Geometry-aware learning of maps for camera localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[33] W. Shi, M. Zhang, R. Zhang, S. Chen and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sensing,* vol. 12, p. 1688, 2020.

[34] K. Li, X. Cao and D. Meng, "A new learning paradigm for foundation model-based remote-sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 62, p. 1–12, 2024.

[35] Y. Qiu, S. Yamamoto, R. Yamada, R. Suzuki, H. Kataoka, K. Iwata and Y. Satoh, "3d change localization and captioning from dynamic scans of indoor scenes," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

[36] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari and R. R. Murphy, "Floodnet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access,* vol. 9, p. 89644–89654, 2021.

[37] R. Gupta and M. Shah, "Rescuenet: Joint building segmentation and damage assessment from satellite imagery," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021.

[38] A. Fujita, K. Sakurada, T. Imaizumi, R. Ito, S. Hikosaka and R. Nakamura, "Damage detection from aerial images via convolutional neural networks," in *2017 Fifteenth IAPR international conference on machine vision applications (MVA)*, 2017.

[39] H. Florea and S. Nedevschi, "TanDepth: Leveraging Global DEMs for Metric Monocular Depth Estimation in UAVs," in *arXiv preprint arXiv:2409.05142* , 2024.

[40] J. Shi, "Good features to track," in *Proceedings of IEEE conference on computer vision and pattern recognition*, 1994.