# One-Shot Indoor Positioning Using 360-Degree Photos

**Daeyoung Gil[a] and Ghang Lee[a, b*]**

[a]Department of Architecture and Architectural Engineering, Yonsei University, Republic of Korea
[b]Technical University of Munich, Institute for Advanced Studies, Munich, Germany
[*]Corresponding author
E-mail: jacob89@yonsei.ac.kr, glee@yonsei.ac.kr

**Abstract**

**Vision-based indoor positioning approaches are increasingly gaining attention due to their cost-effectiveness and scalability. The core principle of vision-based indoor positioning is to identify the most visually similar space to the reference space images. A major challenge in previous vision-based indoor positioning methods arises from the need for large annotated datasets to ensure robust model performance, necessitating coverage of space images from various angles and points of view. To address this limitation, we propose a method that applies feature matching to 360-degree images, enabling indoor positioning with just a single reference image per space. This eliminates the need for multiple reference images by preprocessing the input image through feature-descriptor-based alignment and cube-map projection, allowing the image to be adjusted to better match the position and shape of the reference image. Experiments conducted on six different floor plans achieved an accuracy of 72.57% using only one reference image, confirming the feasibility and efficiency of this lightweight approach to indoor positioning.**

**Keywords –**

**360-degree photo; Image feature matching; Indoor positioning; Computer vision; Few-shot learning; One-shot learning**

## 1 Introduction

Unlike outdoor positioning often relies on global positioning system (GPS) [1], indoor environments require different approaches because GPS signals cannot be received. The most common indoor positioning methods are signal-based techniques such as Wi-Fi [2], ultra-wideband (UWB) [3], and radio frequency identification (RFID) [4], which typically rely on triangulation methods and thereby offer robust performance. However, these techniques can be heavily influenced by indoor objects, such as furniture and walls, that interfere with signal propagation. Moreover, they require the installation of dedicated equipment, which can limit their applicability. In these contexts, vision-based indoor positioning methods present a more intuitive and robust alternative than signal-based methods [5].

Indoor positioning research based on visual information often involves extracting features from images and then determining proximity to predefined reference points [6]. Recent advances in machine learning and deep learning have enabled systems to automatically identify salient features from video frames, which can subsequently be used to estimate a user's location relative to these reference points. By harnessing visual cues, such systems can potentially overcome the limitations of wireless signal-based approaches, offering finer spatial resolution and rich contextual information about the surrounding environment.

Despite the advantages offered by vision-based indoor positioning methods, several challenges remain. A major issue is the heavy dependence on large, labeled image datasets; as the number of positioning reference points increases, so does the need for more extensive data collection. Another concern is the potential confusion arising in visually similar spaces [7]. Even in the same location, images can vary significantly depending on the camera's extrinsic parameters, making it difficult to distinguish between different areas that share analogous visual features. Therefore, to maintain the benefits of image-based indoor positioning while overcoming these challenges, methods are needed that can operate effectively with fewer images and reduce reliance on specific camera parameters.

In this paper, we propose a one-shot indoor positioning method for residential buildings that leverages 360-degree images. Because many residential buildings repeat the same floor plan on every floor, they offer an environment well-suited for vision-based indoor positioning. To address the conventional reliance on large-scale datasets, we advocate the use of 360-degree images instead of standard photographs. Compared to conventional images, 360-degree images exhibit lower dependence on camera angles and positions, and they can be manipulated through projection to highlight specific visual information [8]. Building on these advantages, we

prepare a single reference image for each room as training data and use a corresponding query image to determine the user's current location by comparing their visual similarity. As part of this process, we employ feature extraction to adjust the orientation of the 360-degree images. We then perform cube-map projections on the aligned images, segmenting them so that each direction can be compared directly. Through this approach, robust indoor positioning is achievable with only one representative 360-degree image per room.

## 2 Related Studies

### 2.1 Image-based Indoor Positioning System

Image-based positioning offers the advantage of leveraging dense visual information with minimal equipment while providing intuitive insights into spatial relationships. Traditionally, such methods have relied on feature-based techniques like the Scale-Invariant Feature Transform (SIFT) or Oriented FAST and Rotated BRIEF (ORB) [9] for keypoint detection and matching. By correlating extracted features with reference images or pre-built maps, researchers have demonstrated accurate localization even in the most intricate, multi-room settings. Concurrently, recent advances in neural network architectures have led to the widespread development of similarity-based methods [7]. Although these approaches share the overarching goal of precise localization, they primarily differ in how they extract and represent features from given images.

Despite their promise, these methods often require large-scale image datasets and carefully tuned parameters to ensure robustness under varying conditions. Keypoint-descriptor-based methods (e.g., SIFT, ORB) circumvent extensive training procedures, but tend to be highly sensitive to environmental changes in lighting or perspective [10]. Consequently, additional processes, such as rigorous parameter tuning or ensemble modelling, are often recommended to improve reliability. Moreover, in tall buildings where multiple floors replicate near-identical room layouts, distinguishing among different spaces becomes increasingly problematic. This issue is exacerbated by the limited field of view (FoV) in conventional images, which can capture only a fraction of the environment and thus yield inconsistent results when camera angles differ.

To overcome these challenges, 360-degree imaging has gained momentum as an alternative. By capturing the entire environment from a single vantage point, such images minimize the need for multiple photographs taken at various angles or distances [11]. Not only does this broader coverage enhance feature extraction, but it also simplifies the alignment process during positioning. Furthermore, positioning methods that employ scene understanding of 360-degree photos through segmentation-based techniques have also been introduced [12,13]. However, deep learning-based methods, particularly those leveraging convolutional neural networks and other advanced architectures, demand vast amounts of training data to achieve robust performance. Generating labeled image datasets for every possible indoor environment can be both time-consuming and cost-intensive, which is where few-shot learning becomes highly relevant.

### 2.2 Few-shot Learning in Construction

Few-shot learning techniques enable models to generalize from a limited number of examples [14], thereby reducing the reliance on large, labeled datasets. They encompass various approaches, including meta-learning [14] and similarity-based methods [15], and provide a promising solution by requiring only minimal data per class. In the construction domain, where collecting extensive labeled datasets can be costly and labor-intensive [16], few-shot learning helps alleviate the burdens typically associated with vision-based monitoring techniques [17].

In the context of indoor positioning, few-shot learning substantially decreases the need for exhaustive image capture and labeling in each room or space. By focusing on representative images or leveraging meta-level patterns across different tasks, these approaches maintain high accuracy even in visually similar environments—a feature particularly advantageous in high-rise residential buildings with repeated floor plans. Through the integration of 360-degree imaging and few-shot learning, it becomes possible to capture comprehensive spatial information while avoiding the data-intensive bottlenecks often encountered in deep learning methods.

### 2.3 Research Gap

In summary, two major challenges arise when implementing indoor positioning systems that rely primarily on visual information and image similarity [7]. One involves calibrating the camera's extrinsic parameters [11], and the other concerns assembling the large dataset required to construct the model [8]. As a potential solution, the use of 360-degree images has been proposed in prior work. Within this approach, some studies focus on calibration, while others involve subdividing images to increase the available data. However, the former still requires a feasibility assessment for practical use, and the latter necessitates additional preprocessing, which restricts direct application under few-shot conditions.

In this study, we combine the strengths of these prior methods to propose an indoor positioning technique that

functions with only a small number of images. Through feature matching, we correct the camera's external parameters, and by employing CMP-based 360-degree images, we make full use of omnidirectional visual information. As a result, accurate positioning becomes possible even with minimal data, providing a novel approach to vision-based indoor localization.

## 3 Research Proposal

Our proposed method is divided into two primary steps (Figure 1). The first step is a preprocessing stage performed when the query image is provided. During this stage, we align the 360-degree image and, if necessary, apply inpainting to remove the photographer or other unintended objects. This process corrects for variations in camera angle and eliminates noise that could otherwise skew comparisons, thereby ensuring an equal basis for evaluating similarity.

In the second step, we compare the corrected query image with each reference image. By applying a projection technique, we segment both the query image and the reference image into multiple directions, splitting them into six sections via cube-map projection (CMP), so that images corresponding to the same direction can be compared individually [8]. This approach addresses the distortions inherent in the equirectangular projection (ERP) format of 360-degree images, which can vary based on the camera's position inside the space. We repeat these comparisons for all reference images in the indoor environment, and the final positioning result is determined by selecting the space with the highest average similarity score. Further details on each step of the process are provided in the following subsections.

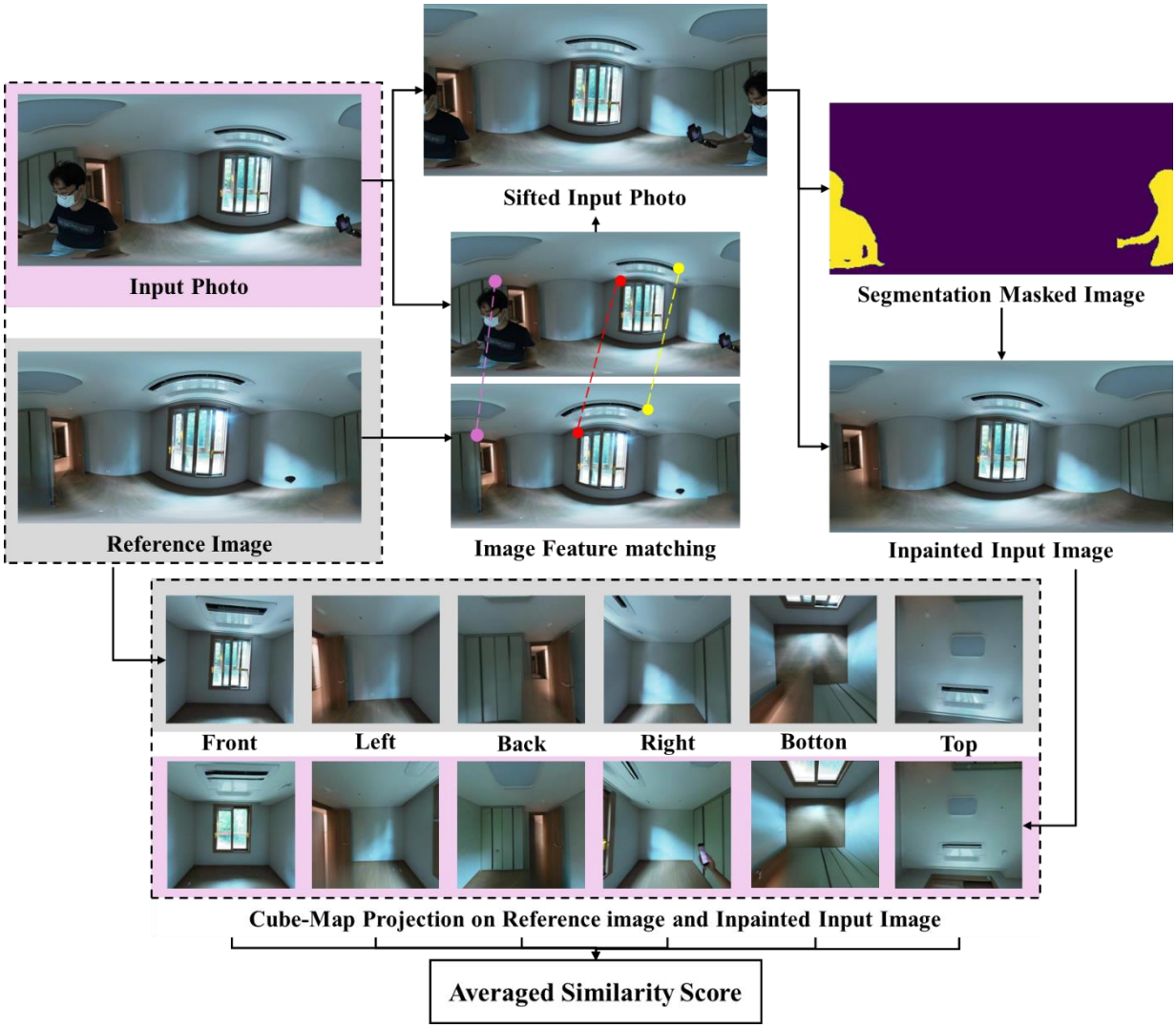


**Figure 1. Overall flow of preprocessing on input image and similarity comparison with reference image**

### 3.1 Preprocessing on Input Image

First, once the input photo is provided, serving as the query image, both the query image and the reference image undergo feature extraction using keypoint-descriptor methods. In this study, we employ an ensemble model comprising ORB [9], BRISK [18], and AKAZE [19]. [20][10]We then measure the horizontal distance between the keypoint-pairs that yield the highest match scores. This distance is used to shift the input image so that it is aligned to face the same direction as the reference image.

Next, if the image contains a human subject, it introduces noise that must be removed. We accomplish this by performing inpainting technique, which requires a segmentation mask to identify the person's location. In this process, we employ semantic segmentation to generate the mask. Specifically, we use the BEiT model [21], pre-trained on the ADE20K dataset [22], for segmentation, and the LaMa model for inpainting [23].

### 3.2 Similarity Comparison

After aligning the orientation of the input image and removing noisy objects, we compare the resulting image with a reference image to evaluate their similarity. In this study, we use a CMP approach for this comparison, aiming to compensate for the substantial distortions in ERP images that can occur even within the same space. In typical residential buildings, each room can be approximated as having a roughly cubic shape; therefore, once the reference image is properly adjusted, each CMP section captures a distinct wall in the room. This setup allows both the CMP-based reference image and the pre-processed input image to be compared on a wall-by-wall basis.

To quantify similarity between images, we employ a pretrained Siamese network that outputs a similarity score [24]. The final similarity score between one input photo and a reference image is the average of the scores obtained from the six CMP sections. Repeating this process for every reference image in each room, we identify the final location as the room whose reference image yields the highest average similarity.

## 4 Experiment & Results

### 4.1 Dataset Preparation and Setting for Experiment

Performance experiments on the proposed method were conducted in residential buildings. Specifically, we selected six housing units, and categorized their spaces into eight areas: ["Bathroom", "Bedroom", "Living Room", "Dress Room", "Balcony", "Utility Room", "Unit Room", "Kitchen"]. In each space, we captured 360-degree images. When a single unit contained multiple spaces of the same type (e.g., multiple bathrooms or bedrooms), we labeled them numerically to distinguish them as separate spaces. As a result, we compiled 288 images across 11 spatial categories for our experimental dataset (Table 1).

In addition, we prepared separate reference images for each room. To correct for potential distortions that could arise during the similarity-check process, these reference images were taken from the center of each room. Once converted via CMP, the resulting six images were oriented to provide frontal views of each wall in the space.

**Table 1. Photo distribution of each room in the test Dataset**

| **Balcony** | **Dressroom** | **Kitchen** | **Livingroom** |
|---|---|---|---|
| 18 | 20 | 30 | 30 |
| **Bedroom** | **Bathroom** | **Unit** | **Utility** |
| 34 / 31 / 35 | 24 / 21 | 24 | 21 |

### 4.2 Performance Check

In our experiments, the primary parameters requiring careful configuration were the number of features used during the feature matching step and the FoV value applied in the CMP process. In this study, we limited the number of matched features to one and fixed the FoV at 90, based on empirical findings that the performance gains from using a larger number of features were outweighed by the degradation caused by increased noise.

Furthermore, for feature matching, we employed three representative keypoint descriptors (ORB, BRISK, and AKAZE) and ensemble model. Propose ensemble model selects the best descriptor on a per-match basis by choosing the one that yields the highest match score, thereby enhancing the overall reliability of the matching process.

Based on the ensemble model, the highest performance among the tested descriptors was 72.57%, largely attributable to BRISK, which was the primary descriptor used in the ensemble (Table 2). Notably, previous studies employing similar approaches relied on AKAZE [11], suggesting that no single descriptor offers superior performance in all scenarios [10]. Instead, a composite approach like the ensemble appears to be more effective overall.

A closer examination of the data reveals varying classification performance across different rooms (Figure 2). Bedrooms 1, 2, and 3 were frequently misclassified due to their high degree of visual similarity, while spaces

with more distinctive features, such as balconies or dressing rooms, showed better performance. In contrast, although bathrooms displayed relatively favorable recall values, their precision scores dropped significantly. This was attributable to low feature-matching scores across all descriptors, which led to reduced overall similarity scores. One likely cause is the large mirror occupying one wall, which may have interfered with the feature-matching process. Future research should address this issue to further enhance system performance.

**Table 2. Average accuracy according to each descriptor**

| Descriptor | ORB | BRISK | AKAZE | Ensemble |
|---|---|---|---|---|
| **Accuracy** | 69.44% | 71.53% | 68.40% | 72.57% |

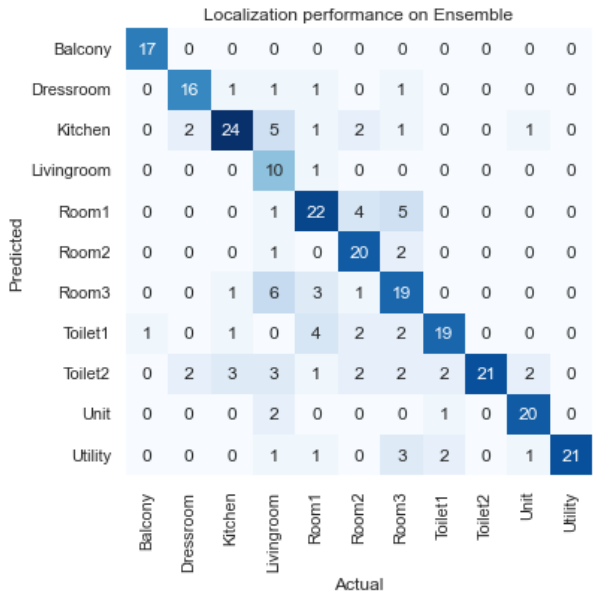


**Figure 2. Confusion matrix of the ensemble model for each room**

### 4.3 Ablation Studies

To evaluate how our proposed method, which utilizes feature matching for image alignment and CMP for image comparison, contributes to overall performance, we conducted an ablation study by selectively removing each of these components. First, (a) we measured accuracy using only the reference image and the input photo, without either feature matching or CMP, resulting in 9.13% accuracy. This is the result of the previous method based on a 360-degree image approach that relies solely on simple feature descriptors [11], excluding all the suggestions from this study. Second, (b) we performed feature matching for image alignment on the 360-degree images but used the ERP images for the comparison step, yielding 8.14% accuracy. Third, (c) we omitted the feature matching step and used only CMP, reaching 8.56% accuracy (Table 3).

**Table 3. Performance changes according to application of suggestions**

| | (a) Without proposal | (b) Only feature matching | (c) Only CMP comparison | (d) Proposed method |
|---|---|---|---|---|
| **Accuracy** | 9.13% | 8.14% | 8.56% | 72.57% |

These results are significantly lower than the 72.57% accuracy achieved by (d) our complete method. Moreover, given that the dataset contains 11 classes, the near 9–10% performance suggests results approaching random chance, thereby underscoring the limited utility of each individual component on its own. Consequently, our findings indicate that the benefits of the proposed method manifest only when feature matching and CMP are applied in tandem. Furthermore, considering that existing studies typically rely on feature matching alone [11], the considerable distortion inherent in 360-degree images can introduce major complications, supporting the necessity of our combined approach for robust indoor positioning.

## 5 Conclusion

This study presents a 360-degree image-based indoor positioning method that can be applied in small, repetitive indoor spaces using a single reference image. Unlike conventional images, 360-degree images provide substantially more visual information, which can alleviate the demands for a large dataset. However, to address the potential distortion and noise inherent in these images, we propose a two-step approach comprising of a preprocessing phase and a comparison phase. The method achieved an accuracy of 72.56% using only a single image per room. This represents a significant improvement compared to traditional feature-matching-based methods that are difficult to apply in practice [11], as confirmed by the ablation study. Furthermore, the performance is comparable to previous studies that required large amounts of training data [7], demonstrating that this study successfully achieved its goal of developing a practical model with minimal image input.

The primary contribution of this work lies in proposing a practical indoor positioning system that requires only one 360-degree image per room, alleviating the burden of preparing a large dataset in previous vision-based methods or the high cost of installing signal-based systems. Another key contribution of this study lies in the automated calibration approach by combining the use of

feature matching and an image similarity comparison technique. When image shifting via feature matching is integrated with CMP-based image similarity, we can achieve a level of performance suitable for practical use, even with only a single image per room. These benefits can also be extended to other studies requiring approximate location tracking, thus broadening the potential applications of our method.

Nevertheless, future research should aim to further enhance positioning accuracy, particularly in spaces like bathrooms that exhibit lower performance. Additionally, although the proposed method demands substantial computational resources and processing time, we anticipate that subsequent optimization and module simplification in future work will help mitigate these limitations.

## 6 Acknowledgement

This work was supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant, funded by the Ministry of Land, Infrastructure and Transport (No. RS-2024-00407028).

## References

[1] A.S. Rao, M. Radanovic, Y. Liu, S. Hu, Y. Fang, K. Khoshelham, M. Palaniswami, T. Ngo, Real-time monitoring of construction sites: Sensors, methods, and applications, Automation in Construction 136 (2022) 104099. https://doi.org/10.1016/j.autcon.2021.104099.

[2] M. Abbas, M. Elhamshary, H. Rizk, M. Torki, M. Youssef, WiDeep: WiFi-based Accurate and Robust Indoor Localization System using Deep Learning, in: 2019 IEEE International Conference on Pervasive Computing and Communications (PerCom, 2019: pp. 1–10. https://doi.org/10.1109/PERCOM.2019.8767421.

[3] P. Dabove, V. Di Pietra, M. Piras, A.A. Jabbar, S.A. Kazim, Indoor positioning using Ultra-wide band (UWB) technologies: Positioning accuracies and sensors' performances, in: 2018 IEEE/ION Position, Location and Navigation Symposium (PLANS), 2018: pp. 175–184. https://doi.org/10.1109/PLANS.2018.8373379.

[4] T. Sanpechuda, L. Kovavisaruch, A review of RFID localization: Applications and techniques, in: 2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008: pp. 769–772. https://doi.org/10.1109/ECTICON.2008.4600544.

[5] Y. Wei, B. Akinci, A vision and learning-based indoor localization and semantic mapping framework for facility operations and management, Automation in Construction 107 (2019) 102915. https://doi.org/10.1016/j.autcon.2019.102915.

[6] P. Pascacio, S. Casteleyn, J. Torres-Sospedra, E.S. Lohan, J. Nurmi, Collaborative Indoor Positioning Systems: A Systematic Review, Sensors 21 (2021) 1002. https://doi.org/10.3390/s21031002.

[7] I. Ha, H. Kim, S. Park, H. Kim, Image retrieval using BIM and features from pretrained VGG network for indoor localization, Building and Environment 140 (2018) 23–31. https://doi.org/10.1016/j.buildenv.2018.05.026.

[8] Y. Wei, B. Akinci, Panorama-to-model registration through integration of image retrieval and semantic reprojection, Automation in Construction 140 (2022) 104356. https://doi.org/10.1016/j.autcon.2022.104356.

[9] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: 2011 International Conference on Computer Vision, 2011: pp. 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544.

[10] S.A.K. Tareen, Z. Saleem, A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK, in: 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2018: pp. 1–10. https://doi.org/10.1109/ICOMET.2018.8346440.

[11] T. Yashiro, H. Hirayama, K. Sakamura, An Indoor Localization Service using 360 Degree Spherical Camera, in: 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech), 2020: pp. 17–18. https://doi.org/10.1109/LifeTech48969.2020.1570617174.

[12] H. Xu, Q. Zhao, Y. Ma, S. Wang, C. Yan, F. Dai, Free-Viewpoint Navigation of Indoor Scene with 360° Field of View, Electronics 12 (2023) 1954. https://doi.org/10.3390/electronics12081954.

[13] J. An, D.H. Lee, H.H. Cho, O.H. Jeong, Indoor Positioning System Using Smartphone and 360° Camera, in: 2021 IEEE International Conference on Smart Internet of Things (SmartIoT), 2021: pp. 342–343. https://doi.org/10.1109/SmartIoT52359.2021.00062.

[14] S. Thrun, L. Pratt, Learning to Learn, Springer Science & Business Media, 2012.

[15] Z. Cui, Q. Wang, J. Guo, N. Lu, Few-shot classification of façade defects based on extensible classifier and contrastive learning, Automation in Construction 141 (2022) 104381. https://doi.org/10.1016/j.autcon.2022.104381.

[16] D. Gil, G. Lee, Zero-shot monitoring of construction workers' personal protective equipment based on

image captioning, Automation in Construction 164 (2024) 105470. https://doi.org/10.1016/j.autcon.2024.105470.
[17] J. Kim, S. Chi, A few-shot learning approach for database-free vision-based monitoring on construction sites, Automation in Construction 124 (2021) 103566. https://doi.org/10.1016/j.autcon.2021.103566.
[18] S. Leutenegger, M. Chli, R.Y. Siegwart, BRISK: Binary Robust invariant scalable keypoints, in: 2011 International Conference on Computer Vision, 2011: pp. 2548–2555. https://doi.org/10.1109/ICCV.2011.6126542.
[19] P. Alcantarilla, J. Nuevo, A. Bartoli, Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces, in: Procedings of the British Machine Vision Conference 2013, British Machine Vision Association, Bristol, 2013: p. 13.1-13.11. https://doi.org/10.5244/C.27.13.
[20] D.G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision 60 (2004) 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.
[21] H. Bao, L. Dong, S. Piao, F. Wei, BEiT: BERT Pre-Training of Image Transformers, (2022). https://doi.org/10.48550/arXiv.2106.08254.
[22] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, A. Torralba, Semantic Understanding of Scenes through the ADE20K Dataset, (2018). https://doi.org/10.48550/arXiv.1608.05442.
[23] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, V. Lempitsky, Resolution-robust Large Mask Inpainting with Fourier Convolutions, (2021). https://doi.org/10.48550/arXiv.2109.07161.
[24] G. Koch, R. Zemel, R. Salakhutdinov, Siamese neural networks for one-shot image recognition, in: ICML Deep Learning Workshop, 2015. https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf.