Deep Learning Approach for Classifying the Built Year and Structure of Individual Buildings by Automatically Linking Street View Images and GIS Building Data

Yoshiki Ogawa, Chenbo Zhao, Takuya Oki, Shenglong Chen, *Graduate Student Member, IEEE*, and Yoshihide Sekimoto

***Abstract*—The built year and structure of individual buildings are crucial factors for estimating and assessing potential earthquake and tsunami damage. Recent advances in sensing and analysis technologies allow the acquisition of high-resolution street view images (SVIs) that present new possibilities for research and development. In this study, we developed a model to estimate the built year and structure of a building using omnidirectional SVIs captured using an onboard camera. We used geographic information system (GIS) building data and SVIs to generate an annotated built-year and structure dataset by developing a method to automatically combine the GIS data with images of individual buildings cropped through object detection. Furthermore, we trained a deep learning model to classify the built year and structure of buildings using the annotated image dataset based on a deep convolutional neural network (DCNN) and a vision transformer (ViT). The results showed that SVI accurately predicts the built year and structure of individual buildings using ViT (overall accuracies for structure = 0.94 [three classes] and 0.96 [two classes] and for age = 0.68 [six classes] and 0.90 [three classes]). Compared with DCNN-based networks, the proposed Swin transformer based on ViT architectures effectively improves prediction accuracy. The results indicate that multiple high-resolution images can be obtained for individual buildings using SVI, and the proposed method is an effective approach for classifying structures and determining building age. The automatic, accurate, and large-scale mapping of the built year and structure of individual buildings can help develop specific disaster prevention measures.**

***Index Terms*—Building identification, deep learning, object detection, street view images (SVIs), Swin transformer.**

I. Introduction

BUILDING-STRUCTURES' built years and materials (reinforced concrete (RC), steel, and wood) are major factors for predicting the level of damage in buildings that have collapsed during disasters, such as earthquakes and tsunamis [1], [2]. These factors are utilized as parameters in building damage fragility curve models for damage assessment based on the field survey results of earthquake disasters [3], [4], [5], [6]. In Japan, after the Building Standard Laws were revised in 1981 and June 2000, the damage levels were found to be significantly different for different buildings [7]. Areas with several old earthquake-resistant buildings in wood-dense areas experienced fires that spread because of building collapses, which caused considerable damage [8]. Therefore, it is important to estimate the built year and structure of individual buildings.

However, most municipalities in Japan cannot use building data with detailed attribute information, such as building structure and building year, because these details are often absent in most public databases. Furthermore, most developing countries do not have sufficient data for such estimations, and hence, detailed damage estimation and disaster prevention planning cannot be conducted; this can endanger human lives, given the possibility of building collapse during large earthquakes, such as the predicted Nankai Trough Earthquake and Tsunami. Especially for developing countries, information on a building's built year and structure can help classify vulnerable properties for earthquake damage simulation and postdisaster recovery [9], [10], [11]. Thus, it is pertinent to develop a method for acquiring the geographic information system (GIS) data of building structures and the built years for each building level, which is applicable to many countries, including developing countries [12].

Although few studies have focused on classifying the built year, several methods have been proposed to classify the number of floors and structures based on field surveys [13], [14], aerial photographs [15], [16], real estate images [17], [18], [19], and methods that use attributes (number of floors and area) of statistical data and GIS building data [20], [21], [22], [23]. However, real estate images bias data collection by essentially covering only trading estate buildings and fewer old buildings. Biljecki and Sindram [23] predicted the built year from a three-dimensional (3-D) GIS dataset comprising attributes, such as building height and area. Rosser et al. [24] proposed to extract

Manuscript received 12 July 2022; revised 23 November 2022 and 23 December 2022; accepted 13 January 2023. Date of publication 16 January 2023; date of current version 8 February 2023. This work was supported by JSPS KAKENHI under Grant JP 20K15001. *(Corresponding author: Yoshiki Ogawa.)*

Yoshiki Ogawa and Yoshihide Sekimoto are with the Center for Spatial Information Science, University of Tokyo, Tokyo 153-8505, Japan (e-mail: ogawa@csis.u-tokyo.ac.jp; sekimoto@iis.u-tokyo.ac.jp).

Chenbo Zhao and Shenglong Chen are with the Department of Civil Engineering, University of Tokyo, Tokyo 153-8505, Japan (e-mail: cbzhao@iis.u-tokyo.ac.jp; chen-sl@iis.u-tokyo.ac.jp).

Takuya Oki is with the School of Environment and Society, Tokyo Institute of Technology, Tokyo 152-8550, Japan (e-mail: oki.t.ab@m.titech.ac.jp).

Digital Object Identifier 10.1109/JSTARS.2023.3237509

measures of the morphology and neighborhood characteristics from topographic mapping, a high-resolution digital surface model (DSM), and statistical boundary data, using them as features within a random forest classifier to infer an age category for each building. However, the 3-D GIS dataset, the real estate dataset, and DSM used in previous studies are not available for many cities, and the building covering rate is low.

Conversely, in recent years, various types of image big data related to architecture and cities have been used for analysis. Examples include exterior images of real estate properties and street view images (SVIs), such as google street view (GSV) and mapillary, which are automatically recorded using an omnidirectional camera mounted on the roof of a vehicle while driving. Therefore, SVI images have the advantage of efficiently collecting huge quantities of comprehensive street image data that can be used for analysis. Pelizari et al. [25] and Wang et al. [26] modeled building information for natural hazard risk management. Pelizari et al. [25] used GSV to classify building materials and the numbers of floors of buildings using a deep convolutional neural network (DCNN) and demonstrated the effectiveness of SVI. Iannelli et al. [27] proposed a CNN-based framework to extract building floor numbers from GSVs in San Francisco, USA. Ghione et al. [28] classified buildings using GSV by applying a CNN model. They defined a corresponding set of typologies—wooden, unreinforced masonry, S, RC, and steel–RC—for the Norwegian building typologies observed in Oslo, which are applicable to the development of a cost-effective building stock model. They achieved an accuracy of 89% for wooden buildings, but the success score for RC buildings was only 35%. These methods manually generate training data, and the method only considers single images for each building even when using a multi omnidirectional image.

In recent years, Sun et al. [29], [30] proposed a built-year estimation model using SVIs and presented an automated workflow for estimating building age from SVIs because of the increasing coverage of SVIs, such as GSV. The proposed model was constructed considering a 30-year time slice without considering the earthquake resilience of buildings; the low accuracy in dense areas within urban areas is due to the image cropping method. Furthermore, their automated labeled method could not label contiguous building areas and complex building shapes because only the façade midpoint was considered. Finally, another automated method demonstrated that a significant number of residential properties do not appear in SVI because of occlusion and misregistration [31], [32]. Thus, we can infer that it is difficult to record building structures and built years efficiently on a large scale, especially in urban areas. In addition, the image cropping method for SVIs has not been evaluated in previous studies even though it affects the accuracy of the training dataset. The method proposed in this article is the first to predict and map with high accuracy the built year for buildings spread over a wide area in urban areas.

Previous studies focused only on developing built-year and structure models based on precropped images and did not apply them to city-scale mapping. Therefore, the scalability and application mechanism for a classification model to designate the built year and structure in GIS-based maps remains a challenge, considering that an earthquake/tsunami damage-risk map cannot be created without built-year and building-structure GIS data. Therefore, the automated building cropping method that uses SVIs is not considered. Thus, we need to develop an automated method for joining SVI and GIS building-footprint data based on spatial relations and crop building images from SVIs.

Extant studies also do not accurately report on their applicability to urban areas with dense buildings over wide areas. Furthermore, previous studies that employed deep learning methods did not use recent DCNN network architectures or deeper network architectures with either more powerful hierarchies or vision transformers (ViT). Therefore, recent DCNN and ViT architectures must be applied for deep learning.

To this end, we propose a new method for large-scale mapping and classifying built years and structures using state-of-the-art deep learning algorithms by combining GIS building data and building images spatially extracted and automatically cropped from SVIs to create building images annotated with the built year and structure for training the model. The cost of applying the method over a wide area can be reduced by combining SVIs using an algorithmic method that automatically connects with GIS building data, which renders the proposed method sustainable. In addition, the results of structure and built-year estimation from the SVI are stored in the GIS data by combining the SVI and GIS building data. Given that recent SVIs have higher resolutions and can be obtained from various angles, detailed building façade information can be collected. We believe that our proposed large-scale approach can contribute to advancements in disaster prevention by upgrading the risk assessment of earthquake and tsunami damage in urban areas worldwide.

The main contributions of this work are listed as follows.

1) We propose a framework that automatically combines GIS building-footprint data with SVI to extract and crop buildings from a vast number of images over a wide area. The building images are then automatically annotated with GIS building data.
2) We developed a novel end-to-end framework using recent deep learning architectures, GIS data, and SVIs to estimate the built years and structures of buildings in a large area automatically and efficiently.
3) We show that high-resolution SVIs are effective in classifying built year and structure by acquiring detailed façade information of buildings from several images.

The rest of this article is organized as follows. Section II provides an overview of the related studies. Section III describes the dataset used in this study, and Section IV describes the experiments conducted and presents the results; Section V discusses the acquired results. Finally, Section VI concludes this article.

II. Data

A. SVI Data

In Japan, the secondary use of GSV is prohibited because of Japanese law. Therefore, instead of GSV, this study used 195 105 SVIs of Kobe City obtained from the Zenrin Corporation (2013), which owns a comprehensive collection of SVIs from all over Japan. The images were captured at 2.5 m intervals using a 360° roof-mounted camera as a vehicle drove along the streets. Each image was annotated and geotagged with textual information,

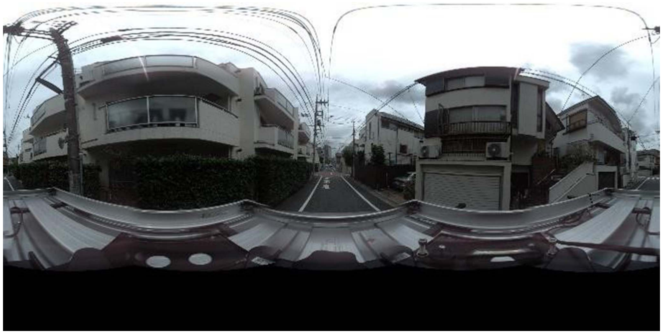


Fig. 1. Example of SVI in this study.

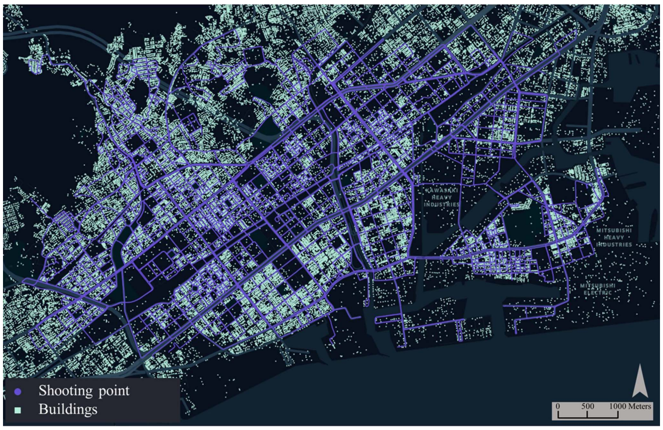


Fig. 2. SVI and GIS building data distribution in this study (Kobe city, Hyogo Prefecture, Japan). The purple and light blue colors represent the distribution of locations where these images were captured and the GIS building data distribution, respectively.

such as the latitude and longitude measured using the GPS, the vehicle's azimuth, and the time of recording. The original panoramic images are in the .jpeg tar format with dimensions of 2700 (height) × 5400 (width) pixels; the resolution is higher than that of GSV. The bottom of the image included the vehicle's roof. Fig. 1 shows an example of an SVI. The image shows that the details of the building facades are important for estimating the built year and structure of the building. Fig. 2 shows a map indicating the image locations (purple) and the distribution of the locations where these images were captured. The data are distributed along roads accessible to vehicles.

B. GIS Building Data

In Japan, there are several GIS building-footprint datasets covering areas all over the country that include building maps from municipalities, open street maps, and residential maps from private companies. In this study, we required the built-year and structure data for modeling, and therefore, the results of the Kobe City Basic Urban Planning Survey conducted in 2015 provided by G-Space Information Center [33] were used as the GIS building-footprint data. The results were provided in the form of polygon data for each building footprint. Attributes, such as building structure, built year, function, and number of floors, were provided. Fig. 2 shows the building distributions (in light blue).

TABLE I
BUILT-YEAR AND STRUCTURE CLASSIFICATION AND PROFILE OF DATASET FOR TRAINING

Experiment	Train	Validation	Test	Total
–1962	2778	1987	1987	4756
1963–1971	2969	1650	1650	4619
1972–1980	3120	1409	1409	4529
1981–1989	2961	1465	1465	4426
1990–2001	3174	2018	2018	5192
2002–	1883	2728	2728	4611
RC	5711	3310	3310	9021
Steel	4895	2905	2905	7800
Wood	6279	5042	5042	11321

III. Overall Workflow

We propose an automatic method for annotating the built years and structures of buildings using SVIs and GIS building data. We developed a deep learning model to classify the structure and building age using a well-developed building image dataset (see Fig. 3).

The training dataset was created by applying an object detection model to extract and crop buildings from SVIs; subsequently, we developed a method to match and crop the two using GIS building data and shooting location information. Furthermore, we used the latest deep learning classification models to classify individual buildings into six classes in terms of decades for the built years and three classes using the training dataset, as summarized in Table I. The built year was classified based on the decade in this study.

According to the Housing and Land Survey, only 9% of the buildings in Japan have been built before 1970. Japan being prone to natural disasters and high humidity, the life expectancy of a house in is ∼22–30 years; that in Europe and the United States is considerably different, ranging from 70 to 140 years. Thus, very few buildings in Japan are older than those in Europe and the United States. Buildings are classified into three classes (pre-1981 and post-1981, and pre-2001 and post-2001) because the building standard law was revised in 1981 and in 2000 in Japan, and because there is a considerable variation between the earthquake resistance of different buildings. We also classified into six classes based on the decade because earthquake resistance is different due to deterioration over time [6]. Therefore, the built year is validated for six classes and three classes. Furthermore, we conducted two class validations related to building structures for the three classes because earthquake resistance differs significantly between wooden and nonwooden buildings.

A. Development of Training Data

Fig. 3 shows the steps involved in the proposed workflow for the automated combination of SVIs with GIS building-footprint data.

1) Buildings were individually extracted and cropped from the panoramic SVIs using a deep learning based object detection approach [mask region-convolution neural network (mask R-CNN)]. The positions and angles of view

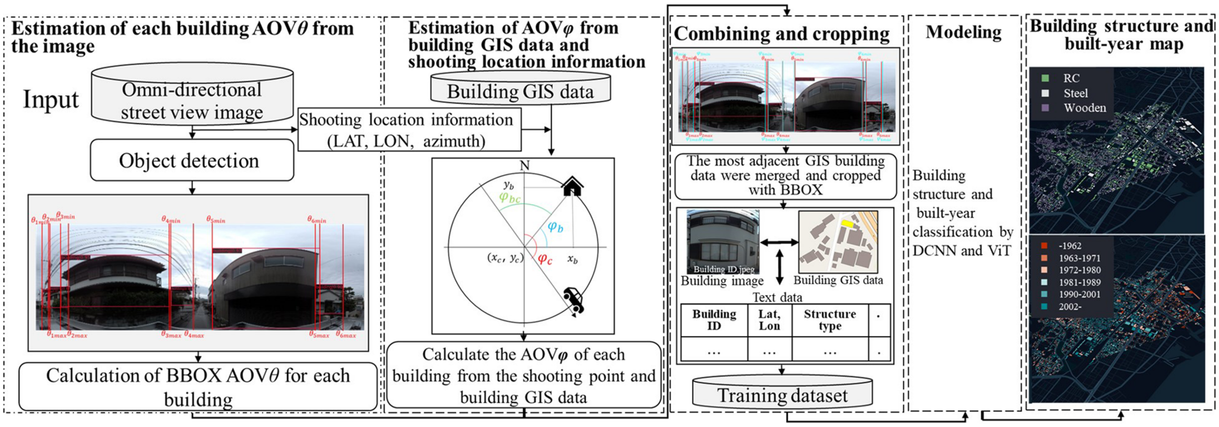

Estimation of each building AOVθ from the image
Input
Omni-directional street view image
Object detection
Calculation of BBOX AOVθ for each building
Estimation of AOVφ from building GIS data and shooting location information
Building GIS data
Shooting location information (LAT, LON, azimuth)
N
Calculate the AOVφ of each building from the shooting point and building GIS data
Combining and cropping
The most adjacent GIS building data were merged and cropped with BBOX
Building image
Building GIS data
Text data
Building ID
Lat, Lon
Structure type
Training dataset
Modeling
Building structure and built-year classification by DCNN and ViT
Building structure and built-year map
RC
Steel
Wooden
-1962
1963-1971
1972-1980
1981-1989
1990-2001
2002-

Fig. 3. Proposed workflow for building-structure and built-year classification at the individual building level.

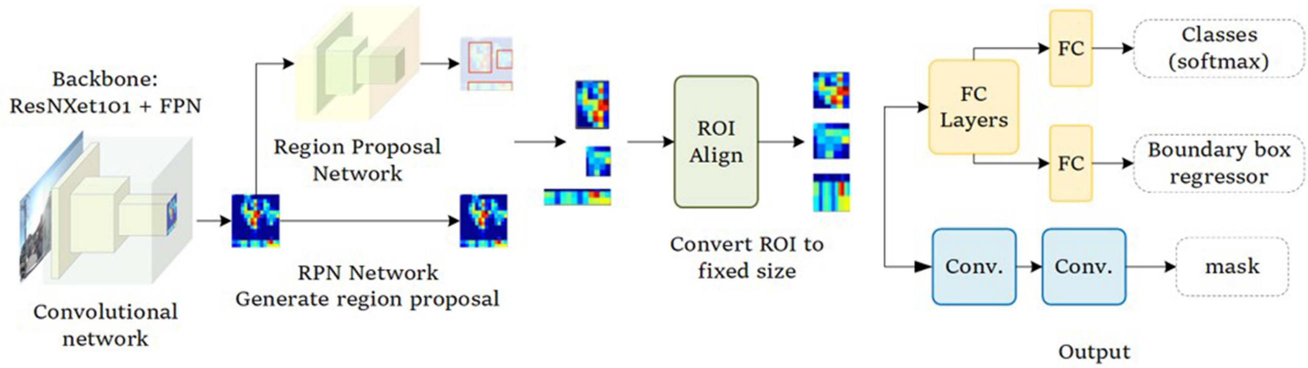


Fig. 4. Architecture of the mask R-CNN.

(AOVs) of both ends of the individual buildings in the SVIs were calculated (AOV1).

2) Based on GIS building data, the theoretical AOVs of the individual buildings were estimated from the shooting point using GIS building data. The shooting point information (latitude, longitude, and azimuth) of each SVI was observed by the in-vehicle GPS (AOV2).

3) Finally, the two angles (AOV1 and AOV2) were compared, and the most adjacent GIS building data were merged and cropped using the bounding box (BBOX) of the target building. Two angles were used to reduce the matching error between the SVI and building-footprint data and to crop a higher quality image that captures all buildings more correctly. If only AOV2 is used, some buildings in the image will be missed or multiple buildings will be extracted because of footprint and GPS errors. If both AOV2 and AOV1 are used, the building footprint can be linked to the image and missing buildings can be eliminated when extracting only target buildings in the image.

The GIS data of the target building were simultaneously assigned to ensure that each image was for an individual building ID from the GIS data when combined. Subsequently, we prepared the image dataset annotated with the building structure and built year.

1) Estimation of an Individual Building's AOV Based on the Panoramic SVI (AOV1): Multiple buildings were identified in the panoramic SVI, and the BBOXes of individual buildings were extracted by object detection using a mask region-convolution neural network (mask R-CNN) [34].

We constructed a new model suitable for panoramic SVIs based on a mask R-CNN network to determine the BBOX more accurately because the performance of the two-stage object detection framework (e.g., faster R-CNN [35] and mask R-CNN) was superior to that of a one-stage network (e.g., YOLO [36] or RetinaNet [37]). In addition, the instance segmentation branch promotes the object detection branch [37].

Fig. 4 shows the architecture of the mask R-CNN. First, input images are sent to the backbone for feature extraction, and the backbone feature map is passed through the region proposal network to extract the possible target regions of interest (ROIs). Compared with faster R-CNN, the backbone of the mask R-CNN has superior multiscale feature extraction capability. We applied a ResNeXt network with a depth of 101 layers combined with a feature pyramid network (FPN) as the backbone; this is referred to as ResNXet-101-FPN. These RoIs are mapped into fixed-dimensional feature vectors using an RoI alignment layer, which improves the detection accuracy of the boundary boxes. First, the input RoI is obtained as a 7 × 7 × 1024 RoI feature, and then it is upscaled to 1024 channels. Two branches are employed for classifying and regressing the target boundary box through a fully connected layer. Furthermore, the masked branch is upsampled by multiple convolution operations for obtaining a segmented region image. First, the ROI is changed to 14 × 14 × 256 features; then, the same operation is performed five times followed by the deconvolution operation. Finally, a 28 × 28 × 80 mask is output.

We used one thousand randomly selected SVIs with buildings that were annotated manually as training data. The entire training set consisted of only one building class. The panoramic image was used directly for training purposes; however, the image had a more intense perspective because of its integrity and because the distortion of the building was within acceptable limits.

A model [38] pretrained on the COCO dataset was used for transfer learning considering that the dataset was not sufficiently large to train a mask R-CNN model end-to-end from the start. The algorithmic network was implemented on an open-source platform for the state-of-the-art detection and segmentation algorithms provided by Meta; the pretrained model was provided by Detectron2 [39]. The BBOX for the final prediction of each building was obtained using the object detection branch of the mask R-CNN. The AOV1 of the building (left end: θ_min, right end: θ_max) was estimated from the pixel values at both ends of each building boundary box. In addition, the building image from the SVI was cropped based on the BBOX.

2) Estimating the Building AOV Based on the GIS Building Data and Shooting Location (AOV2): We estimated the corresponding theoretical AOV2 for each building from the GIS building-footprint data and shooting location data acquired from the GPS. The vertices were extracted from building polygons to preprocess the GIS building data.

The latitude and longitude of the shooting point were $(x_c,\ y_c)$, and the forward azimuth (north is 0° and south is 180°) was φ_c. The latitude and longitude of the building *b* vertex *v (v* = 1, 2, …, *n)* obtained from the GIS building-footprint data was $(x_b^v,\ y_b^v)$. The theoretical AOV2 of the building $((x_b^v,\ y_b^v) \ni (x_r,\ y_r))$ located within the radius *r* of the shooting point was calculated using these values as follows.

The radius was set to 30 m because of computational cost restrictions. The distance between the shooting point and each building's centroid from the building vertex (D_b^v m) was calculated using

$$\begin{aligned} &D_b^v \\ &= E * \cos^{-1}\left(\sin y_b^v\ \sin y_c + \cos y_b^v \cos y_c \cos\left(x_b^v\ -\ x_c\right)\right) \end{aligned} \tag{1}$$

where *E* represents the equatorial radius (6.73×10^6 m). If the distance is within the radius, the building is considered within the search target. The vertex angle of each building with respect to the shooting point is calculated as follows:

$$\varphi_b^v = \tan^{-1}\left(x_b^v - x_c,\ y_b^v - y_c\right). \tag{2}$$

TABLE II
COMPUTE CONSUMPTION OF DIFFERENT NETWORKS

Network	Parameters (M)	Flops (G)
ShuffleNet V2	2.28	0.149
MobileNet V2	3.5	0.319
VGG19	143.67	19.67
ResNet101	44.55	7.85
ResNeSt101	48.28	10.27
T2T-ViT T-19	10.27	7.8
Swin transformer small	49.61	8.52
Swin transformer base	87.77	15.14

The AOV of each building from the shooting point (car) φ_{bc} (left is 0° and right is 360° in the image) based on the GIS footprint data is calculated according to the forward azimuth of car φ_c as follows (see Fig. 5):

$$\begin{aligned} &\text{i) } x_b^v > x_c,\ y_b^v > y_c : \\ &\quad \varphi_{bc}^v = 270^\circ - \varphi_c - \varphi_b^v\ (0 \le \varphi_c \le 180) \end{aligned} \tag{3}$$

$$\varphi_{bc}^v = 630^\circ - \varphi_c - \varphi_b^v\ \left(180^\circ < \varphi_c \le 360^\circ\right) \tag{4}$$

$$\begin{aligned} &\text{ii) } x_b^v > x_c,\ y_b^v < y_c : \\ &\quad \varphi_{bc}^v = 270^\circ - \varphi_c + \varphi_b^v \end{aligned} \tag{5}$$

$$\begin{aligned} &\text{iii) } x_b^v < x_c,\ y_b^v < y_c : \\ &\quad \varphi_{bc}^v = 450^\circ - \varphi_c - \varphi_b^v \end{aligned} \tag{6}$$

$$\begin{aligned} &\text{iv) } x_b^v \left\langle x_c,\ y_b^v \right\rangle y_c; \\ &\quad \varphi_{bc}^v = 90^\circ - \varphi_c + \varphi_b^v\ \left(0 \le \varphi_c \le 180^\circ\right) \end{aligned} \tag{7}$$

$$\varphi_{bc}^v = 450^\circ - \varphi_c + \varphi_b^v\ \left(180^\circ < \varphi_c \le 360^\circ\right). \tag{8}$$

3) Combining Street View With GIS Building Data From AOV: Adjacent building IDs were joined to crop the building image based on the BBOX with respect to the AOV1 of the BBOX on the SVI $(\theta_{\min}, \theta_{\max})$ and the AOV2 of the building φ_{bc}^v calculated from the GIS building-footprint data and trajectory history, given as follows:

$$\begin{aligned} &\text{Target building ID} \\ &= \operatorname{argmin}\ \left\{\left(\varphi_{bc\text{max}}^v - \theta_{b\ \text{Max}} + \varphi_{bc\text{min}}^v - \theta_{b\ \text{min}}\right),\ D_b^v\ \right\}. \end{aligned} \tag{9}$$

B. Image Classification With Deep Learning Algorithms

Deep learning significantly outperforms traditional algorithms in the field of computer vision (CV) [40]. The ImageNet large-scale visual recognition challenge held since 2010, which contains a dataset of 1.4 *M* natural images labeled across 1000 classes, is a benchmark competition in the field of image classification. In this competition, the application of a DCNN significantly improved the score of ILSVRC from AlexNet [40]. In addition to the excellent performance of the DCNN, the ViT, which is another type of DL algorithm [41], [42], achieved better results in several CV tasks. Table II summarizes the DCNNs and ViT implemented in this study.

1) Implemented Models: Since 2020, ViTs have been incorporated in various state-of-the-art methods in classification, segmentation, and detection tasks [42], [43], [44]. The position of specific items and the information interaction among pixel patches should be considered the key factors for the

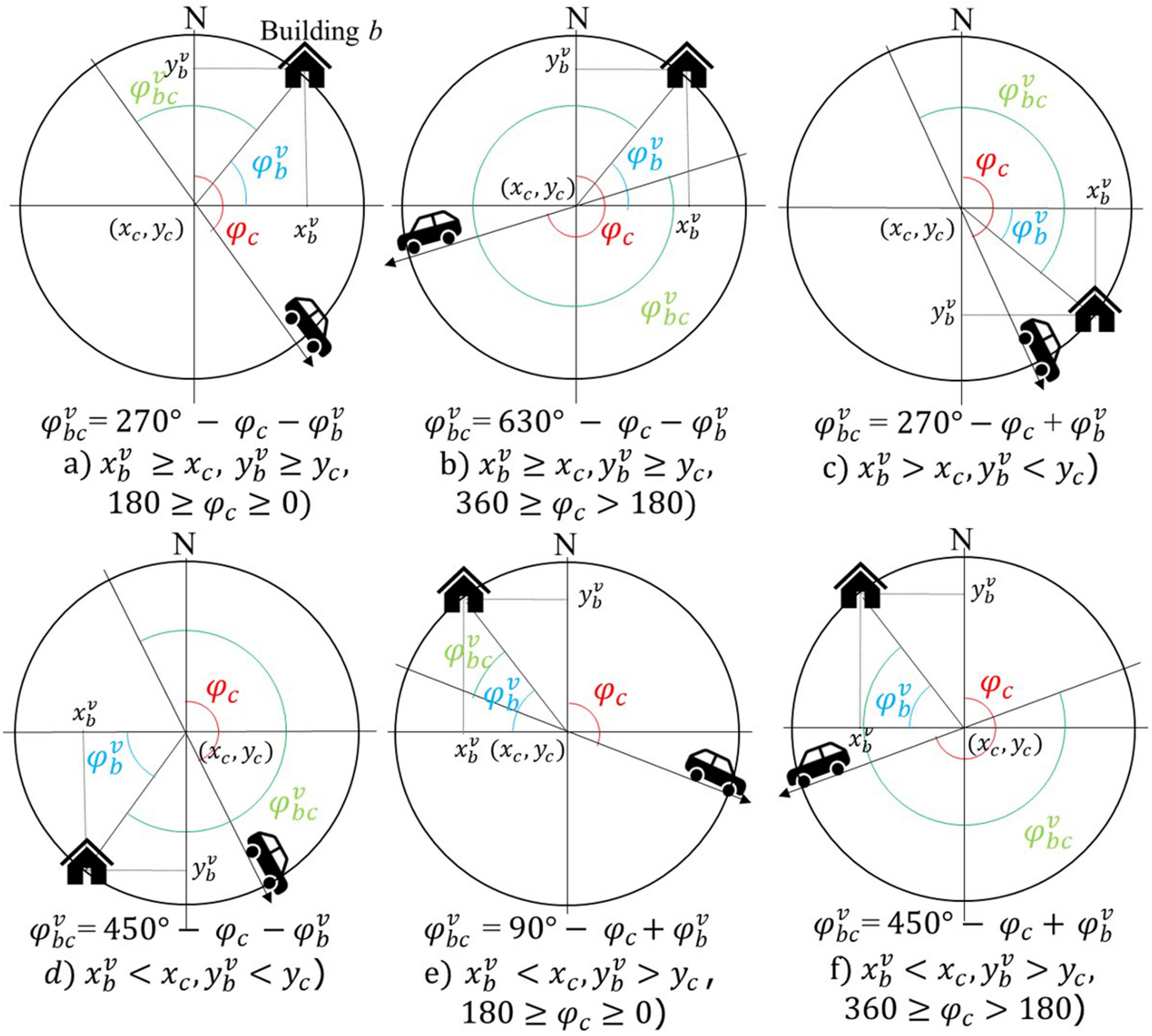

N
Building b
y_b^v
φ_{bc}^v
φ_b^v
φ_c
(x_c, y_c)
x_b^v
$\varphi_{bc}^v = 270° - \varphi_c - \varphi_b^v$
a) $x_b^v \geq x_c, y_b^v \geq y_c, 180 \geq \varphi_c \geq 0)$
$\varphi_{bc}^v = 630° - \varphi_c - \varphi_b^v$
b) $x_b^v \geq x_c, y_b^v \geq y_c, 360 \geq \varphi_c > 180)$
$\varphi_{bc}^v = 270° - \varphi_c + \varphi_b^v$
c) $x_b^v > x_c, y_b^v < y_c)$
$\varphi_{bc}^v = 450° - \varphi_c - \varphi_b^v$
d) $x_b^v < x_c, y_b^v < y_c)$
$\varphi_{bc}^v = 90° - \varphi_c + \varphi_b^v$
e) $x_b^v < x_c, y_b^v > y_c, 180 \geq \varphi_c \geq 0)$
$\varphi_{bc}^v = 450° - \varphi_c + \varphi_b^v$
f) $x_b^v < x_c, y_b^v > y_c, 360 \geq \varphi_c > 180)$

Fig. 5. Estimation of the building AOV φ_{bc}^v based on the GIS.

acquired classification results because positional encoding and self-attention are the key factors for ViTs excellent performance. In our built-year and structure classification tasks, the relationship among the key items of the building (e.g., overhang, nonopening wall, etc.) is critical for classifying according to the building standards law in Japan (see Fig. 16). This relationship aids in positional encoding and the ViTs self-attention, being suitable for the task in this study, is considerably better than that of the DCNN. Therefore, we implement the Swin transformer (small and base) [41], which is a representative SOTA model of ViT. We deployed experiments on a small version of the Swin transformer (Swin-transformer small), wherein the output channel number of the different stages was $C = 96$ when the Swin-transformer base was $C = 128$. Besides, we compared one ViT algorithm (T2T-ViT) with five DCNN algorithms (ShuffleNet V2 [45], MobileNet V2 [46], VGG19 [47], ResNet101 [48], and ResNeSt101 [49]) to identify the models suitable for our panorama image dataset and for a wide-area inference demand.

2) Transfer Learning: The volume of the dataset (see Table I) used to develop our deep learning model by training from scratch was limited, and it was far less than the ImageNet dataset (1.4 M). In addition, models pretrained on ImageNet fit the natural image source domain D_s; however, we needed a panorama image target domain D_t. Transfer learning was appropriate for our tasks considering that $D_s \neq D_t$.

In our study, all deep learning models were trained based on a model pretrained on the ImageNet dataset [50], which improved the accuracy of all deep learning algorithms in this study.

3) Assemble Modeling From Several Images: Using SVIs captured every 2.5 m, multiple images from different shooting positions were obtained for each building. Therefore, to improve the robustness of the classification, the prediction result was calculated based on the number of images. To this end, we used the maximum probability averaging method to obtain the class solution by referring to the method of Kang et al. (2018) [31] for improving the robustness of classification. Assuming that

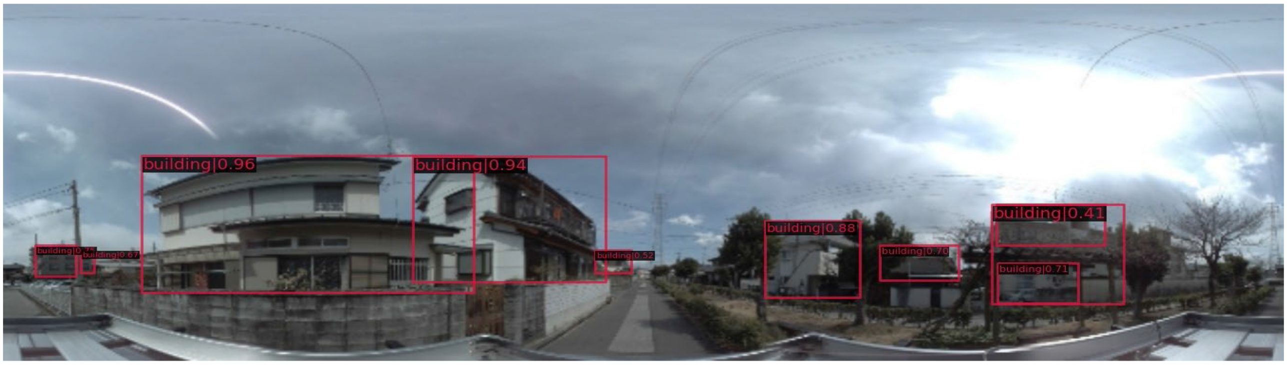


Fig. 6. Example of individual building extraction results of object detection.

$N\,(i = 1, 2, \ldots, N)$ images are acquired from the building object, i.e., from the class probability $p_j^{(i)}$, which is each class j of image i output from SoftMax, the final building class H is determined as follows:

$$H \;=\; \arg\,\max_{j} \left\{ \frac{1}{N} \sum_{i\,=\,1}^{N} p_i^{(j)} \right\}. \tag{10}$$

Hereafter, the results of the building and image units are referred to as building and image bases, respectively.

IV. Experiments and Results

A. Labeling Street-Level Imagery

1) Building Object Detection: The system used for the study included eight NVIDIA A100 GPUs with 40 GB of RAM and the Ubuntu 20.04 64-bit OS. The maximum number of iterations was set to 8000 based on the outcome of relevant experiments and hardware conditions. The batch size was set to 4, and the initial learning rate was set to 0.0001. The Adam optimizer was used for network optimization. All parameters were initialized according to the orthogonal distribution. Fig. 6 shows an example in which object detection was applied to SVIs, and the building class was determined individually; however, only buildings were individually identified. We randomly selected 1000 images for the object detection task and randomly selected 300 images (30%) for validation set.

The pixelwise accuracy was evaluated using a validation set, and a mean average precision of 35.7 was obtained; this value corresponded to the average precision of the intersection over union (IoU) range of 0.5–0.95 with a step size of 0.05. This is a key metric used in the COCO dataset. The IoU is a measure of the extent to which two regions overlap. Furthermore, the accuracy of the BBOX was verified. Precision, recall, and F1 score are the main metrics for evaluating object detection (11), (12), (13), which define these three metrics, where TP, FP, and FN refer to true positive, false positive, and false negative, respectively. Precision indicates the percentage of all data predicted to be TP, whereas recall indicates the amount of data that is predicted to be TP

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\text{F1} = \frac{2\ \times \text{Precision}\ \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{13}$$

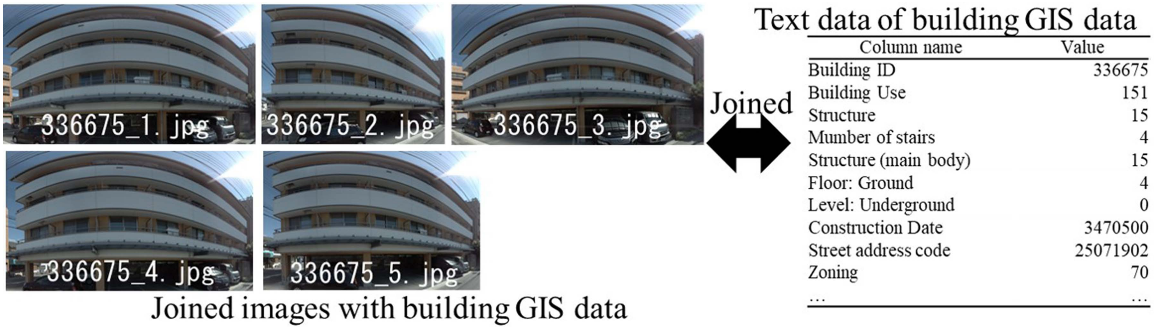
The BBOX prediction produced an excellent precision of 0.941, recall of 0.901, and an F1 of 0.914, whereas the IoU reached 0.85. This suggests that the proposed model can segment buildings accurately and precisely.

2) Result of Labeling SVI: The proposed method obtained images with building IDs annotated with building attributes, such as the built-year and structure type, because the building images extracted via object detection were combined with the GIS building-footprint data based on the building IDs (see Fig. 7). Thus, data obtained in this study can be used to identify the target building in the GIS based on attribute data (text data) by referring to the building ID in the file name for each building image. In addition, the mapping and analysis with GIS data attribute are possible by referring to building IDs.

Fig. 8 shows an example of a map of the combined relationship between the SVI and GIS building-footprint data, represented by lines. Kepler.gl [51] was used to draw the maps. Each building along the street is associated with shooting point data. For this method, SVIs were captured at 2.5 m increments, which allows multiple images to be acquired for an individual building.

3) Accuracy of Labeling Street View With the GIS Building Data: We evaluated the accuracy and building coverage of the proposed method that combines SVIs and GIS building data. There were 22 392 buildings in the study area of which 17 698 were located along roads where the vehicle with the mounted camera passed. A total of 16 183 buildings were combined using this method, which represents 72.3% of all buildings in the target area.

An objective of this study was to cover 91.4% of roadside buildings. The number of building images that could be joined was 169 086, and this confirmed that approximately 10.4 images per building could be acquired. We confirmed that over 37% of the buildings could provide ten or more images by totaling the number of images per building (see Fig. 9). Next, we evaluated the accuracy of the joined data through visual inspections to check whether the correct building images were combined and



Column name	Value
Building ID	336675
Building Use	151
Structure	15
Mumber of stairs	4
Structure (main body)	15
Floor: Ground	4
Level: Underground	0
Construction Date	3470500
Street address code	25071902
Zoning	70
…	…

Fig. 7. Example of combining SVIs with GIS building data.

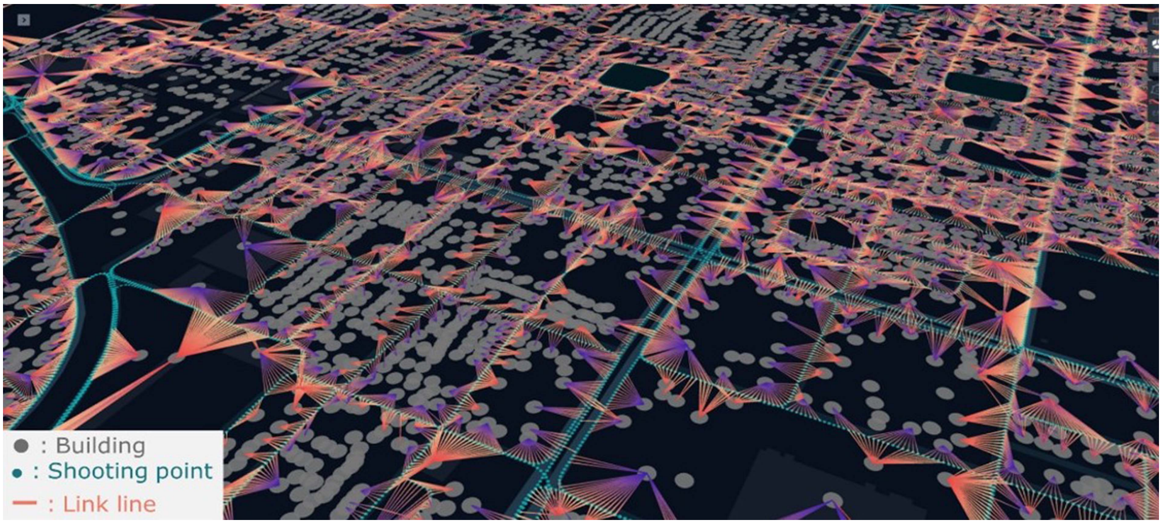


Fig. 8. Map combining the relationship between SVIs and the GIS building-footprint data.

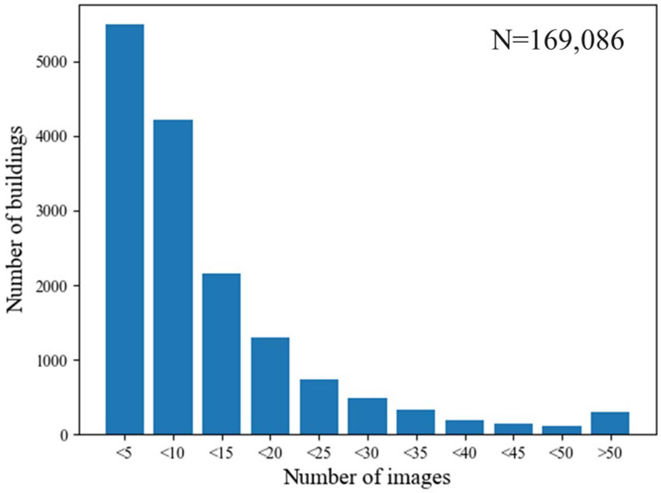


Fig. 9. Histogram of the number of images in each building.

tested the images of 1000 buildings The overall accuracy was 90.5%, which demonstrated the effectiveness of this method.

B. Classification Experiments

We distributed our building-level façade panorama image dataset into six classes based on year (–1962, 1963–1971, 1972–1980, 1981–1989, 1990–2001, and 2002– [see Fig. 10]) and three classes based on structure (RC, steel, and wooden [see Fig. 10]) for training to create the building structures and built-year classification models.

However, the quality of the source dataset was poor with many images having unnatural aspect ratios. Furthermore, some buildings were obscured by obstacles or there was more than one building in an image. In addition, class imbalance was another problem in our dataset. These factors misled the training of our deep learning model, and therefore, data filtering and balancing were required before training.

1) Data Filtering and Balancing: We randomly selected a part of the image from the wide-area source data, removed images with an $\text{aspect ratio} > 2$ or $\min\,(H,\,W) < 384$ (input images for networks), and manually removed images with multiple buildings in one image or where the buildings were obscured by obstacles.

Furthermore, we randomly selected images after manual filtering and kept the number of each class type almost identical to minimize the imbalance. We obtained a dataset, as summarized in Table I. The labeled data were split into training, validation, and test sets at a 60: 20: 20 ratio. We then conducted our test on a wide-area source dataset.

2) Training of Deep Learning Algorithms: Several preprocessing techniques, training strategies, and postprocessing techniques were applied to prevent overfitting and to maximize the performance of the algorithms.

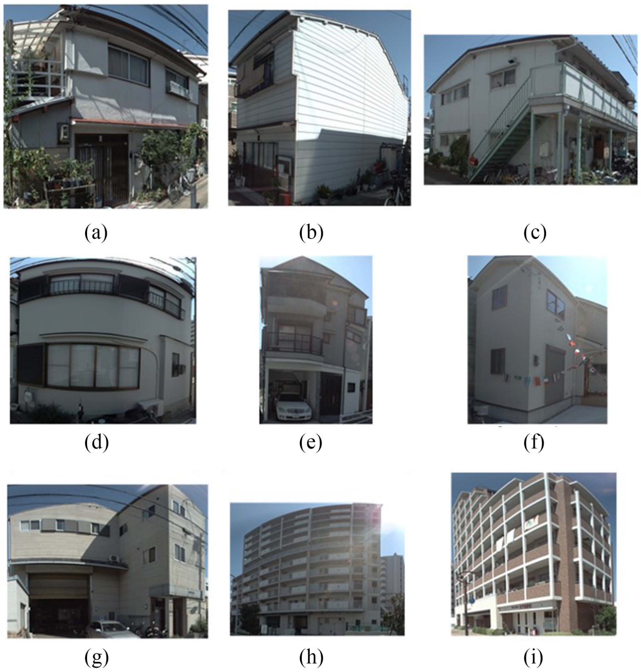


Fig. 10. Example of benchmark dataset for building structure and built year. (a) Wooden (built in 1962 or earlier). (b) Wooden (built from 1963 to 1971). (c) Wooden (built from 1972 to 1980). (d) Wooden (built from 1981 to 1989). (e) Wooden (built from 1990 to 2001). (f) Wooden (built in 2002 or later). (g) Steel (built 2002 or later). (h) RC (built from 1990 to 2001). (i) RC (built in 2002 or later).

TABLE III
Results of the Building Base Performances of the Eight Trained Networks on the Three Building-Structure Classes (Test Data)

Network	Recall	Precision	mF1	Accuracy
ShuffleNet V2	0.809	0.809	0.809	0.836
VGG19	0.859	0.845	0.850	0.873
MobileNet V2	0.879	0.873	0.875	0.894
ResNet101	0.869	0.865	0.867	0.888
ResNeSt101	0.894	0.903	0.898	0.911
T2T-ViT T-19	0.900	0.897	0.898	0.914
Swin transformer small	0.912	0.909	0.910	0.924
Swin transformer base	**0.923**	**0.925**	**0.924**	**0.935**

TABLE IV
Overall Accuracy, Recall, Precision, and F1 Score of Swin Transformer Based on Three and Two Structure Classes on the Test Data (Building Base)

	Recall	Precision	F1
RC	0.920	0.918	0.919
Steel	0.884	0.897	0.891
Wooden	0.965	0.959	0.962
Overall accuracy (3 classes) 0.935			
Non-wooden	0.949	0.956	0.952
Wooden	0.965	0.959	0.962
Overall accuracy (2 classes) 0.958			

We resized images to 224×224 or 384×384 following the bicubic mode and loaded pretrained models before inputting the images into the networks. Normalization, random resized cropping, random flipping $(\mathrm{rate} = 0.5)$, random erasing $(\mathrm{prob} = 0.25, \ \mathrm{area} = 0.02 \sim 0.33)$, and CutMix and MixUp $(\mathrm{prob} = 0.5)$ were applied for data augmentation. For training, we selected AdamW $(lr = 0.0005, \ \mathrm{decay} = 0.05, \ \mathrm{betas} = (0.9, \ 0.999))$ as the optimizer, cosine annealing $(\mathrm{warmup\_iters} = 1000)$ as the learning rate scheduler, and layer normalization in ViTs. We used test time augmentation as our postprocessing method.

The experiments used $\mathrm{Nvidia\ A100\ GPU} \times 8$ with $40\,\mathrm{GB} \times 8$ memory on a high-performance nationwide platform named "mdx," which accelerates Society 5.0 and can be operated by the University of Tokyo and eight other universities and two research institutes [52]. Considering the computational resources, we set the $\mathrm{batchsize}$ as 64 for each GPU and 100 epochs for each algorithm.

C. Classification Results

1) Building-Structure Classification: Table III presents the overall accuracy, recall, precision, and mean-F1 (mF1) scores for individual building units of the building-structure classification model attained by the learning approaches of the eight architectures, where mF1 represents the mean value of the F1 score calculated using (18) on each class. The results show that the three ViT networks (T2T-ViT T-19, Swin-transformer small, and Swin-transformer base) outperformed the other five DCNN networks. Furthermore, the building-structure classification results of the Swin-transformer base showed the highest accuracy.

Fig. 11 shows the image and building base confusion matrices of the structures obtained using the Swin-transformer base evaluated on the test data. Table IV summarizes the recall, precision, and F1 scores for each structure of the Swin-transformer base. All classes indicate accuracies of approximately 89% or higher. In addition, the building (see Table IV) and image bases show high accuracy with a slight difference between them. This result indicates that the structure classification task does not affect the number of images in each building. The specific classification accuracies are 0.963, 0.939, and 0.894 for wood, RC, and steel, respectively, for the image base, and 0.962, 0.919, and 0.891 for wood, RC, and steel, respectively, for the building base.

2) Built-Year Classification: Table V lists the overall accuracy, recall, and mF1 score of the building base for the built-year classification model determined using the learning approaches of the eight networks. The results showed that the built-year classification performance and building-structure classification of the ViT networks were better than those of the other five DCNN networks. Furthermore, the built-year classification results of the Swin-transformer base exhibited the highest accuracy and structural classification.

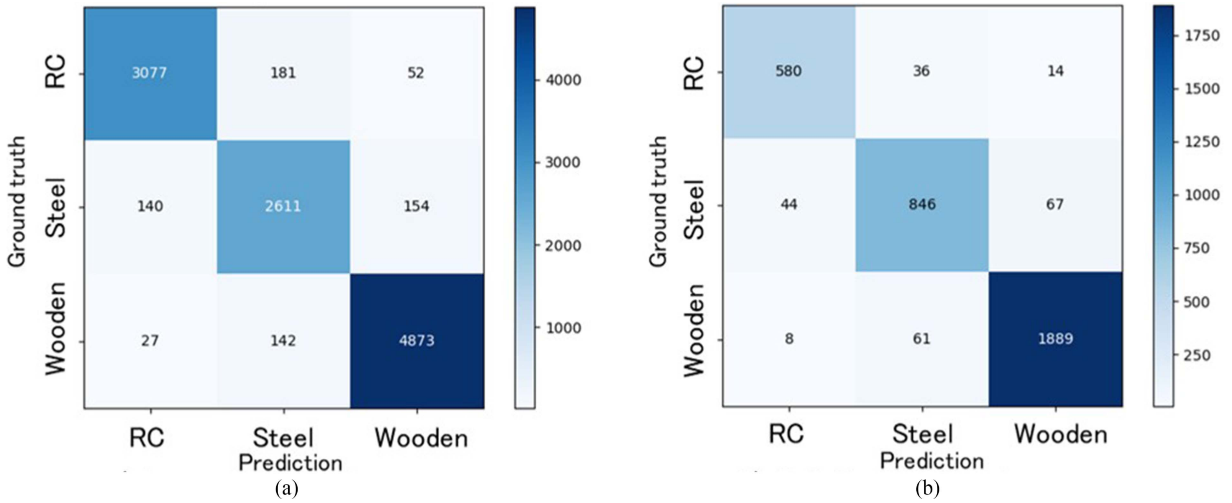


Fig. 11. Confusion matrices of the building-structure classification obtained by the Swin-transformer base networks evaluated on the test data. (a) Image base three classes; (b) Building base, three classes.

TABLE V
RESULTS OF THE BUILDING BASE PERFORMANCES OF THE EIGHT TRAINED NETWORKS ON THE SIX BUILT-YEAR CLASSES (TEST DATA)

Network	Recall	Precision	mF1	Accuracy
ShuffleNet V2	0.514	0.509	0.510	0.574
MobileNet V2	0.545	0.544	0.543	0.611
VGG19	0.546	0.541	0.542	0.606
ResNet101	0.555	0.555	0.553	0.621
ResNeSt101	0.571	0.573	0.571	0.645
T2T-ViT T-19	0.574	0.573	0.571	0.643
Swin transformer small	0.591	0.600	0.590	0.666
Swin transformer base	**0.597**	**0.602**	**0.597**	**0.679**

TABLE VI
RECALL, PRECISION, AND F1 SCORE OF SWIN TRANSFORMER ON THE SIX BUILT-YEAR CLASSES AND THREE BUILT-YEAR CLASSES ON THE TEST DATA (BUILDING BASE)

	Recall	Precision	F1
–1962	0.634	0.675	0.654
1963–1971	0.421	0.417	0.419
1972–1980	0.466	0.364	0.409
1981–1989	0.332	0.453	0.383
1990–2001	0.781	0.757	0.769
2002–	0.896	0.909	0.903
Overall accuracy (6 classes) 0.662			
–1980	0.949	0.914	0.931
1981–2001	0.848	0.882	0.864
2002–	0.896	0.909	0.903
Overall accuracy (3 classes) 0.902			

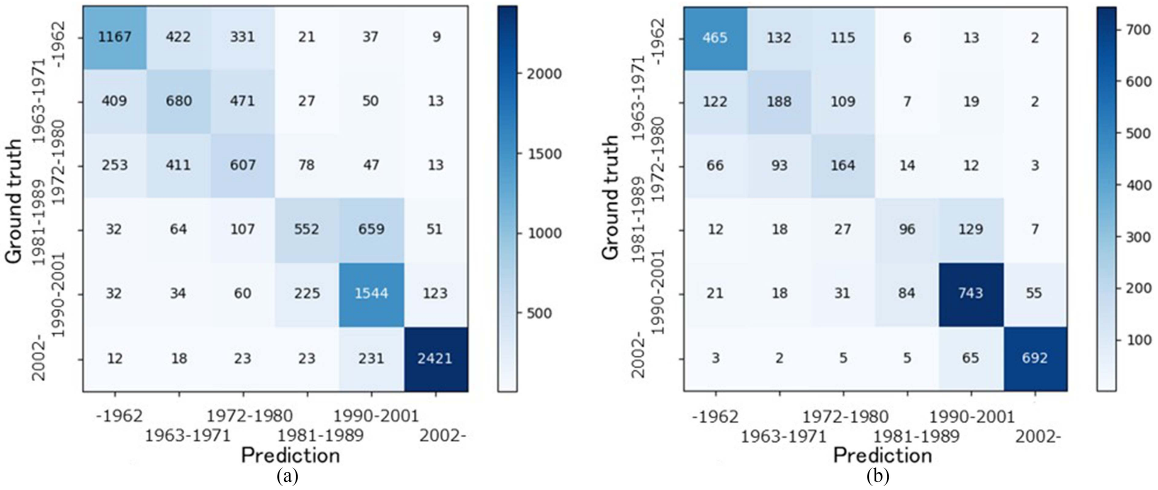
Fig. 12 shows the confusion matrix of the image and building bases for the built-year classification by the Swin-transformer base evaluated on the test data. Table VI provides the recall, precision, and F1 scores of the Swin-transformer base for the built year. This can be attributed to the same building in SVI being used for both classes, which suggests that the ability to acquire multiple SVIs of the same building can improve classification accuracy. The built year with the highest accuracy rate varies significantly depending on the period. The classification accuracy was high for 2002, followed by 1990–2001, and the classes before 1962. The classification accuracy for 1963–1989 was low; the classification accuracy for 1981 and 2001, when the seismic structural standards changed significantly, was 0.902 (three classes), which is high. Thus, the model can discriminate correctly, especially in terms of seismic resistance; therefore, its accuracy is very high when aggregated across the three classes.

D. Building-Structure and Built-Year Maps of Study Areas

We chose Kobe City for this study and evaluated the feasibility of mapping building structures and built years over a wide area from SVIs to demonstrate the efficacy of the proposed method in urban environments with high building densities.

Fig. 13 shows the predicted city-scale building-structure and built-year classification map along with the ground truth. In total, 195 105 panoramic images were used to classify 22 392 buildings. The comparison of the spatial distribution of the ground truth and prediction reveals that the distribution of the building structure and built year of the buildings is similar. The newer RC and steel buildings formed a coherent cluster along the railroad tracks in the center of Kobe City. Older wooden buildings were distributed in suburban residential areas far from the city center. The trend in the change in built-year and structural class varied depending on the street unit, which, in turn, depended on the area of use. Consequently, there may be regional trends, such as larger numbers of older buildings, in certain areas. It is difficult to determine a clear pattern in building units because, even in such areas, some buildings are undergoing structural conversions or renovations.

Although there may be certain regularities in structures, occasioned by regulations, zoning stipulations, or earthquake



(a)

Ground truth \ Prediction	-1962	1963-1971	1972-1980	1981-1989	1990-2001	2002-
-1962	1167	422	331	21	37	9
1963-1971	409	680	471	27	50	13
1972-1980	253	411	607	78	47	13
1981-1989	32	64	107	552	659	51
1990-2001	32	34	60	225	1544	123
2002-	12	18	23	23	231	2421

(b)

Ground truth \ Prediction	-1962	1963-1971	1972-1980	1981-1989	1990-2001	2002-
-1962	465	132	115	6	13	2
1963-1971	122	188	109	7	19	2
1972-1980	66	93	164	14	12	3
1981-1989	12	18	27	96	129	7
1990-2001	21	18	31	84	743	55
2002-	3	2	5	5	65	692

Fig. 12. Confusion matrices of built-year classification by the Swin-transformer base networks evaluated on the test data. (a) Image base, six classes. (b) Building base, six classes.

preparation measures along emergency transportation routes, this is only one aspect to be considered.

The trend in the change in built-year and structural class varied depending on the street unit, which, in turn, depends on the area of use. This map indicates that the spatial distribution of the built year and structure is uneven in cities, which implies that the detailed urban data can be developed using SVIs. Furthermore, the seismic risk is affected strongly by differences in the building structure and built year; this implies that the seismic risk may be higher in the suburbs where older and smaller buildings are densely located.

V. Discussion

A. Automated Annotation for the Development of Training Data

To date, street view imaging has generated the highest resolution images because of the development of sensing technologies, such as image sensors. This creates new opportunities for the detailed understanding of the physical environment of cities. However, the successful integration of different sources, such as GIS building data, SVIs, and the combination of GIS data with diverse types of information, is yet to be fully explored. Therefore, in many studies, the annotation of semantic information in SVIs is performed manually; however, preparing a training image dataset is costly and time-consuming [25], [26], [32].

In this study, we demonstrated the automatic annotation of GIS attribute information to individual building images from SVIs to annotate image datasets efficiently when the information required for estimation is available as GIS building data. It is desirable to acquire images from many angles to understand a building in detail. The proposed method can acquire ten images per building for over 37% of the buildings at 2.5 m intervals; the automatic annotation method combined with SVIs and GIS building data can provide considerable amounts of training data for developing image-based models.

When GIS building data and SVIs are combined for a high-density area (such as an urban area), simply linking the nearest building and shooting point may not help to annotate the building accurately if the vehicle's travel direction and AOV are not considered. Previous studies used images recorded with the camera direction perpendicular to the travel direction, and this helped calculate the compass direction of each detected building easily based on its relative position in the image [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40]. However, a 360° SVI in a high-density building area cannot correctly annotate the buildings with GIS building data because the buildings are densely packed. The data captured by GSVs have uneven intervals and 360° or wider AOVs; hence, as observed in the proposed method, the AOVs of the buildings and the footprint of the GIS building data will reduce the error.

Although the building extraction and development of training data are automated, the constructed extraction method still suffers from issues in terms of application to dense urban wide areas. This problem can be attributed to three specific reasons.

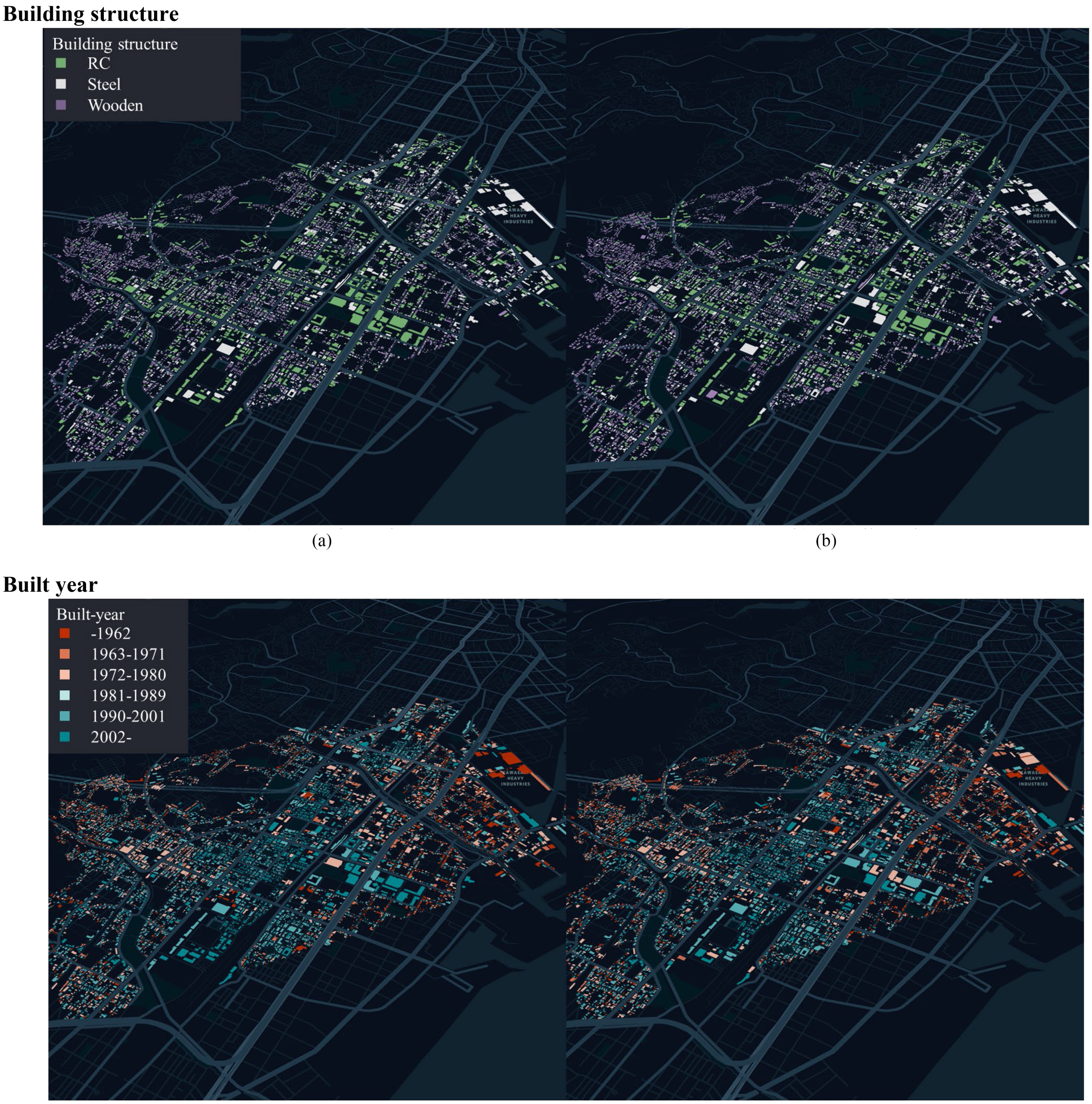
1) Multiple buildings are detected as one building because of the poor building detection accuracy of the object detection model.
2) Buildings that are not the target are included, especially because of the different periods between the data.
3) Images include obstacles, such as trees and trucks, that overlap with buildings.

For 1 and 2, images with large aspect ratios (i.e., high-rise buildings in the foreground and background detected as one building) can be deleted automatically using the proposed method. For 2 and 3, there are two possible solutions for future studies.

1) Perform binary classification by learning incorrect and correct images to automatically filter them.
2) Perform segmentation again and filter the images based on the percentage of pixels that are not buildings, but the computational cost will be high.

B. Classification of the Built Year and Structure of Individual Buildings

Building-structure and built-year data are essential for assessing earthquake and tsunami risks. In terms of the classification



(a) (b)

(c) (d)

Fig. 13. City-scale predicted building-structure (b) and built-year classification map (d) along with the ground truth (a and c), where different colors represent different building-structure classes. The total number of buildings in this area is 22 392. A total of 6209 buildings are not classified considering that no corresponding SVIs are found or buildings are not located along the road. In addition, the map shows that the spatial distribution of the predicted building structure and built year by the proposed model matches the ground truth. (a) Ground truth. (b) Predicted. (c) Ground truth. (d) Predicted.

accuracy of the Swin transformer based on the building-structure test data, the two-class accuracy (0.958) was estimated to be equal to that of the three classes (0.935). Oki and Ogawa [17] found it difficult to classify steel and RC with high accuracy using low-resolution real estate images using a CNN (F1 = 0.66 and 0.84). However, in the current study, both classes were classified with high accuracy (F1 = 0.891 and 0.919), which encourages the combined use of high-resolution SVIs and the ViT architecture based on the high accuracy in classifying steel and RC buildings. In addition, this model can classify the building structure and age by learning the texture of the building façade, the overall building shape, thickness and ratio of columns and beams, window sizes and their designs, and other factors because it employs DCNN and ViT. Therefore, although the accuracy may be degraded slightly for buildings whose facades have been remodeled, such as by repainting, it should not be considered a significant problem that questions the usefulness of the model.

It is important to discuss the examples of false predictions to better understand the difficulties the model encounters. In the case of misclassification, the images have large trees [see Fig. 14(a)] and large buildings in the background [see

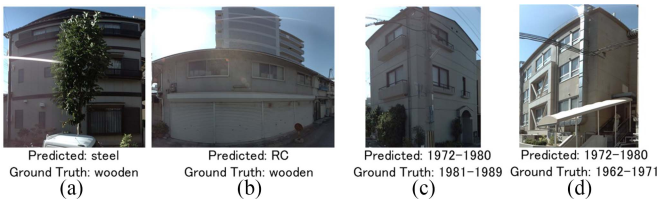


Fig. 14. Examples of misclassified buildings. Misclassifications of (a) steel as wooden, (b) RC as wooden, (c) built years 1972–1980 as 1981–1989, and (d) built years 1972–1980 as 1962–1971.

Fig. 14(b)]. We considered excluding images with trees and other plants because the buildings were captured from many angles. The latter type of misclassification depends on the accuracy of object detection. Therefore, it is possible to deal with this problem by increasing the variety of training data to remove the buildings in the background.

In addition, the accuracy of the three building-based classes (0.902) was considerably higher than that of the six classes (0.662) based on the classification accuracy of the Swin transformer for the test data of the built-year classes. This can be attributed to the seismic standards of the building standard law in Japan revised in 1981 and 2002. The revisions in the seismic standards added restrictions on the placement of windows and columns, which suggests that the design of the house is reflected in the façade and is relatively easy to recognize. For seismic risk assessment, classification accuracy is important for the three classes of buildings around 1981 and 2002. Buildings with the same seismic standards, such as 1981–1989 and 1990–2001, include classes that cannot be classified easily. Therefore, the misclassification of the six-class classification of the built year occurs mostly in adjacent classes with boundaries between 1981 and 2002 (see Fig. 12). In other words, a similar appearance before 1980 is represented in the confusion matrix and F1 score, which could affect the classification results. The façade in Fig. 15 shows an example of a typical image in which 1981–1989 is misclassified as 1972–1980. Additionally, seismic retrofitting has been promoted in recent years, and its typical examples include reinforced buildings with bracing and other wall reinforcements that increase the strength of the columns. The large buildings with these reinforced braces and frames installed on the façade may have caused the misclassification [see Fig. 14(d)].

We compared our model with an existing model proposed by Oki and Ogawa [17] that combines images and attributes. The results showed that the accuracy of the multimodal model [Transform with Sparse modeling (SpM)], which added building attributes (number of floors and area of the building) to the Swin transform as features, had comparable results for both the building structure and built year (see Fig. 15). This indicates that, from the SVI, the Swin transform can learn the number of floors and area information from the façade. In addition, comparing the accuracy with that achieved in the previous study demonstrated significant improvement over the results of using accurate Japanese real estate images with built-year and building-structure classifications of 0.367 and 0.786, respectively [17]. A previous study using SVI data achieved an accuracy of 0.869–0.871 in material classification in Chile [25] and an accuracy of 0.614 and 0.81 in built age classification in Austria and in Amsterdam, respectively [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30].

SVIs have the advantages of low cost, ease of use, and time-saving, whereas their limitation is image resolution [53]. However, in this study, street view imaging has a shorter acquisition interval and higher resolution than GSV, and hence, it was used to acquire images from successive angles for a single ground object. Therefore, conventional limitations are overcome, which indicates that the development of sensing technology can contribute to the development of research using SVIs.

C. Visual Explanations of Classification Models

Deep learning approaches are known for being black boxes and are criticized for lacking interpretability. The inputs and labels are fed into the network, and a model can be trained without any knowledge of the learning decisions derived using deep learning. Moreover, if the training and inferencing can be visualized, additional deep learning details can be understood. To address this problem, this study implemented Grad-CAM [54] to further explain the proposed built-year and structure classification model.

Grad-CAM is an extension of the original class activation map (CAM) method [55]. Unlike CAM, Grad-CAM considers the feature map gradients in the backpropagation instead of the output of the global average pooling (GAP) layers as weights to process the feature maps. Therefore, Grad-CAM can be deployed for all deep learning methods even without GAP layers, particularly the Swin transformer. As for the visualization results, Grad-CAM can determine the point of focus for a network. In this study, the localization map highlighting the important regions in the image visualization results (see Fig. 16) shows that the attention of the Swin transformer focused on the key building points where the Building Standards Law in Japan [56] had shape and standards' requirements. The red regions correspond to the high score for the class.

According to the Building Standards Law implemented in 1981 in Japan, buildings are required to have a well-shaped floor plan and no overhangs in terms of elevation. An overhang refers to a multistory building with an upper floor wider than the lower floors. Furthermore, the proportion of nonopening walls on each side has been defined since 1981. In addition, because of the Building Standards Law revision in 2002, buildings must consider the balance between the amount and arrangement of wall surfaces depending on the quadrant method. The width of each building floor in the east–west and north–south directions is divided into four equal parts, and the ratio and balance of the wall volume at the sides of these parts are specified. Furthermore, the building regulations for earthquake resistance have been revised to meet the stipulated seismic performance and the specifications for the thickness and number of columns (amount of reinforcing steel) and braces for the column joints. These specifications are reflected in the façade design.

Consequently, the Swin transformer focused on the key façade points and successfully learned the key semantic features, which resulted in a high performance in classifying the built year and structure.

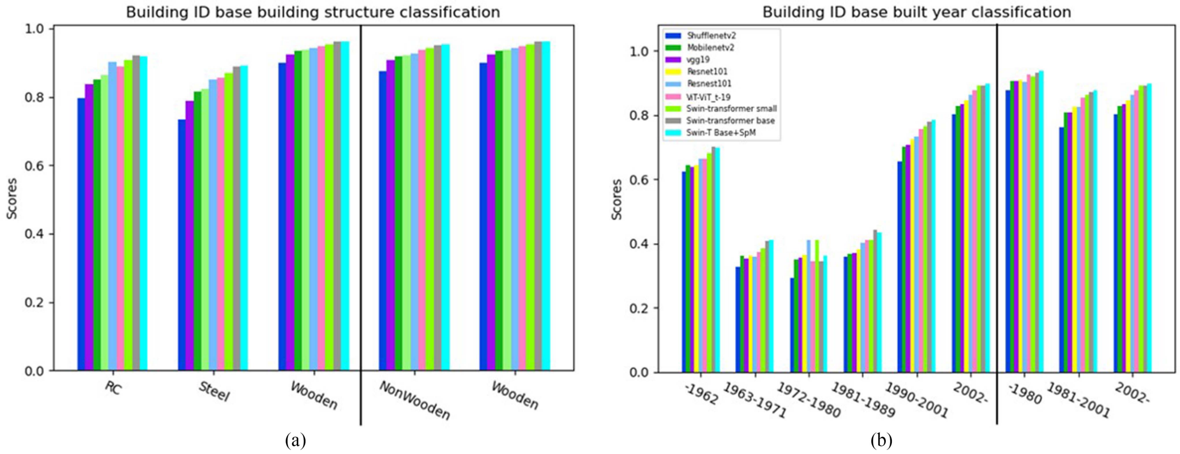


Fig. 15. F1-scores of the eight trained networks and Swin-transformer base + the SpM model in the building base. (a) Three building-structure classes (left) and two building-structure classes (right), and (b) six built-year classes (left) and three built-year classes (right). For almost all classes, the Swin-transformer base achieves the highest F1 score, similar to the Swin transformer + the SpM (multimodal model of using images and building attribute (area and number of floors)).

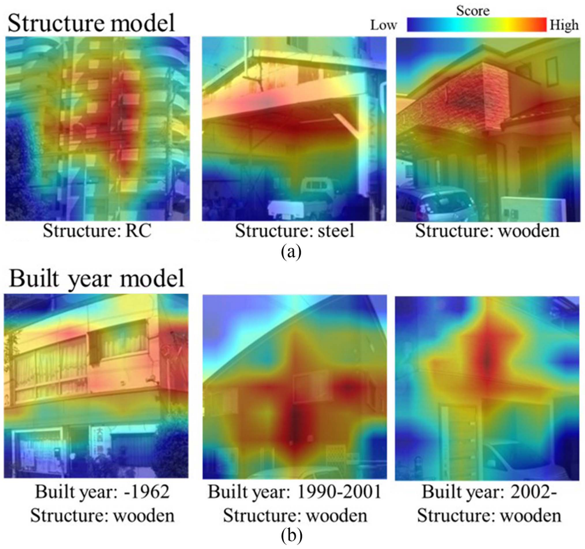


Fig. 16. Visualization results of Grad-CAM. (a) Structure model. (b) Built-year model.

D. ViT Versus DCNN

In CV, attention is either applied in conjunction with convolutional networks or used to replace certain components of convolutional networks while maintaining their overall structure in place. The Swin transformer is a ViT, which shows that this reliance on DCNNs is unnecessary, and a pure transformer applied directly to the sequences of image patches can perform significantly well in image classification. In addition, ViTs achieve excellent results compared with the state-of-the-art convolutional networks when they are pretrained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (e.g., ImageNet and VTAB). Therefore, they require substantially fewer computational resources to train.

However, in many existing studies that have used SVIs, models based on the DCNN architecture (e.g., VGG and ResNet) have been adopted as the most accurate. Thus, our experimental results indicate that the ViT architecture can potentially improve the classification accuracy of many existing studies using SVI.

E. Limitations

A limitation of this study is that it uses SVIs taken by vehicles, and it does not provide information on narrow locations through which vehicles cannot pass. This can be said for all studies that use street image sources, such as GSV in urban areas where there are dense narrow areas with wooden areas [7]. This problem can be resolved by capturing images with a smartphone fixed to a bicycle. However, a method is needed to ensure that its stability matches that of the vehicle camera [57]. Furthermore, it is important to consider the computational cost considering that we used large numbers of high-resolution SVIs. As indicated in Table II, the computational cost differs significantly depending on the network. Therefore, networks with a slightly lower classification accuracy must be used to reduce the cost (ShuffleNet V2 and MobileNet V2) with limited computational resources when applying the method over a wide area. However, the trained model's applicability to other cities in Japan must be verified.

The built-year and structure classification results were significantly influenced by the design requirements of the Japanese Building Standards Law, and as the street view styles and building regulations and laws differ by country, it is difficult to use our trained model elsewhere. We believe that if big data of street images are exclusively collected using the method in this study, with fine-tuning, similar accuracy can be expected with a relatively small number of image samples. If building professionals have difficulty determining the built year from a building's exterior, the built year may not significantly affect the building's seismic resistance or other specifications. Therefore, the important built-year classes for each country should be considered when applying this method in other countries.

Furthermore, spending considerable time on minor changes or adjustments to the Swin-transformer architecture and continued overfitting for slight accuracy improvements on the Japanese dataset is discouraged. In the future, the robustness and generalizability of our method will be verified using datasets from other countries because our approach produced excellent performance on the Kobe wide-area dataset.

VI. Conclusion

This article presents a large-scale framework for mapping and classifying building structures and built years obtained from SVIs and GIS building data to be used for earthquake and tsunami disaster risk assessment. We achieved relatively high accuracy in classifying the structure and built year of individual buildings using the existing GIS building data to assemble an annotated dataset using an automatic annotation method. Associated experiments were conducted over a wide area.

Based on the results of testing eight different architectures used in Kobe City, we selected the Swin-transformer base to estimate the built-year and building-structure classification of individual buildings on an urban scale. Furthermore, we compared the model with images only and a multimodal model (Swin transformer+SpM) that combines images and basic GIS building attributes. The results showed that the building attributes had little effect, which implies that the model may learn from the image to include the volume and number of floors of a building if multiple high-resolution SVIs can be obtained for an individual building. We were able to achieve relatively high accuracy levels (0.94 for three classes and 0.96 for two classes of structure, and 0.66 for six classes and 0.96 for three classes of built year) on the test data and mapped the entire city of Kobe from SVIs, which illustrates the possibility of efficiently developing building-structure and age information for individual buildings, which is essential for earthquake and tsunami risk assessments. Such building structures and built-year maps can potentially be used for disaster risk assessment and urban analysis with very high resolution and precision, and they can replace conventional statistical data (aggregated data).

We believe the proposed method can identify dense wooden-building areas and old built-year areas and can be used in combination with the existing building damage estimation models to identify areas that are at a higher risk. These should be considered as a priority for disaster prevention planning.

For future works, we plan to consider an appropriate GSV shooting interval to improve the computational cost and maintain accuracy while reducing the volume of images. To improve the building coverage rate, we will develop a method for capturing SVIs with the same quality as that of vehicles by using cyclists in narrow streets where vehicles cannot pass.

References

[1] T. Nagao, F. Yamazaki, and M. Inoguchi, "Analysys of building damage in Kashiwazaki city due to the 2007 Niigata-ken Chuetu-oki earthquake," in *Proc. 32nd Asian Conf. Remote Sens.*, 2011, p. 6.

[2] C. del Gaudio et al., "Empirical fragility curves from damage data on RC buildings after the 2009 L'Aquila earthquake," *Bull. Earthq. Eng.*, vol. 15, pp. 1425–1450, 2017.

[3] N. Yamaguchi and F. Yamazaki, "Fragility curves for buildings in Japan based on damage surveys after the 1995 Kobe earthquake," in *Proc. 12th World Conf. Earthq. Eng.*, 2000, Paper 2451.

[4] A. Suppasri et al., "Building damage characteristics based on surveyed data and fragility curves of the 2011 Great East Japan tsunami," *Natural Hazards*, vol. 66, pp. 319–341, 2013.

[5] S. Medina, J. Lizarazo-Marriaga, M. Estrada, S. Koshimura, E. Mas, and B. Adriano, "Tsunami analytical fragility curves for the Colombian Pacific coast: A reinforced concrete building example," *Eng. Struct.*, vol. 196, 2019, Art. no. 109309.

[6] Cabinet Office, "The damage estimation on the Nankai trough megathrust earthquake," 2012, Accessed: Dec. 20, 2021. [Online]. Available: https://iisee.kenken.go.jp/symposium/10thIWSMRR/10.pdf

[7] The Building Center of Japan, "Introduction to the building standard law—Building regulation in Japan," 2013, Accessed: Dec. 20, 2021. [Online]. Available: https://www.bcj.or.jp/upload/international/baseline/BSLIntroduction201307_e.pdf

[8] T. Osaragi and T. Oki, "Wide-area evacuation simulation incorporating rescue and firefighting by local residents," *J. Disaster Res.*, vol. 12, pp. 296–310, 2017.

[9] S. Mangalathu et al., "Classifying earthquake damage to buildings using machine learning," *Earthq. Spectra*, vol. 36, pp. 183–208, 2020.

[10] M. Wieland, M. Pittore, S. Parolai, J. Zschau, B. Moldobekov, and U. Begaliev, "Estimating building inventory for rapid seismic vulnerability assessment: Towards an integrated approach based on multi-source imaging," *Soil Dyn. Earthq. Eng.*, vol. 36, pp. 70–83, 2012.

[11] S. Steimen, D. Fah, D. Giardini, M. Bertogg, and S. Tschudi, "Reliability of building inventories in seismic prone regions," *Bull. Earthq. Eng.*, vol. 2, pp. 361–388, 2004.

[12] Y. Ogawa, Y. Sekimoto, and R. Shibasaki, "Estimation of earthquake damage to urban environments using sparse modeling," *Environ. Plan. B, Urban Anal. City Sci.*, vol. 48, pp. 1075–1090, 2021.

[13] Federal Emergency Management Agency, *Rapid Visual Screening of Buildings for Potential Seismic Hazards: A Handbook*, 3rd ed. Washington, DC, USA: Federal Emerg. Manage. Agency, 2015, p. 154.

[14] F. Rivera, M. Hube, H. Santa-Maria, and C. Alvarez, "Use of remote digital surveys to generate exposure models of residential structures in Chile," in *Proc.16th World Conf. Earthq. Eng.*, 2017, Paper 2414.

[15] M. Matsuoka et al., "Development of building inventory data and earthquake damage estimation in Lima, Peru for future earthquakes," *J. Disaster Res.*, vol. 9, pp. 1032–1041, 2014.

[16] H. Miura and S. Midorikawa, "Updating GIS building inventory data using high-resolution satellite images for earthquake damage assessment: Application to metro Manila, Philippines," *Earthq. Spectra*, vol. 22, pp. 151–168, 2006.

[17] T. Oki and Y. Ogawa, "Model for estimation of building structure and built year using building façade images and attributes obtained from a real estate database," in *Urban Informatics and Future Cities*. Berlin, Germany: Springer, 2021, pp. 549–573.

[18] M. Zeppelzauer, M. Despotovic, M. Sakeena, D. Koch, and M. Doller, "Automatic prediction of building age from photographs," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2018, pp. 126–134.

[19] R. Nagasawa, E. Mas, L. Moya, and S. Koshimura, "Model-based analysis of multi-UAV path planning for surveying postdisaster building damage," *Sci. Rep.*, vol. 11, 2021, Art. no. 18588.

[20] Y. Ogawa, Y. Akiyama, H. Sengoku, and R. Shibasaki, "Evaluation of catastrophic earthquake damage throughout Japan using estimated micro data," in *Proc. CUPUM Conf. Papers*, 2013, vol. 103, pp. 1–30.

[21] B. Borzi et al., "Vulnerability study on a large industrial area using satellite remotely sensed images," *Bull. Earthq. Eng.*, vol. 9, pp. 675–690, 2011.

[22] Y. Sakai, N. Fukukawa, and K. Arai, "Development of classification and story building data for accurate earthquake damage estimation," *J. Jpn. Assoc. Earthq. Eng.*, vol. 9, no. 5, pp. 21–28, 2011.

[23] F. Biljecki and M. Sindram, "Estimating building age with 3D GIS," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. IV-4/W5, pp. 17–24, 2017.

[24] J. F. Rosser, D. S. Boyd, G. Long, S. Zakhary, Y. Mao, and D. Robinson, "Predicting residential building age from map data," *Comput., Environ. Urban Syst.*, vol. 73, pp. 56–67, 2019.

[25] P. A. Pelizari, C. Geib, P. Aguirre, H. S. Maria, Y. M. Pena, and H. Taubenbock, "Automated building characterization for seismic risk assessment using street-level imagery and deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 180, pp. 370–386, 2021.

[26] C. Wang et al., "Machine learning-based regional scale intelligent modeling of building information for natural hazard risk management," *Autom. Construction*, vol. 122, 2021, Art. no. 103474.

[27] G. C. Iannelli and F. Dell'Acqua, "Extensive exposure mapping in urban areas through deep analysis of street-level pictures for floor count determination," *Urban Sci.*, vol. 1, 2017, Art. no. 16.

[28] F. Ghione, S. Maeland, A. Meslem, and V. Oye, "Building stock classification using machine learning: A case study for Oslo, Norway," *Front. Earth Sci.*, vol. 10, 2022, Art. no. 886145.

[29] M. Sun, F. Zhang, and F. Duarte, "Automatic building age prediction from street view images," in *Proc. 7th IEEE Int. Conf. Netw. Intell. Digit. Content*, 2021, pp. 102–106.

[30] M. Sun, F. Zhang, F. Duarte, and C. Ratti, "Understanding architecture age and style through deep learning," *Cities*, vol. 128, 2022, Art. no. 103787.

[31] J. Kang, M. Korner, Y. Wang, H. Taubenbock, and X. X. Zhu, "Building instance classification using street view images," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 44–59, 2018.

[32] S. Zou and L. Wang, "Detecting individual abandoned houses from google street view: A hierarchical deep learning approach," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 298–310, 2021.

[33] G-Space Information Center, Accessed: Dec. 20, 2021. [Online]. Available: https://www.geospatial.jp/gp_front/

[34] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, vol. 1, pp. 91–99.

[36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.

[38] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[39] Y. Wu et al., Detectron2, 2019, Accessed: Dec. 20, 2021. [Online]. Available: https://github.com/facebookresearch/detectron2

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.

[41] Z. Liu et al., "Swin transformer V2: Scaling up capacity and resolution," 2021, *arXiv:2111.09883*.

[42] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," 2022, *arXiv:2205.01917*.

[43] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[44] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," 2021, *arXiv:2106.04803*.

[45] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 122–138.

[46] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[49] H. Zhang et al., "ResNeSt: Split-attention networks," 2020, *arXiv:2004.08955*.

[50] X. Zhai et al., "Scaling vision transformers," 2021, *arXiv:2106.04560*.

[51] Kepler.gl, Accessed: Dec. 20, 2021. [Online]. Available: https://kepler.gl/

[52] T. Suzumura et al., "mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations," in *Proc. IEEE Int. Conf. Cloud Big Data Comput.*, 2022, pp. 1–7.

[53] A. Rzotkiewicz, A. L. Pearson, B. V. Dougherty, A. Shortridge, and N. Wilson, "Systematic review of the use of google street view in health research: Major themes, strengths, weaknesses and possibilities for future research," *Health Place*, vol. 52, pp. 240–246, 2018.

[54] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[55] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[56] The Building Standards Law, Accessed: Dec. 20, 2021. [Online]. Available: https://www.mlit.go.jp/

[57] S. Sony, S. Laventure, and A. Sadhu, "A literature review of next-generation smart sensing technology in structural health monitoring," *Struct. Control Health Monit.*, vol. 26, 2019, Art. no. e2321.

Yoshiki Ogawa received the Ph.D. degree in environment from The University of Tokyo, Tokyo, Japan, in 2016.

He is currently a Lecturer with the Center for Spatial Information Science, The University of Tokyo. His area of research is big data analysis of geographic information system data and remote sensing data and developing integrated simulation system of gigantic earthquake and tsunami disasters using urban data.

Chenbo Zhao received the master's degree in engineering from Wuhan University, Wuhan, China, in 2020. He is currently working toward the Ph.D. degree with the Department of Civil Engineering, University of Tokyo, Tokyo, Japan.

His research interests include Big data analysis of GIS data remote sensing data and computer vision algorithms.

Takuya Oki received the Ph.D. degree in engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 2018.

He is currently an Associate Professor with the School of Environment and Society, Tokyo Institute of Technology. His research interest is developing methods that utilize big data and AI technologies to provide evidence on which to base architectural planning and urban development.

Shenglong Chen (Graduate Student Member, IEEE) received the master's degree in engineering from Tongji University, Shanghai, China, in 2018. He is currently working toward the Ph.D. degree with the Department of Civil Engineering, University of Tokyo, Tokyo, Japan.

His research focuses on GeoAI of remote sensing.

Yoshihide Sekimoto is currently a Professor with the Center for Spatial Information Science, University of Tokyo, Tokyo, Japan. In December 2020, he started his main duties at the Center for Spatial Information Science and the university-wide Research Organization for Digital Spatial Society, where he became the Director of the organization in April 2021. He is also continues to work at the Institute of Industrial Science as a specially appointed Professor. His research interests include mobile sensing, people movement analysis infrastructure data management.