ICIC Express Letters
Part B: Applications
Volume **15**, Number **6**, June **2024**

ICIC International ©2024 ISSN 2185-2766
pp. 575–584

ESTIMATION OF SEDIMENT DISASTER PRONE AREAS FROM MAP IMAGES USING MASK R-CNN

Keigo Miwa[1], Noritaka Shigei[1,*], Satoshi Sugimoto[2], Yoichi Ishizuka[2] and Hiromi Miyajima[1]

[1]Graduate School of Science and Engineering
Kagoshima University
1-21-40 Korimoto, Kagoshima City, Kagoshima 890-0065, Japan
{ k7306768; k2356323 }@kadai.jp; *Corresponding author: shigei@ibe.kagoshima-u.ac.jp

[2]Graduate School of Engineering
Nagasaki University
1-14 Bunkyo, Nagasaki City, Nagasaki 852-8521, Japan
{ s-sugi; isy2 }@nagasaki-u.ac.jp

Received September 2023; accepted November 2023

Abstract. *In Japan, where sediment disasters such as debris flow, landslides, and slope failures frequently occur, for disaster prevention purposes, the government publishes areas at risk of sediment disasters as Sediment Disaster Prone Areas. Since this requires time and effort for experts to conduct field surveys, attempts are being made to estimate the collapse hazard areas from elevation maps using deep learning. However, the conventional method of estimating collapse hazard areas pixel by pixel has the problem that it is difficult to identify the area clearly. To solve these problems, we propose using instance segmentation and object detection in computer vision techniques. This study investigates effective methods for estimating areas at risk of sediment disasters using a Mask R-CNN model that simultaneously processes segmentation and object detection. Specifically, we develop 1) a method for generating mask images, 2) a learning method combining Red Relief Image Map images, aerial photographs, standard maps, and land use maps as input images, and 3) a method for overlapping prediction. Numerical experiments conducted under multiple conditions evaluate the accuracy and the predicted images and demonstrate the effectiveness of the developed methods.*

Keywords: Sediment disaster prone area, Mask R-CNN, Red Relief Image Map, Object detection, Instance segmentation

1. **Introduction.** In Japan, where 70% of the land is mountainous, sediment disasters such as landslides are frequent due to typhoons, earthquakes, and heavy rainfall. For disaster prevention, the government publishes areas where steep slopes are in danger of collapsing as Sediment Disaster Prone (SDP) Areas [1]. However, these surveys are conducted manually by experts and others and require time and labor. In contrast, attempts have been made to estimate risk topography from land elevation maps using deep learning [2, 3]. In [2], an attempt has been made to estimate the location of collapse hazard areas on steep slopes from RRIM (Red Relief Image Map) [4] images and aerial photographs using a CNN (Convolutional Neural Network). In [3], methods using Pix2pix, a deep learning model capable of image transformation, have been investigated. It has been shown that 1) the estimation accuracy was better when the SDP Areas was used as the correct location rather than the Steep Slope Areas in Danger of Failure and 2) although Pix2pix can eliminate missed collapse hazard areas on steep slopes, the accuracy needs to be improved. All the methods proposed in these studies determine hazard areas pixel-by-pixel. However, such an approach has the problem that it is difficult to

DOI: 10.24507/icicelb.15.06.575

identify areas clearly. A possible solution to this problem is to use semantic segmentation, Instance Segmentation (IS) or Object Detection (OD) methods to target hazardous areas for segmentation and detection. In [5], although semantic segmentation is applied for RRIM, it uses an improved version of U-Net [6] and is used to detect roads and rivers instead of hazardous areas. In [7, 8, 9], IS and OD methods using Mask R-CNN (Mask Region-based CNN) have successfully applied to post-disaster landslide detection from aerial photographs. However, the methods do not estimate the area where landslides are likely to occur. In [10], landslide detection is performed from high-definition 1m resolution DEM data before collapse using a CNN. The generated landslide susceptibility map is consistent with the actual distribution trends. However, it is difficult to apply this method to the entire Japan because the needed high-definition DEM data is not always available everywhere. Although the effectiveness of landslide detection using IS and OD has been shown, no method has been proposed for pre-disaster detection such as SDP Areas detection that can be applied to arbitrarily large areas.

In this study, we investigate effective methods using Mask R-CNN [11] for estimating collapse hazard areas on steep slopes. Unlike U-Net, Mask R-CNN provides IS in addition to OD and class identification. It is highly versatile and has a variety of uses, such as detecting people and buildings from photographs. This paper proposes 1) a method for generating mask images, 2) a learning method combining RRIM images, aerial photographs, standard maps, and land use maps as input images, and 3) a method for overlapping prediction. Numerical experiments conducted under multiple conditions evaluate the accuracy and the predicted images and demonstrate the effectiveness of the developed methods.

2. **Mask R-CNN.**

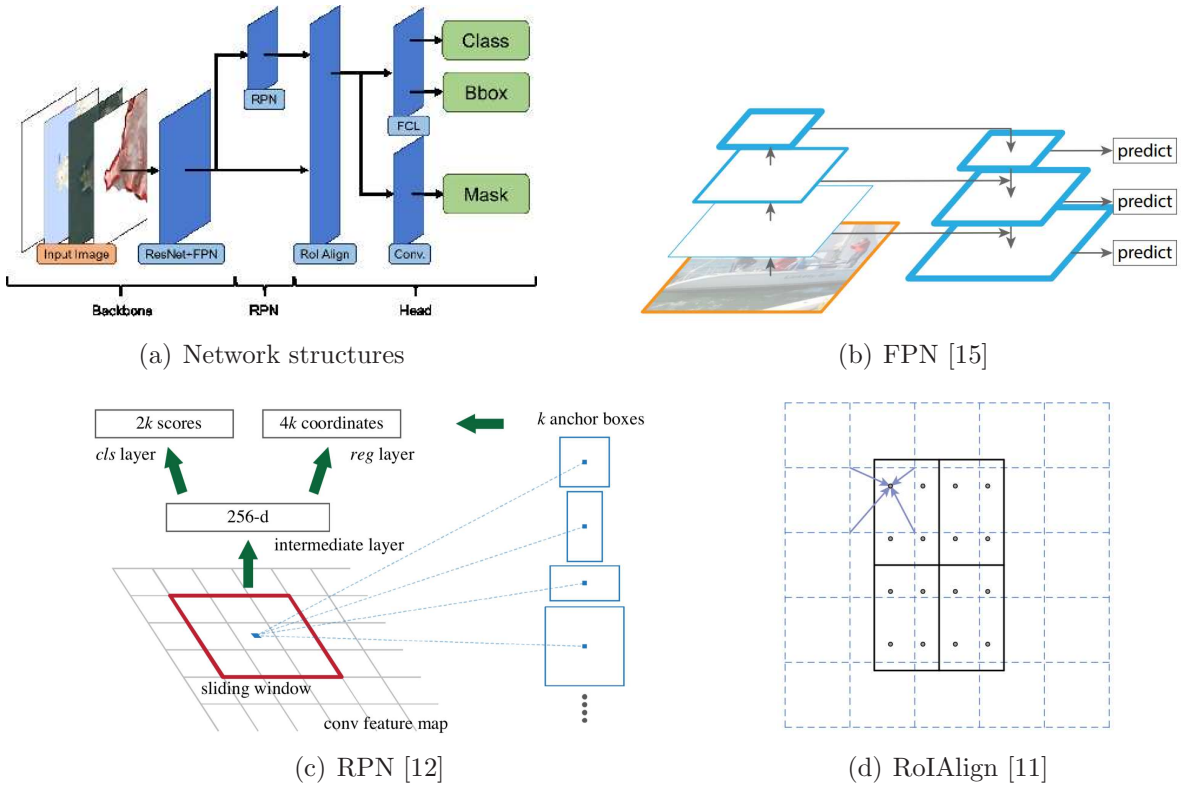
Mask R-CNN extends Faster R-CNN (Faster Region-based CNN) [12] to provide IS in addition to OD and class identification. It is possible to detect regions in the image that are considered to be objects and the class that these regions represent, and to identify classes at the pixel level within the resulting regions. The network structure of the Mask R-CNN is shown in Figure 1(a). Mask R-CNN consists of three major layers: Backbone, RPN (Region Proposal Network), and Head. Backbone extracts features from the input image, RPN proposes RoIs (Region of Interests) where objects to be detected are located, and Head performs object classification, bounding-box regression and mask prediction.

2.1. **Backbone.**

The backbone of the Mask R-CNN implementation [13] used in this study is ResNet50 (Residual Network 50) with the addition of FPN (Feature Pyramid Network). Backbone extracts features from the input image, starting with low-level features such as edges and corners specified in the initial layer and successively detecting higher-level features in later layers. ResNet50 is a deep CNN proposed in [14], and processing images in the convolution layer of this model results in a pyramid of feature maps with high semantic value, although the spatial resolution decreases as the number of convolution stages increases. FPN proposed in [15] further processes the output layer of ResNet50 in a top-down path, as shown in Figure 1(b), to build higher resolution layers from semantically valuable layers. This allows access to low-level and high-level features at each stage, improving the representation of objects at multiple scales.

2.2. **RPN.**

The network structure of the RPN is shown in Figure 1(c). For the feature map extracted by Backbone, a bounding box of region candidates and a score representing the object-like nature of the region are output. The loss used for training is as follows: $L_{CLS} + L_{BOX} = \frac{1}{N_{cls}} \sum_i (L_{cls}(p_i, p_i^*)) + \lambda \frac{1}{N_{reg}} \sum_i (p_i^* L_{reg}(t_i, t_i^*))$, where i is the index of the anchor in the mini-batch, p_i is the predicted probability that anchor i is an object, and p_i^* indicates actually whether it is an object or not. t_i is a vector representing the coordinates of the predicted bounding box and t_i^* is a vector representing the coordinates of the correct



(a) Network structures (b) FPN [15]

(c) RPN [12] (d) RoIAlign [11]

FIGURE 1. Details of the network structure of Mask R-CNN

rectangular region. L_{cls} is the log loss for the two classes (object/non-object) and L_{reg} is the regression loss, expressed as $L_{reg}(t_i, t_i^*) = R(t_i - t_i^*)$ using the robust function R. N_{cls} is the size of the mini-batch, and N_{reg} is the number of anchors, weighted by the balance parameter λ.

2.3. **Head.** The network structure of Head consists of a RoIAlign layer, a segmentation layer, a class identification layer, and a region extraction layer. RoIAlign is a method to improve pixel misalignment, which was a problem with RoIPooling in Faster R-CNN. Figure 1(d) shows RoIAlign. RoIAlign does not just thin out the feature map but creates a fixed-size vector by interpolation to account for sub-pixel-level information. The branch estimating the mask from the RoI feature vector consists of a small FCN (Fully Convolutional Network), which estimates an $m \times m$ object region mask by convolution layer. Assuming that there are K types of classification objects, binary classification is performed for each of the K classes, resulting in a $K \times m \times m$ output for the mask estimation. In this case, the threshold value θ_{out} determines whether the object is regarded as an object or not. The loss used for training is as follows: $L = L_{CLS} + L_{BOX} + L_{MASK}$. Apply a per-pixel sigmoid and define L_{MASK} as the averaged binary cross-entropy loss.

3. Estimation of Sediment Disaster Prone Areas Using Mask R-CNN.

3.1. **Sediment disaster prone areas.** In Japan, landslide hazard warning areas including SDP Areas are lands that are recognized as having the potential to endanger the lives or bodies of residents, etc., in the event of a steep slope collapse, etc. The area is designated based on the "Law Concerning Promotion of Sediment Disaster Countermeasures in Sediment Disaster Precaution Areas, etc.", which came into effect on April 1, 2001 in Japan.

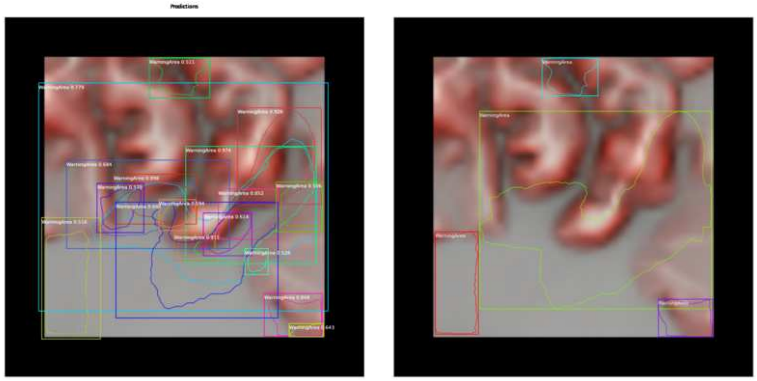
3.2. **Proposed methods.** In the past studies, Pix2pix and traditional CNN have been applied. All the models perform pixel-by-pixel detection. On the other hand, our proposed model is capable of OD as well as pixel-by-pixel detection through IS. Since the detection area is displayed as a large object, it is easy to grasp the detection area.

3.2.1. *Mask image generation.* The mask image for estimating SDP Areas represents the alert areas. The data of SDP Areas are provided in shapefile format at the digital national land information download site [16]. Normally, in OD, the mask information is extracted by manually creating a json file with a tool. In the implementation used in this study, the mask information is extracted from the binarized mask image. The binary raster images are generated by converting the vector format image obtained from the shapefile formatted file. Mask images can be easily generated from shapefile format files without manual processing. During training, all polygons drawn with a Boolean value of 1 are extracted from the mask image. Let N be the number of polygons extracted and p_n be the nth polygon ($1 \leq n \leq N$) extracted. For each $n \in \{1, 2, \ldots, N\}$, a mask image is generated from polygons p_n and a list consisting of N class IDs is simultaneously generated.

3.2.2. *Input images.* Two selected from four types of images are synthesized as multi-channel image and the synthesized image is used as an input image. The images used for the synthesis should contain the features of steep slopes and human activity areas so that the features of the SDP Areas can be extracted. We consider to use the four types of images: RRIM, aerial photograph, standard maps and land-use map. All of them are color images with 3-channel. RRIM can be considered to contain the feature of steep slopes and the rest of three images can be considered to contain the feature of human activity areas. RRIM developed by Asia Air Survey Co., Ltd. [4] is a 3D visualization method that represents slope and ridge-valley by using saturation and brightness of red color. RRIM images shall be used for features of steep slopes. Aerial photographs, standard maps, and land-use maps shall be used for features of human activity areas, which means that the areas may endanger the lives or bodies of residents. Aerial photographs, standard maps, and land-use maps, in that order, are considered to better represent the characteristics of human activity areas. The land-use map is created from the land-use subdivision mesh data [16] of the digital national land information. The land-use subdivision mesh data is data that shows land-use conditions expressed by 11 types of items judged from map symbols and satellite image color tones for each 100 m (1/10 subdivision) mesh unit, of which 6 types of land: rice fields, other agricultural land, building use, roads, railroads, and other sites are shown as red color areas. Since only RRIM can be considered to contain the future of steep slopes, we propose to use RRIM and one selected from the other three types of images, that is, three combinations, "RRIM and aerial photograph", "RRIM and standard map", and "RRIM and land-use map", are considered. For any combination, the synthesized image can be considered to contain both of features of steep slopes and human activity areas, and the number of channels of the synthesized image is six for any combination.

3.2.3. *Integration of detected areas by logical OR.* Bounding boxes for detected areas with IoU (Intersection over Union) greater than 0.7 are excluded as duplicates. However, as shown in Figure 2(a), the remaining bounding boxes still have some overlap. This rarely happens with a single object with clear boundaries, such as a vehicle or a person. In contrast, since SDP Areas generally have unclear boundaries, a single SDP Area can be easily detected as multiple areas. This characteristic reduces the advantage of OD in visibility. To address this problem, we propose a method to integrate multiple overlapping detection areas by logical OR operations, as shown in Figure 2(b). The algorithm is as follows.

Step 1: Let M be the number of mask images detected for SDP Areas. Let $X_m = (x_{m,i,j})$ be the mth mask image ($1 \leq m \leq M$), where each pixel $x_{m,i,j}$ takes the logic value 1 if it is in the SDP Area, otherwise it takes the logic value 0. Then integrate M mask images into a single image $X = (x_{i,j})$ by letting $x_{i,j} = x_{1,i,j} + x_{2,i,j} + \cdots + x_{M,i,j}$.
Step 2: Extract from image X all polygons drawn with logic value 1. Let M' be the number of the extracted polygons. Let p_m be the mth extracted polygon ($1 \leq m \leq M'$). For each $m \in \{1, 2, \ldots, M'\}$, generate the mask image X'_m from the polygon p_m, where X'_m represents the mask and bounding box information for the mth detected area integrated by logical OR operations.



(a) Raw prediction (b) OR prediction

FIGURE 2. Integration of predictions by logical OR

4. Experimental Evaluation.

4.1. Evaluation indices.

The reproducibility, fit rate, and $F1$-*score* calculated by Equation (1) are used as Evaluation Indices (E.I.) for the simulation results. TP is the number of correctly predicted SDP Areas as SDP Areas, TN is the number of correctly predicted backgrounds as backgrounds, FP is the number of incorrectly predicted backgrounds as SDP Areas, and FN is the number of incorrectly predicted SDP Areas as backgrounds. *Recall* is the percentage of correctly predicted SDP Areas out of those that are actually SDP Areas. *Precision* is the percentage of the predicted SDP Areas that are actually SDP Areas. The $F1$-*score* is the harmonic mean of *recall* and *precision*. TP, TN, FP and FN are calculated in two ways. The first way counts the pixel-by-pixel matches between the mask image and the prediction. The second way is to determine whether each SDP Area is correctly predicted or not and to treat a detection as correct if the bounding box of the detected area overlaps with the bounding box of true area by more than a certain threshold $\theta = 0.5$. The percentage of this overlapping area is calculated by Equation (2) by obtaining the x and y coordinates of the center of each area and the width and height of the area of T, w_T and h_T, respectively, with T and P as the bounding boxes of the correct answer and prediction, respectively. $x_{T,\max}$ and $x_{T,\min}$ denote the maximum and minimum values of the x-coordinate of T, respectively; the same is true for P and y.

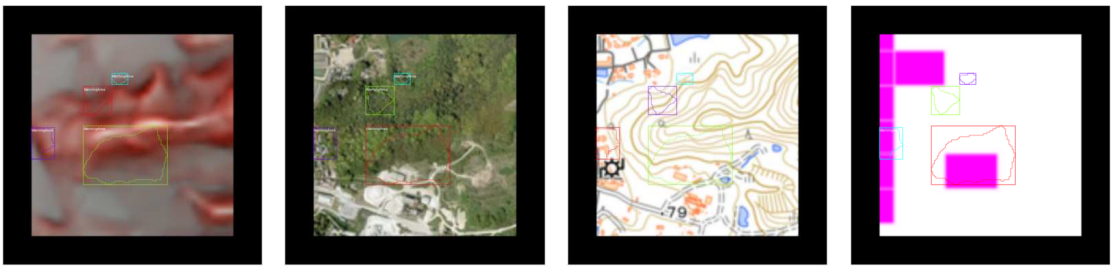
$$recall = \frac{TP}{TP + FN}, \quad precision = \frac{TP}{TP + FP}, \quad F1\text{-}score = \frac{2 \cdot recall \cdot precision}{recall + precision} \tag{1}$$

$$DR = \frac{T \cap P}{T} = \frac{dxdy}{w_T h_T} \tag{2}$$

$$dx = \min(x_{T,\max}, x_{P,\max}) - \max(x_{T,\min}, x_{P,\min}) \tag{3}$$

$$dy = \min(y_{T,\max}, y_{P,\max}) - \max(y_{T,\min}, y_{P,\min}) \tag{4}$$

4.2. **Simulation conditions.** The data used in the experiment consist of 9022 pairs of RRIM images, aerial photographs, standard map images, land-use maps, and binarized mask images with a zoom level of 16 and a size of 128×128 patch images. These patch images are created by sampling from northern Fukuoka Prefecture in Japan, where latitudes range from 33.3333 to 33.9999 degrees and longitudes range from 130.0000 to 131.0000 degrees. RRIM images are from Red Relief Image Map RRIM®5+ created by Asia Air Survey Corporation, aerial photographs and standard maps are from Geographical Survey Institute tiles, and land-use maps and SDP Areas are from land-use subdivision mesh data and landslide hazard warning area data from the digital national land information data of the National Land Policy Bureau, Ministry of Land, Infrastructure and Transport. Figure 3 shows an example of a training image with mask information notated. These 9022 pairs of images are divided into training, validation, and testing, with 7222, 900, and 900 images, respectively. The learning rate is set to 0.001, the number of epochs is 10, the batch size is 2, and the number of steps per epoch is 3500. The threshold θ_{out} for the probability of being an alert area in the output is set to 0.5.



(a) RRIM (b) Aerial photographs (c) Standard maps (d) Land-use maps

FIGURE 3. Example of training image

In the training of Mask R-CNN, we use transfer learning in order to balance accuracy and processing time. The base model for transfer learning is the model provided in [13] and trained on the COCO dataset [17]. The COCO dataset consists of 80 class labels, which are composed of categories such as people, bicycles, elephants, glass, sky, and road.

4.3. **Simulation 1.** To examine the effectiveness of the transfer learning, a comparison was made between the cases with and without transfer learning. The used data is an RRIM images. For each case, the Mask R-CNN model is trained with the number of epochs of 15, and the accuracy of the trained model is evaluated on test data of 900 images. The accuracy is calculated as the pixel-level accuracy obtained from the mask information.

In the case without transfer learning, *recall* is 77.64% and *precision* is 26.72%. With transfer learning, the *recall* is 72.72% and *precision* is 33.85%. The results suggest that the use of transfer learning results in a higher percentage of *precision* than without transfer learning. $F1$-*score* is 39.76% without transfer learning and 46.20% with transfer learning, thus indicating more accurate collapse hazard areas on steep slopes of the test image when transfer learning is used. Therefore, we use transfer learning in all remaining simulations.

4.4. **Simulation 2.** To investigate effective input image combinations, the following four input image combinations described in Section 3.2.2 are compared: ① Only RRIM images, ② Combination of RRIM images and aerial photographs, ③ Combination of RRIM images and standard maps, and ④ Combination of RRIM images and land-use maps. For each of the four combinations, the mask R-CNN model is trained, and the accuracy of the trained model is evaluated on the test data of 900 images. Two types of accuracy, pixel-level accuracy obtained from mask information and OD accuracy obtained from bounding box overlap, are calculated as the average of three trials. Furthermore, the prediction results

are shown for two sample images, one with and one without steep slopes, and the accuracy is shown for the sample image with steep slopes.

Table 1 shows the evaluation result on test data of 900 test images, Table 1(a) shows the accuracy for the pixel level obtained from the mask information, and Table 1(b) shows the accuracy of OD obtained from bounding boxes overlap. Table 1(a) shows that the highest *recall* is 78.51% for ③. For *precision* and *F*1-*score*, the highest percentages of 31.20% and 43.49% are obtained respectively when ① is used. Table 1(b) also shows that the highest *recall* is 88.86% for ③. For *precision* and *F*1-*score*, ③ is also the highest, at 53.50% and 66.57%, respectively.

Table 1. Evaluation results on test data of 900 images

(a) For mask information

Comb. / E.I.	①	②	③	④
Recall (%)	72.55	70.75	78.51	75.78
Precision (%)	31.20	28.06	27.49	28.06
*F*1-*score* (%)	43.49	40.64	40.18	40.94

(b) For bounding boxes

Comb. / E.I.	①	②	③	④
Recall (%)	80.33	84.64	88.86	88.65
Precision (%)	41.89	45.62	53.50	46.39
*F*1-*score* (%)	54.51	59.26	66.57	60.86

For the four combinations, prediction results on the test sample image with steep slopes are shown in Figure 4, and prediction results on the test sample image without steep slopes are shown in Figure 5. Table 2(a) shows the pixel level accuracy obtained from the prediction results in Figure 4, and Table 2(b) shows the OD accuracy obtained from the prediction results in Figure 5. As can be seen from Figure 4, Table 2(a), and Table 2(b), the combination ③ achieves the highest *F*1-*score*s 63.11% and 66.67% for the pixel level and the OD accuracy, respectively. Figure 5 shows that except for ②, false detection rarely occurs for a test sample image with no steep slopes. This suggests that combinations other than ② are effective. The cause of the false detection in combination ② can be considered to be due to the features obtained from the aerial photographs. Therefore, the use of aerial photography is not considered suitable for the proposed method.

Table 2. Evaluation results on a test sample image shown in Figure 4

(a) For mask information

Comb. / E.I.	①	②	③	④
Recall (%)	90.08	84.75	89.63	93.19
Precision (%)	47.64	45.58	48.70	41.33
*F*1-*score* (%)	62.32	59.28	63.11	57.26

(b) For bounding boxes

Comb. / E.I.	①	②	③	④
Recall (%)	50.0	100.0	100.0	100.0
Precision (%)	100.0	50.00	50.00	40.00
*F*1-*score* (%)	66.67	66.67	66.67	57.14

4.5. **Simulation 3.** For the combination ③, which obtained the best results in Simulation 2, we experimentally investigate the effective threshold θ_{out} at which an area is determined to be an alert area. The threshold θ_{out} is examined for 0.5, 0.6, 0.7, 0.8, and 0.9.

Table 3 shows the evaluation results of the pixel level accuracy for ③, and Table 4 shows the evaluation results of the OD accuracy for ③. Tables 3 and 4 show that, as the threshold increases, *recall* decreases and *precision* increases. In Table 4, with a threshold $\theta_{\text{out}} = 0.5$, *recall* is already close to 90% and *precision* is close to 50%, which means that almost no SDP Areas are missed and almost as many new alert areas are detected as existing SDP Areas. Therefore, the threshold θ_{out} of 0.5 for the simulation condition is the best among the tested candidates.

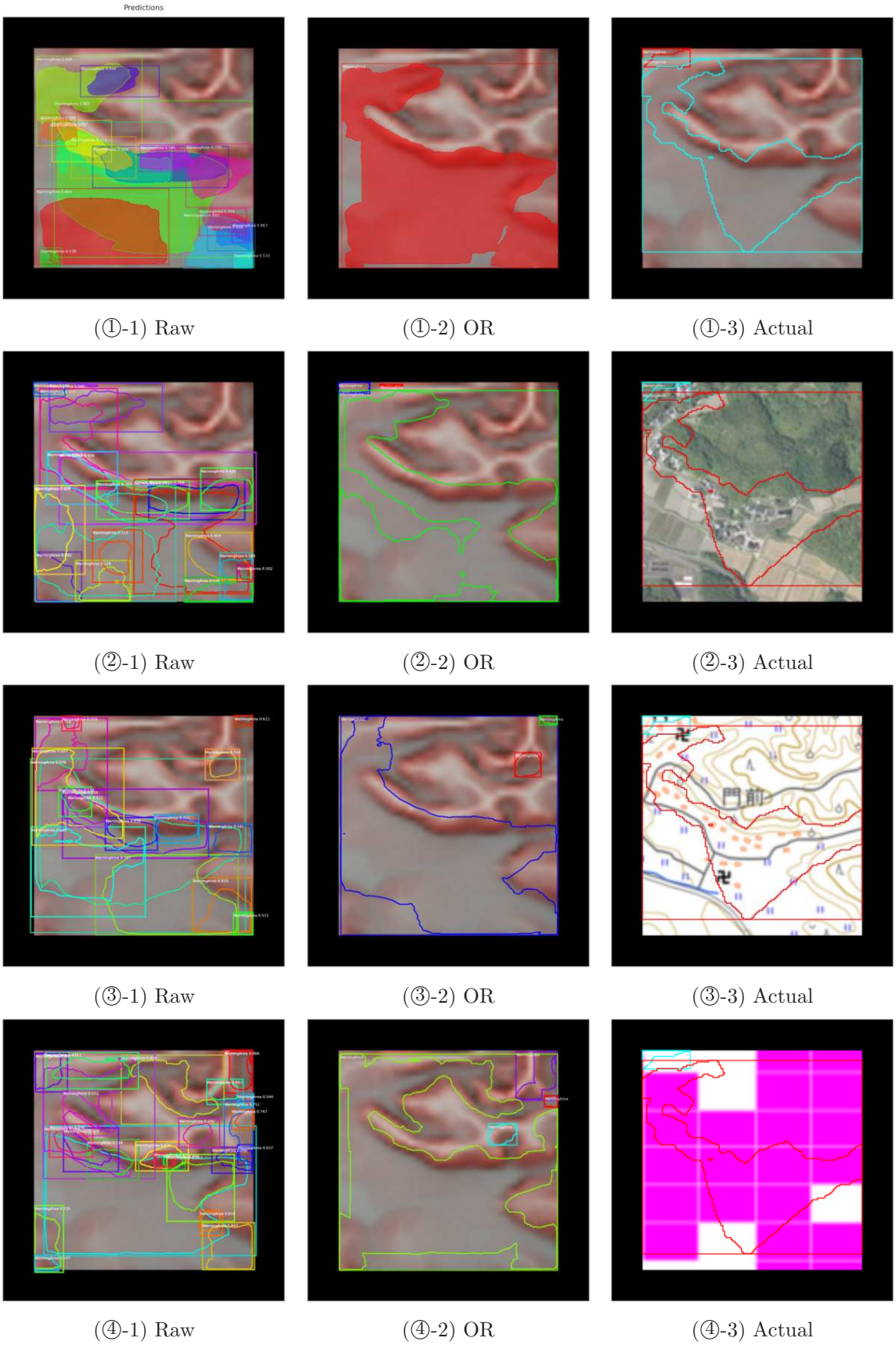


FIGURE 4. Results on a test sample image where steep slopes exist

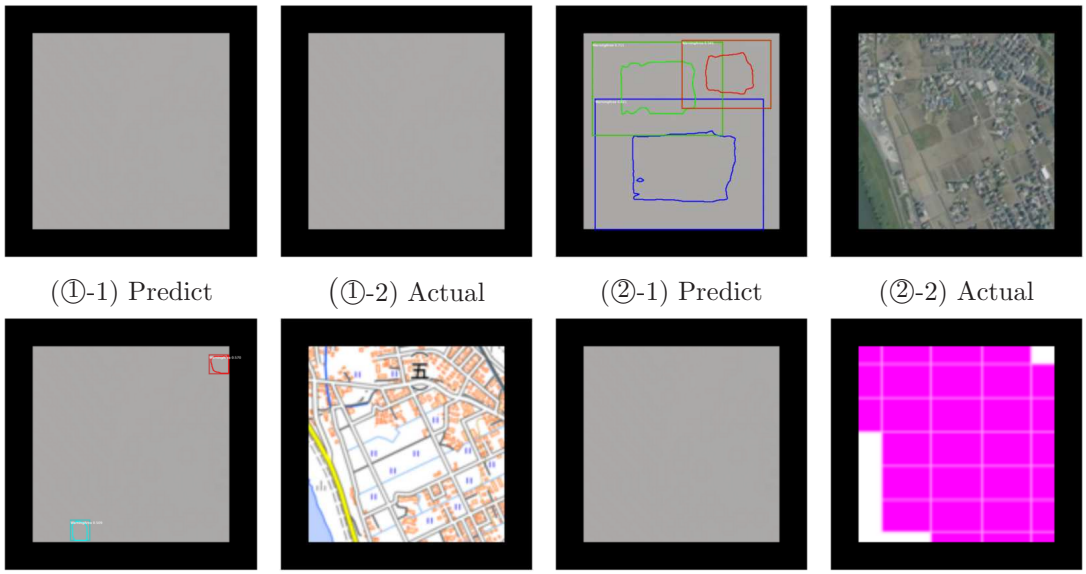


FIGURE 5. Results on test sample image where steep slopes do not exist

TABLE 3. Evaluation results of ③ with different thresholds for mask information

E.I. \ Threshold	0.5	0.6	0.7	0.8	0.9
Recall (%)	81.92	79.29	76.76	73.19	66.53
Precision (%)	28.73	30.90	33.37	36.37	41.36
*F*1-*score* (%)	42.16	44.47	46.52	48.59	51.01

TABLE 4. Evaluation results of ③ with different threshold for bounding boxes

E.I. \ Threshold	0.5	0.6	0.7	0.8	0.9
Recall (%)	91.49	87.52	82.76	75.89	61.16
Precision (%)	46.34	49.85	54.41	60.03	67.98
*F*1-*score* (%)	61.52	63.52	65.66	67.03	64.39

5. **Conclusion.** In this study, we examined an effective method for estimating SDP Areas from map images such as RRIM images using Mask R-CNN. We studied four different combinations of RRIM images, aerial photographs, standard maps, and land-use maps, and examined which combination was best, as well as a method of integrating predicts for visual clarity. Simulations confirm that looking at bounding boxes improves the fit rate to nearly 30% at the pixel level and nearly 50% at the bounding box, while maintaining the repeatability rate near 90%. In the pixel-by-pixel results, ③ had the highest *recall*, 78.51%, with *precision* at 27.49% and *F*1-*score* at 40.18%. In [3], the method based on Pix2pix has been proposed, and its best *F*1-*score*, *recall*, and *precision* were 17.01%, 49.67%, and 10.27%, respectively. Our proposed Mask R-CNN model outperforms the Pix2pix model. From the predicted images, it was observed that the combination of RRIM images and aerial photographs resulted in the predicted for areas where no steep slopes exist, while the other three methods almost never resulted in the predicted for areas where no steep slopes exist, indicating the effectiveness of the combination of these methods. From a practical standpoint, future issues include confirming whether the same accuracy can be obtained for wide-area images at lower zoom levels.

REFERENCES

[1] Ministry of Land, Infrastructure, Transport and Tourism, *Designation of Sediment Disaster Alert Areas*, https://www.mlit.go.jp/mizukokudo/sabo/linksinpou.html, Accessed on 7 November, 2022 (in Japanese).

[2] Y. Yamashita, N. Shigei, S. Sugimoto, R. Takaesu and Y. Ishizuka, Estimation of landslide hazard areas from 3D maps and aerial photographs using CNN [Translated from Japanese], *Proc. of Japan Society of Civil Engineers-West Annual Meeting*, pp.339-340, 2019 (in Japanese).

[3] Y. Nishi and N. Shigei, Estimation methods of slope failure hazard areas by using Pix2pix, *"Hinokuni – Land of Fire" Information Processing Symposium*, A11-2, 2022 (in Japanese).

[4] Asia Air Survey Co., Ltd., *RRIM*, https://www.rrim.jp/en/, Accessed on 30 April, 2023.

[5] T. Komiyama, K. Hotta, K. Oda and S. Kakuta, Semantic segmentation in red relief image map by UX-Net, *Proc. of the 7th International Conference on Pattern Recognition Applications and Methods*, pp.597-602, 2018.

[6] O. Ronneberger, P. Fischer and T. Brox, U-Net: Convolutional networks for biomedical image segmentation, *arXiv Preprint*, arXiv: 1505.04597, 2015.

[7] S. Kubo, T. Yamane and P.-J. Chun, Study on accuracy improvement of slope failure region detection using Mask R-CNN with augmentation method, *Sensors*, vol.22, 6412, 2022.

[8] S. L. Ullo, A. Mohan, A. Sebastianelli, S. E. Ahamed, B. Kumar, R. Dwivedi and G. R. Sinha, A new Mask R-CNN-based method for improved landslide detection, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp.3799-3810, 2021.

[9] R. Yang, F. Zhang, J. Xia and C. Wu, Landslide extraction using Mask R-CNN with background-enhancement method, *Remote Sens.*, vol.14, 2206, 2022.

[10] T. Kikuchi, K. Sakita, S. Nishiyama and K. Takahashi, Landslide susceptibility mapping using automatically constructed CNN architectures with pre-slide topographic DEM of deep-seated catastrophic landslides caused by Typhoon Talas, *Natural Hazards*, vol.117, pp.339-364, 2023.

[11] K. He, G. Gkioxari, P. Dollár and R. Girshick, Mask R-CNN, *arXiv Preprint*, arXiv: 1703.06870, 2017.

[12] S. Ren, K. He, R. Girshick and J. Sun, Faster R-CNN: Towards teal-time object detection with region proposal networks, *arXiv Preprint*, arXiv: 1506.01497, 2015.

[13] W. Abdulla, Mask R-CNN for object detection and segmentation on Keras and TensorFlow, *GitHub Repository*, https://github.com/matterport/Mask_RCNN, Accessed on 21 October, 2022.

[14] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016.

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, Feature pyramid networks for object detection, *arXiv Preprint*, arXiv: 1612.03144, 2017.

[16] Ministry of Land, Infrastrucutre, Transport and Tourism, *Digital National Land Information Download Site* [Translated from Japanese], https://nlftp.mlit.go.jp/ksj/, Accessed on 30 April, 2023 (in Japanese).

[17] T.-Y. Lin et al., Microsoft COCO: Common objects in context, *arXiv Preprint*, arXiv: 1405.0312, 2015.