SEMANTIC SEGMENTATION AND CHANGE DETECTION BY MULTI-TASK U-NET

Shungo Tsutsui[1,2]*, Tsubasa Hirakawa*[3]*, Takayoshi Yamashita*[3]*, Hironobu Fujiyoshi*[3]

[1] Department of Computer Science, Graduate School of Engineering,
Chubu University, Kasugai, Aichi, Japan
[2] Asia Air Survey Co., Ltd., Kawasaki, Kanagawa, Japan
[3] Center for Mathematical Science and Artificial Intelligence, Chubu University, Kasugai, Aichi, Japan

ABSTRACT

Change detection involves extracting the changed regions from images taken of the same place at different times. Potential applications are automatically updating of HD maps or identifying damages caused by natural disasters. However, conventional change detection methods merely detect changed regions without classifying them. In this paper, we propose a change detection method that can estimate the object class of a changed region. Our method extends a U-Net as a multi-task learning framework and estimates changed regions and semantic segmentation simultaneously. We propose using the pixel-wise classification probabilities of semantic segmentation for detecting changed regions rather than the conventional L2 norm-based difference of feature maps. In our experiments, we show that our method can improve change detection performance and estimate the classes of corresponding changed objects.

Index Terms— Change Detection, Semantic Segmentation, Multi-task Learning

1. INTRODUCTION

Change detection is an important task in the field of computer vision, which has been widely investigated due to its potential for various applications such as automatically updating high definition (HD) maps and identifying damage situations caused by natural disasters. We input two images of the same place at different times for a change detection model to estimate the changed regions. In identifying damage situations after natural disasters, to restore facilities and buildings damaged by natural disasters, it is necessary to quickly identify the damaged areas. The recent development of in-vehicle cameras and drones equipped with cameras have enabled us to take wide range of images, making it easier to collect large-scale data in an urban setting. This makes it easy to identify the affected area and can respond quickly, but it takes human cost to select the necessary data from a huge amount of data.

Previously proposed methods [1, 2, 3, 4, 5] provide detection results as class labels or confidence values pixel-wisely. The aforementioned application requires class information of changed region. It is necessary to categorize changed objects or regions to automatically update HD maps. Also, providing the changed object category can further accelerate the quick identification of damaged regions after natural disaster. Therefore, understanding the semantics of the changed regions is expected to be beneficial for these applications.

In this paper, we propose a change detection method that can estimate the object class of a changed region. The proposed method consists of a U-Net that has been expanded to a multi-task learning framework. The decoder of the proposed method is split into branches for change detection and semantic segmentation tasks. Hence, we can obtain change detection and semantic segmentation results simultaneously. The proposed network is trained by using a loss function calculated by the weighted losses of change detection and semantic segmentation. To attain more robust change detection for environmental changes such as changes in viewpoint and lighting, we evaluate the accuracy of two detection methods based on (i) the differences of feature maps and (ii) the semantic segmentation probabilities. Moreover, to train and evaluate the proposed method, we build semantic segmentation labels for a panoramic change detection (PCD) dataset [6]. The experimental results show that our network can improve change detection performances.

2. RELATED WORK

In this section, we briefly introduce the conventional change detection methods.

2.1. CNN features and superpixel segmentation based change detection

Sakurada *et al.* [6] proposed a CNN-based change detection method that is based on a VGG16 network [7]. This method consists of the following three steps. In the first step, images are split into grid cells and each cell is input into a VGG16 network trained by ImageNet dataset [8]. The use of split grid cells can make the detection more robust to small displacements between a pair of images. Then, the dissimilarity for each grid is computed by using features extracted by the

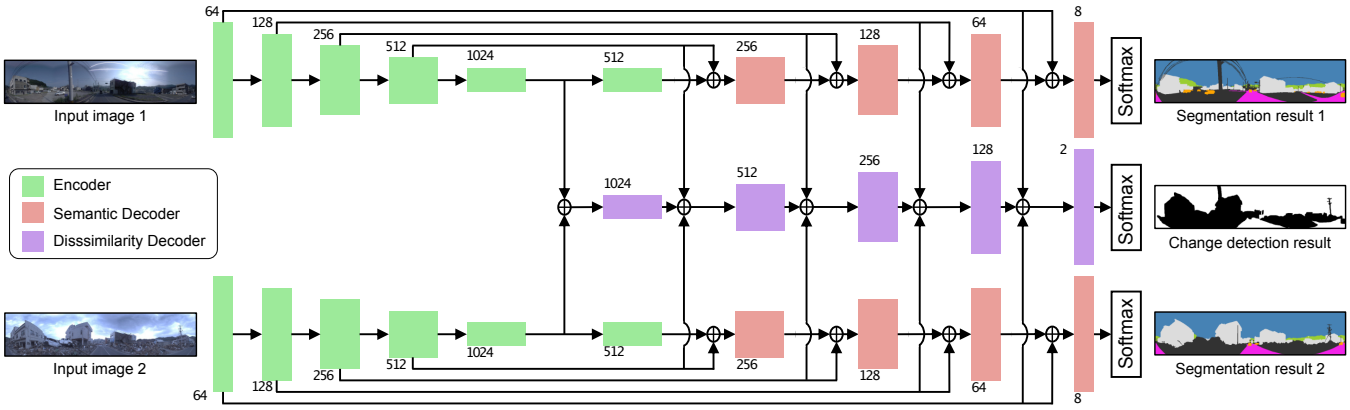

Fig. 1. Network structures of proposed method with classification probabilities of semantic segmentation

VGG16 network. In the second step, a superpixel segmentation [9] is introduced in order to detect the changed regions precisely. The superpixels are estimated for two input images and the dissimilarities of each superpixel are computed by using the grid dissimilarities. Then, a dissimilarity map is computed from the dissimilarities obtained for each superpixel. In the final step, the change detection results derived from the images of the sky and ground regions, e.g. clouds or lighting changes, are removed. A scene parsing method is used to estimate the sky and ground regions [10].

2.2. SSCDNet

The previously developed silhouette-based semantic change detection network (SSCDNet) [11] is largely similar to our proposed method. The SSCDNet is a method for simultaneously detecting changed regions and estimating the class of the changed object. This method consists of two networks: a correlated Siamese change detection network (CSCDNet) that detects changed region from an image pair, and an SSCDNet, which uses the change detection results obtained from the CSCDNet and estimates the object class of the detected region. These networks are based on a U-Net architecture [12, 13]. This method trains a CSCDNet to detect a changed region and uses the trained CSCDNet to obtain a change probability mask. The change probability mask is created from the change probability at each pixel. Then, the image pair and the change probability mask are input into the SSCDNet and the class of the changed region is estimated. The SSCDNet simultaneously trains the relationship between input images, change probability, and semantic segmentation, enabling us to provide the change detection region with the corresponding object class.

3. PROPOSED METHOD

Herein, we describe the details of the proposed method.

3.1. Multi-task U-Net

As shown in Fig. 1, our method consists of three modules: an encoder, semantic decoder, and dissimilarity decoder. The architecture of the decoder is the same as that of the conventional U-Net. The semantic decoder outputs the result of semantic segmentation and the dissimilarity decoder detects the changed region. Note that the decoders are independent and do not share network weights. Because two tasks are performed with a single network by a multi-task learning framework, the network is smaller than when using independent networks for each task. The multi-task U-Net inputs a pair of images taken at the same place at the different times. To train the network, we calculate the loss computed by the weighted sum of losses for each task. By training the network with multi-task learning, change detection performances can be improved.

3.2. Semantic Decoder

The semantic decoder outputs segmentation results using the feature map obtained from the encoder. We use cross entropy loss for the loss function of the segmentation task, as with conventional U-Net.

3.3. Dissimilarity Decoder

The dissimilarity decoder detects the changed regions from a pair of input images. The decoder architecture is similar to the aforementioned Semantic Decoder. We input feature maps obtained from the encoder. To recognize the boundary between objects, we concatenate the feature maps obtained from the encoder that corresponds in size to the decoder. The output feature map is fed into the softmax layer to obtain the pixel-wise probability of change detection.

3.4. Training

We train the network in an end-to-end manner. The loss function for training is computed by adding the losses of the semantic decoder L_{Seg} and the dissimilarity decoder L_{Diff}, which is defined as follows:

$$L = w_{Seg} \cdot L_{Seg} + w_{Diff} \cdot L_{Diff}, \quad (1)$$

where w_{Seg} and w_{Diff} are scale parameters for balancing loss and gradients for updating network parameters, respectively. These parameters are decided manually, and we evaluate the segmentation and change detection accuracies over different scale parameters in our experiment.

4. EXPERIMENT

In this section, we evaluate the performance of the proposed method in terms of change detection and semantic segmentation. We also evaluate the performances by changing the scale parameters of the loss function.

4.1. Dataset

We use a panoramic change detection (PCD) dataset [6]. The PCD dataset consists of image pairs taken of a place damaged by a tsunami in Japan. The dataset contains panoramic image pairs and the ground truth indicating the changed region as a binary mask. The number of samples (image pairs) are 100. However, the PCD dataset is only constructed for change detection tasks; there is no ground truth for semantic segmentation tasks, so we annotated segmentation labels. We annotated images with the following eight classes: sky, road, car, building, plant, bicycle, person, and other.

4.2. Experimental Settings

In the PCD dataset, we use 180 images (90 image pairs) for training and 20 images (10 image pairs) for evaluation. Due to the lack of training samples, we apply data augmentations such as resize, image shift to horizontal axis, smoothing, noise addition, flipping, contrast, and gamma conversions, and we set the number of iterations to 100,000. Also, the conditions are unified for all networks. However, CNN-feat, which is one of the comparison methods, uses the trained model of VGG16 and does not perform new training. In addition, U-Net, which is used to delete erroneous detection results, has the same conditions as other networks.

For evaluation metrics, we used the F1 score for the change detection task, and global accuracy, class accuracy, and mean IoU for the semantic segmentation task. To compare the performance of the proposed method, we used the following methods.

CNN-feat [6]: This method originally used geometric context [10] to remove the change detection results of sky and road.

Table 1. F1 scores of proposed method with different scale parameters

Params.		Proposed method	
w_{Seg}	w_{Diff}	Diff.	Prob.
0.6	0.4	0.781	0.862
0.7	0.3	**0.785**	**0.867**
0.8	0.2	0.776	0.862

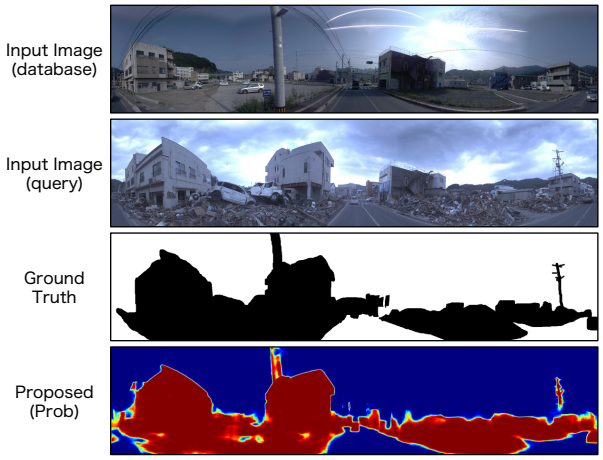

Fig. 2. Results of change detection. (best viewed in color)

In our experiments, we substitute the results obtained from the U-Net and remove incorrect detection results.

CSCDNet [11]: This method consists of CSCDNet for detecting change regions and SSCDNet for estimating the object class of the changed regions. We compare the change detection accuracy of CSCDNet.

ChangeNet [14]: This method is a change detection method based on the transfer learning approach. In our experiment, since we use a dataset that detects only the changed region, we treat it as a binary classification.

U-Net [12]: We use U-Net to compare semantic segmentation and change detection. We prepare and train the U-Net for each task independently. For the change detection task, we input a pair of images by concatenating images channel-wisely and output the change detection result.

Proposed: **Diff.** indicates the method with the L2 norm-based difference of feature map, and **Prob.** is the method with class probability of semantic segmentation. For Diff., we need to determine a threshold to decide the changed region. Hence, we investigate an optimal threshold by changing the threshold value from 0.1 to 0.9 at intervals of 0.1.

4.3. Results for different w_{Diff} and w_{Seg}

We first evaluate the performance of the proposed method by changing scale parameters w_{Diff} and w_{Seg}. Table 1 shows the F1 scores with different parameters. In Diff., the loss of

Table 2. Accuracies of change detection and semantic segmentation

Task	Metric	CNN-feat	CSCDNet	ChangeNet	U-Net		Proposed	
					Seg	Diff	Diff	Prob
Change detection	F1 score	0.723	0.859	0.82	0.585	0.833	0.794 ± 0.037	**0.862** ± 0.003
Semantic segmentation	Global accuracy	-	-	-	**92.06**	-	90.15 ± 0.09	90.79 ± 0.07
	Class accuracy	-	-	-	**68.71**	-	65.84 ± 0.34	65.83 ± 0.28
	Mean IoU	-	-	-	**59.93**	-	56.17 ± 0.22	56.81 ± 0.17

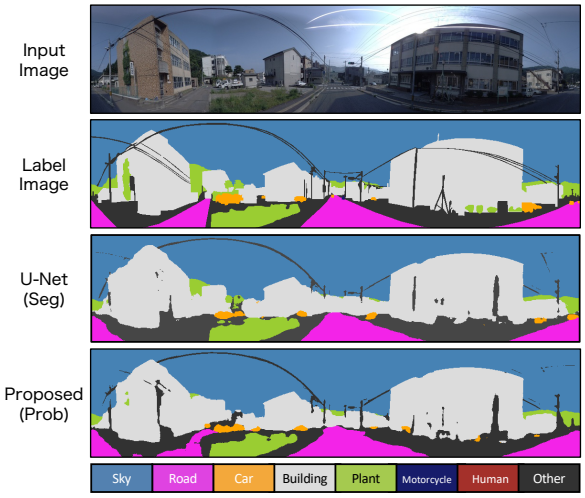

Fig. 3. Results of semantic segmentation. (best viewed in color)

change detection tends to become larger than that of the segmentation task. In case of Prob., the change detection task can be assumed as a binary segmentation problem, which is easier than an 8-class segmentation task. Because a large w_{Diff} decreases the segmentation performance, we do not use a w_{Diff} larger than 0.5. As shown in Table 1, the results of $w_{Seg} = 0.7$ and $w_{Prob} = 0.3$ were highest for both Diff. and Seg. In the following experiments, we use these values to train the proposed method.

4.4. Comparison of change detection results

Table 2 shows the scores of the change detection and semantic segmentation tasks using each method. U-Net (Seg) represents the result with a U-Net trained for semantic segmentation tasks, and U-Net (Diff) represents change detection tasks. To detect changed regions with U-Net (Seg), we split feature maps obtained from U-Net into grids and compute the differences for each grid, which is similar to CNN-feat.

First, we evaluate the two proposed methods. The result of the proposed (Prob) is higher than that of the proposed (Diff). Our method (Proposed (Prob)) achieved the highest result. By extending U-Net to a multi-task learning framework, the proposed method can acquire efficient features at the encoder. Although our method does not share decoder parameters, the change detection accuracy of our method is higher than that of U-Net (Diff). The CSCDNet produced similar results to that of the proposed method. The CSCDNet requires two networks for change detection and semantic segmentation, whereas our method can handle both tasks with a single network. Fig. 2 shows the change detection results. The proposed (Prob) successfully detects changed regions around the boundaries of objects.

4.5. Comparison of semantic segmentation results

Next, we evaluate the performance of semantic segmentation. From Table 2, U-Net (Seg), which is only trained for semantic segmentation, achieved the highest accuracy for all metrics. Fig. 3 shows the segmentation results. The proposed method estimates object classes as with U-Net. Although the performance of the proposed method is quantitatively lower than that of U-Net, the proposed method successfully estimates objects such as debris, which are difficult to estimate.

5. CONCLUSION

In this paper, we proposed a method for change detection while estimating semantic segmentation. Our method extends a U-Net architecture to a multi-task learning framework. Furthermore, we annotated semantic segmentation labels for the PCD dataset. In the experiments, our method achieved the highest results in the change detection tasks and successfully detected the boundaries of changed regions. The results show that multi-task learning with semantic segmentation improves the accuracy of change detection. While we used a U-Net as a base network, our method can potentially be applied to other encoder-decoder networks, which we will investigate in a future work. We also intend to further improve change detection and segmentation accuracies by sharing the decoder and the internal features.

6. REFERENCES

[1] A. Taneja, L. Ballan, and M. Pollefeys, "Image based detection of geometric changes in urban environments," in *International Conference on Computer Vision (ICCV)*, 2011, pp. 2336–2343.

[2] Simon Stent, Riccardo Gherardi, Björn Stenger, and Roberto Cipolla, "Detecting change for multi-view, long-term surface inspection.," in *The British Machine Vision Conference (BMVC)*, 2015.

[3] Salman Khan, Xuming He, Fatih Porikli, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri, "Learning deep structured network for weakly supervised change detection," in *The International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 2008–2015.

[4] Sergey Zagoruyko and Nikos Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4353–4361.

[5] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch, "Fully convolutional siamese networks for change detection," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4063–4067.

[6] Ken Sakurada and Takayuki Okatani, "Change detection from a street image pair using cnn features and superpixel segmentation," in *The British Machine Vision Conference (BMVC)*, 2015.

[7] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[8] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[9] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[10] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," in *International Conference on Computer Vision (ICCV)*, 2005, vol. 1, pp. 654–661.

[11] Ken Sakurada, "Weakly supervised silhouette-based semantic change detection," *arXiv preprint, arXiv:1811.11985*, 2018.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[13] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, Alexander Dovzhenko, Olaf Tietz, Cristina Dal Bosco, Sean Walsh, Deniz Saltukoglu, Tuan Leng Tay, Marco Prinz, Klaus Palme, Matias Simons, Ilka Diester, Thomas Brox, and Olaf Ronneberger, "U-net: deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, 2019.

[14] Ashley Varghese, Jayavardhana Gubbi, Akshaya Ramaswamy, and P Balamuralidhar, "Changenet: A deep learning architecture for visual change detection," in *European Conference on Computer Vision (ECCV)*, 2018.