Weakly Supervised Silhouette-based Semantic Change Detection

Ken Sakurada
National Institute of Advanced Industrial Science and Technology (AIST)
k.sakurada@aist.go.jp

Abstract

This paper presents a novel semantic change detection scheme with only weak supervision. A straightforward approach for this task is to train a semantic change detection network directly from a large-scale dataset in an end-to-end manner. However, a specific dataset for this new task, which is usually labor-intensive and time-consuming, becomes indispensable. To avoid this problem, we propose to train this kind of network from existing datasets by dividing this task into change detection and semantic extraction. On the other hand, the difference in camera viewpoints, for example images of the same scene captured from a vehicle-mounted camera at different time points, usually brings a challenge to the change detection task. To address this challenge, we propose a new siamese network structure with the introduction of correlation layer. In addition, we create a publicly available dataset for semantic change detection to evaluate the proposed method. Both the robustness to viewpoint difference in change detection task and the effectiveness for semantic change detection of the proposed networks are verified by the experimental results.

1. Introduction

In the field of computer vision and remote sensing, change detection methods have been comprehensively studied and applied to many kinds of tasks, such as detecting anomaly using surveillance and satellite cameras, inspecting infrastructure [41], managing disaster [34, 35], and automating agriculture [12]. The existing methods specify a few detection targets, such as pedestrians and vehicles, for each application. However, in cases when images contain various kinds of scene changes, more semantic information except for these targets is required for better discrimination in other advanced applications, such as updating city model for autonomous driving [1].

There are several strategies to achieve semantic change detection from an image pair. One of the most straightforward methods is independent detection and classification of changes. However, it requires the estimation of the input image that contains the changes. If training datasets are available, it is possible for end-to-end learning methods to directly estimate semantic changes from an image pair. Nevertheless, it is labor-intensive to create large-scale semantic change detection datasets for each class definition of applications in terms of collecting and labeling images.

In order to overcome these difficulties, we propose a novel semantic change detection scheme with only weak supervision by dividing this task into change detection and semantic extraction. As shown in Fig.1, the proposed method is composed of the two convolutional neural networks (CNNs), a correlated siamese change detection network (CSCDNet) and a silhouette-based semantic change detection network (SSCDNet). To the best of our knowledge, this is the first method to estimate semantic scene changes. Our main contributions are as follows:

- We propose a novel semantic change detection network that can be trained with only weak supervision from existing datasets.
- We propose a novel siamese change detection network which uses correlation layers that can deal with difference in camera viewpoints.
- We create the first publicly available dataset for semantic scene change detection.

![Figure 1. Overview of the proposed method. First, the CSCDNet takes an image pair as input, which is trained using a change detection dataset, and outputs one change probability mask. Thereafter, the input image pair and the estimated change mask are fed into the SSCDNet, which is trained using a dataset synthesized from a semantic image segmentation dataset. Finally, the SSCDNet estimates the pixel-wise semantic labels of each input image.](image-url)
We propose the SSCDNet that can be trained with the dataset synthesized from commonly available semantic image segmentation datasets, such as the Mapillary Vistas dataset [29], to avoid creating a new dataset for semantic change detection. The estimation accuracy of the SSCDNet depends on that of change detection. However, in case of images captured from a vehicle mounted camera at different time points, existing change detection methods suffer from estimation errors due to difference in camera viewpoints. Hence, we propose a novel siamese network architecture with the introduction of correlation layers, named as the CSCDNet. The CSCDNet can deal with difference in camera viewpoints and achieves state of the art performance for the panoramic change detection (PCD) dataset [33]. Additionally, we incorporate the data augmentation for the input change mask in the training step to improve the robustness of the SSCDNet to change detection errors (Sec. 4.2.2). For evaluating the proposed methods, we have created the panoramic semantic change detection (PSCD) dataset in the hopes of accelerating researches in the filed of dynamic scene modeling.

This paper is organized as follows. In Sec. 2 we summarize the related work. Section 3 explains the details of the proposed network and the training method. Section 4 shows the experimental results. Section 5 presents our conclusions.

2. Related Work

Many methods for temporal scene modeling have been proposed. However, most of them focused on detecting changes or estimating the length of time that each part of a scene exists for. Semantic recognition is required for advanced applications based on dynamic modeling, such as autonomous driving and augmented reality. This section explains the reason for the proposal of the semantic change detection method using commonly available semantic image segmentation datasets in terms of existing methods and datasets in the following.

Change Detection

Change detection methods are classified into several categories depending on types of target scene changes and available information. Change detection in 2D (image) domain is the most standard approach, especially for surveillance and satellite cameras [8, 20, 30, 31], which are accurately aligned. A typical approach models the appearance of the scene from a set of images captured at different times, against which a newly captured query image is compared to detect changes [47]. Scene models are usually designed using the images from the same viewpoint to detect target changes while accounting for irrelevant appearance changes such as differences in illumination conditions.

There are studies that formulate the problem in a 3D domain. Schindler et al. proposed the probabilistic temporal inferences model based on the visibility of each 3D point reconstructed from images taken from multiple viewpoints at different times [37]. The work by Matzen et al. [26] is classified into the same category. In terms of application, the work by Taneja et al. [42, 43] and Sakurada et al. [34] might be the closest to our research.

In recent years, significant efforts have been made to change detection using machine learning, especially for deep neural networks (DNNs) [1, 16, 24, 33, 41]. There are mainly two types of formulations, “patch similarity estimation” and “pixel-wise segmentation”, which can be converted to each other. Patch similarity estimation has been studied for not only change detection but also feature, stereo and image matchings [5, 25, 39, 49, 50]. The work by Zagoruyko et al. [49] showed in their experiments that one-stream networks, which take different time images concatenated in channel dimension as input, outperforms two-stream networks such as the siamese network [6], and multi-scale inputs improve the estimation accuracy. Pixel-wise change detection has been further studied in the context of anomaly detection, background subtraction and moving object detection [4, 24, 45].

Recently, to update city model for autonomous driving, several change detection methods using vehicular imagery has been proposed [1, 33, 36]. Sakurada et al. [33] proposed the change detection method that differentiates feature maps extracted from input images using a CNN such as VGG [40], which is trained using large-scale image recognition datasets, and refines the coarse detection results using superpixel segmentation [15]. The work by Alcantarilla et al. [1] tackles the same problem of viewpoint changes between different times using depth map estimated from multi-view images in an end-to-end manner with CNN. For single view setting, the dense optical flow based network also has been proposed [36].

Semantic Change Detection

There are few studies on semantic change detection because most of change detection studies that specify their target domain, such as moving object, forest, and do not explicitly recognize semantic classes of change. The work by Kataoka et al. [22] does not consider the problem of detecting changes and estimating which of the input images contains change objects. Daudt et al. [9] detected land surface changes between satellite images. In the case of land surface change detection of satellite images, unlike scene change detection, it is not necessary to estimate which of the input images contains change objects because the change region between the two images is the same. However, for the scene change detection, the estimation is necessary because scene objects can appear, disappear and move.
Table 1. Details of the datasets used in the experiments. *(The CSCDNet is trained with only image pairs of a scene and their change masks of the PSCD dataset.)*

| Dataset          | TSUNAMI | GSV | Vistas [29] | PSCD (This work) |
|------------------|---------|-----|-------------|------------------|
| Number of images | 100     | 100 | 20,000      | 500              |
| Original size    | 1024 × 224 |   | various     | 1024 × 224       |
| Crop size        | 224 × 224 |   | -           | 224 × 224        |
| Size in training | 256 × 256 | 256 × 256 | 256 × 256 |                 |
| Pared            | ✓       | ✓   | ✓           | ✓                |
| Change mask      | ✓       | ✓   | ✓           | ✓                |
| Semantic label   | -       | ✓   | -           | -                |
| Alignment        | medium  | coarse | -        | coarse           |
| Training target  | CSCDNet | SSCDNet | CSSCDNet | *(CSCDNet)*       |

### Change Detection Dataset

The main reason why there are few studies of semantic change detection is that there are no publicly available datasets for semantic scene change detection. CDnet2014 [48] is one of the largest change detection datasets. It contains videos for background subtraction and moving object detection with various challenges, such as dynamic background, camera jitter, and shadow [4, 45]. In the field of remote sensing, many change detection datasets exist, such as the aerial imagery change detection (AICD) dataset [3], forest change detection dataset [23] and AIST building-level change detection (ABCD) dataset [16]. Basically, in the aforementioned datasets, change label is annotated as a binary value depending on the differences between background and query images.

Furthermore, the PCD [33] dataset contains image pairs with differences in camera viewpoints and one change mask for each image pair on which scene changes of both input images are superimposed as binary values. Thus, the problem has not been considered that estimates which of input images contain change objects.

### 3. Weakly Supervised Silhouette-based Semantic Change Detection

The method proposed in this paper makes semantic change detection dataset for estimating semantic change unnecessary. There are many types of label definitions for semantic image segmentation depending on the applications; for example, ground-level images of indoor and outdoor scenes [2, 29], aerial and satellite images [21, 28]. Additionally, the definition of change (e.g., whether changes of moving objects, display of digital screens, the light of a lamp, transparent barriers, growth of plants, a pool of water, and seasonal changes of vegetation are ignored or not) depends on the application. Thus, there is a large number of combinations of change and semantic definitions. Clearly, it is time-consuming to create semantic change detection datasets for each application. Furthermore, as mentioned above (Sec. [2]), it is necessary to estimate which of the input images contain change objects because the existing change detection datasets do not explicitly contain that information.

To solve these problems, the proposed method includes two CNNs, namely, the CSCDNet and the SSCDNet. This separated architecture enables the method to train the semantic change detection system with change detection datasets and commonly available semantic image segmentation datasets. The rest of this section explains the details regarding the weakly supervised method.

#### 3.1. Overview

Figure 1 shows an overview of the proposed semantic change detection method [1]. First, the CSCDNet takes an image pair as input, which is trained using a change detection dataset and outputs the change probability of each pixel as one change mask image. Subsequently, the input image pair and the estimated change mask are fed into the SSCDNet, which is trained using a dataset synthesized from a semantic image segmentation dataset. Finally, the SSCDNet estimates the pixel-wise semantic labels of each input image. It should be noted here that the SSCDNet can estimate semantic change labels and which of input images contains the change objects simultaneously.

We conjecture that these semantic label estimations and splitting the change mask into the input images can be trained using a commonly available semantic image segmentation dataset, such as the Mapillary Vistas dataset [29], and that semantic information can improve the accuracy of the change mask estimation. Table 1 shows the details of the datasets used in this paper. The experimental results show the effectiveness of this strategy (Sec. [4]). The details of the training dataset synthesis and the network architectures are explained in the following subsections.

#### 3.2. Dataset Synthesis from Semantic Segmentation Dataset

Here, we consider the problem of estimating pixel-wise semantic change labels of each input image from an im-

---

1 The source code of our method will be publicly available.
Figure 3. Network architectures of the CSCDNet, the SSCDNet, and the CSSCDNet. The architecture of the CSSCDNet is based on the CSCDNet and its output layer is replaced with that of the SSCDNet.

3.3. Network Architecture

Correlated Siamese Change Detection Network (CSCDNet)

We propose the CSCDNet to overcome the limitation of the camera viewpoints of the previous methods. Figure 3 shows the overview of the network architecture of the proposed method (see the supplementary material for the detail). As mentioned in Sec.2, Sakurada et al. [33] found that the comparison between feature maps extracted from input images using a CNN trained with large-scale image recognition datasets [36] is effective for scene change detection task. To incorporate this advantage, we chose the siamese network architecture based on the ResNet-18 [19] which was pretrained on the ImageNet [11] dataset as the encoder of the CSCDNet. Each feature map extracted from two input images in the encoder is concatenated with each pixel value is normalized in the final convolution layer, the feature maps are evaluated by imposing the randomly sampled semantic labels as binary silhouettes $M$.

Initial weights pretrained with ImageNet

Silhouette-based Semantic Change Detection Network (SSCDNet)

age pair and the change mask. There are several possible methods of generating training datasets to solve this problem. A simulator using a photorealistic rendering, such as Virtual KITTI [17], SYNTHIA [32] and SceneNet RGB-D [27] datasets, is one solution. Although photorealistic images might be effective for pre-training, fine-tuning is necessary to address the domain gaps between synthetic and real images. To bridge the gap, Shrivastava et al. proposed the method to learn a model to improve the realism of a simulator’s output using unlabeled real data [38], however, it is difficult to directly apply this method to natural scene images, which are more complicated than their target domains. Alternatively, synthesis using real images can be applied. Dwibedi et al. proposed the synthetic method to generate large annotated instance datasets in a cut and paste manner [14]. Their study might be the closest to our method.

Figure 2 shows an overview of the proposed training dataset synthesis for the SSCDNet from a semantic image segmentation dataset. First, two RGB images $I_1, I_2$ and their semantic label images $L_1, L_2$ are randomly sampled from the semantic image segmentation dataset. Thereafter, the change semantic label images $L_1’, L_2’$ are generated by sampling $n$ semantic labels randomly and removing the others from each semantic label image ($1 \leq n \leq \min(n_{\text{max}}, N_i)$). $N_i$ represents the number of the classes that the semantic label image $L_i$ contains. The maximum number of class samplings $n_{\text{max}}$ should be decided depending on the number of classes of the semantic segmentation dataset. Finally, the change mask is generated by superimposing the randomly sampled semantic labels as binary silhouettes $M$. 

Figure 2 shows an overview of the proposed training dataset synthesis for the SSCDNet from a semantic image segmentation dataset. First, two RGB images $I_1, I_2$ and their semantic label images $L_1, L_2$ are randomly sampled from the semantic image segmentation dataset. Thereafter, the change semantic label images $L_1’, L_2’$ are generated by sampling $n$ semantic labels randomly and removing the others from each semantic label image ($1 \leq n \leq \min(n_{\text{max}}, N_i)$). $N_i$ represents the number of the classes that the semantic label image $L_i$ contains. The maximum number of class samplings $n_{\text{max}}$ should be decided depending on the number of classes of the semantic segmentation dataset. Finally, the change mask is generated by superimposing the randomly sampled semantic labels as binary silhouettes $M$. 

Note: CSCDNet
by the following pixel-wise binary cross entropy loss:
\[
L_c = -\sum_{x} t(x) \ln(p_c(x)) + (1-t(x)) \ln(1-p_c(x)),
\]
where \(t(x)\) represents the ground truth, \(p_c(x)\) represents predictions computed using each output feature maps by a pixel-wise softmax.

Silhouette-based Semantic Change Detection Network (SSCDNet)

The architecture of the SSCDNet is based on the combination of U-Net based on ResNet-18 \cite{18, 19}. Their main differences are the input and output parts. The SSCDNet takes images \(I_1, I_2\) and \(M\), which are concatenated in the channel dimension as a seven-channel image, for the input. Moreover, after the final convolution layer, the output feature maps are split in half (the bottom of Fig.3), and each of the feature maps is evaluated by the following pixel-wise cross entropy loss:
\[
L_s = -\sum_{x} \sum_{k} t_1(x,k) \ln(p_1(x,k)) + t_2(x,k) \ln(p_2(x,k)),
\]
where \(k\) is an index of classes (1 \(\leq k \leq K\); the number of classes), \(t(x,k)\) represents the ground truth with 1-of-\(K\) coding scheme, \(p(x,k)\) represents predictions computed from each output feature maps by a pixel-wise softmax.

Correlated Siamese Semantic Change Detection Network (CSSCDNet)

For a comparative study, we proposed the CSSCDNet as a naive method in the case that the semantic change detection dataset is available. The architecture is based on the CSCDNet. After the final convolution layer, the output feature maps are split in half, and each of the feature maps is evaluated by the pixel-wise cross entropy loss in the same manner as the SSCDNet (in the dash line box of Fig.3).

4. Experiments

To evaluate the effectiveness of our approach, we performed three experiments. The first experiment is the accuracy evaluation of the change detection with the CSCDNet, which affects the prediction accuracy of the SSCDNet, using the PCD dataset \cite{33}. The proposed siamese change detection networks with and without correlation layers and other existing methods are compared. The second experiment is an accuracy evaluation of the semantic change detection with the SSCDNet using datasets synthesized from the Mapillary Vistas dataset \cite{29}. The data augmentation of change mask is also evaluated, which improves the robustness of the SSCDNet to change detection error of the CSCDNet. In the final experiment, we applied our semantic change detection method to the PSCD dataset, which is different from the training dataset of the SSCDNet, and show the effectiveness of our approach.

4.1. Panoramic Semantic Change Detection (PSCD) dataset

For the quantitative evaluation of the proposed approach, we have created a new dataset named as the PSCD dataset, which opens up new vista for semantic change detection. Figure 4 shows examples of the PSCD dataset. The PSCD dataset comprises 500 panoramic image pairs. Each pair consists of images \(I_1, I_2\) taken at two different time points \(t_1\) and \(t_2\). These panoramic images, which are taken in urban and tsunami-damaged areas, are downloaded from Google Street View.

The PSCD dataset contains the change binary masks \(C_1, C_2\), the semantic labels \(S_1, S_2\), the instance labels \(D_1, D_2\), the attributes \(A_1, A_2\) (3D object, 2D texture, (digital) display) and the visibilities \(V_1, V_2\) (glass, mirror, and wire fence). The image annotation was performed by a team of 37 well-trained image annotators, and the average annotation time was approximately 156 minutes per image pair.

We defined the 67 semantic classes based on those of the Mapillary Vistas dataset \cite{29}, and integrated the original classes into the 11 classes based on the map updating applications as shown in Tables 3 and 4. The annotation data and the metadata for downloading the Google Street View images will be made publicly available.

4.2. Experimental Settings

4.2.1 Training dataset generation

We generated training datasets for the CSCDNet, the SSCDNet, and the CSSCDNet from the PCD, the Mapillary Vistas, and the PSCD datasets, respectively. Table 1 shows the details of the dataset. The PCD dataset is composed of panoramic image pairs \(I_1, I_2\) taken at two different time points \(t_1\) and \(t_2\), and the change mask \(M_g\). From the im-

Figure 4. Two example image pairs of the panoramic semantic change detection (PSCD) dataset.
Table 2. $F_1$ score and mIoU of change detection for TSUNAMI and GSV datasets. Siamese-CDResNet represents the CSCDNet without correlation layers. The CSCDNet consistently outperforms the other methods.

| Method          | TSUNAMI $F_1$ score | TSUNAMI mIoU | GSV $F_1$ score | GSV mIoU | Average $F_1$ score | Average mIoU |
|-----------------|---------------------|--------------|-----------------|----------|---------------------|--------------|
| DenseSIFT [33]  | 0.649               | 0.528        | 0.589           |          | –                   | –            |
| CNN-feat [33]   | 0.723               | 0.639        | 0.681           |          | –                   | –            |
| DeconvNet [1]   | 0.774               | 0.614        | 0.694           |          | –                   | –            |
| WS-Net [24]     | –                   | –            | –               | –        | 0.477               | –            |
| FS-Net [24]     | –                   | –            | –               | –        | –                   | 0.588        |
| CDNet [36]      | 0.848               | 0.695        | 0.772           | 0.811    | 0.672               | 0.741        |
| Siamese-CDResNet (Ours) | 0.850 | 0.718 | 0.784 | 0.815 | 0.691 | 0.753 |
| CSCDNet (Ours)  | 0.859               | 0.738        | 0.799           | 0.824    | 0.706               | 0.765        |

The age set $[I_1, I_2, M_g]$, patch images are cropped by sliding and resized. Furthermore, data augmentation is performed by rotating the patches. Thus, 12,000 sets of image patches were generated. The PSCD dataset is resized, and cropped and data augmentation is performed in the same way as the PCD dataset.

We also generated training datasets for the SSCDNet from the Mapillary Vistas dataset [29]. The Mapillary Vistas dataset for research use contains 20,000 scene images and the pixel-wise semantic labels with 66 semantic classes (including an unlabeled class). We integrate them into the following 11 classes: animal, vehicle, barrier, area, structure, lane marking, vegetation, traffic, others, debris, and no change. (See the supplementary material for the detail of the class integration.) Furthermore, the PSCD dataset and the subset TSUNAMI of the PCD dataset used in the final experiment contain much debris, hence, we added 150 debris images (100 and 50 images for training and validation, respectively) into the dataset used for the dataset synthesis mentioned in Sec. 3.2. We selected the value of $n_{max}$ as 3 because the silhouette information may be lost if too many classes are sampled. Figure 2 shows an example of the dataset synthesized by the proposed method.

### 4.2.2 Data augmentation for robustness to change detection error

If the change masks that are synthesized from semantic segmentation datasets are directly used in the training of the SSCDNet, the trained SSCDNet can be vulnerable to errors in change detection. To improve the robustness of the SSCDNet to change detection error, we perform the data augmentation for change mask in the training. Specifically, the change mask is randomly applied to one of the four morphological transformations (erosion, dilation, opening, closing) with a random kernel size $k$ ($1 \leq k \leq 20$). We expect that the semantic label information can reduce the error of semantic change detection due to the error of change detection by simulating the change mask.

### 4.2.3 Training details

The CSCDNet, the SSCDNet, and the CSSCDNet are trained using eight Nvidia Tesla P100 GPUs using the PyTorch framework. We used the batch size of 32. The numbers of iteration for the CSCDNet, the SSCDNet, and the CSSCDNet are $3 \times 10^4$, $1 \times 10^5$, and $1 \times 10^5$, respectively. The Adam algorithm with a learning rate of $2 \times 10^{-4}$ is used. The evaluations of the estimation accuracies of the CSCDNet using the PCD dataset and the CSSCDNet using the PSCD dataset are performed using the five-fold cross-validation.

### 4.3 Evaluation

#### 4.3.1 Change detection for the PCD dataset

Table 2 shows $F_1$ scores and mean intersection-over-union (mIoU) of each method for TSUNAMI and GSV datasets. The CSCDNet outperforms the other methods in terms of both $F_1$ scores and mIoU. Furthermore, the improvements of the scores for GSV are more significant than those of TSUNAMI. The main reason is that GSV contains more precise changes and the camera viewpoint differences are relatively larger than TSUNAMI because of the differences in their scene depths. The CSCDNet can accurately detect the precise scene changes dealing with the differences of camera viewpoints.

#### 4.3.2 Accuracy of the SSCDNet for synthetic data

Figure 5 shows an example of the results estimated using the SSCDNet. The SSCDNet can accurately estimate semantic changes on each input image even if there are overlapping areas of change between input images. Table 3 shows the IoU of each class and the mIoU of the SSCDNet for the synthetic validation data from the Mapillary Vistas dataset. There are four combinations of training and test datasets with or without the data augmentation of the aforementioned change mask. In the case of test data without data augmentation, namely, the input change mask is quite accurate, the SSCDNet trained using the dataset without the data augmentation performs better than one trained with the augmentation. However, in the case of test data with the
Table 3. mIoU of the SSCDNet for synthetic data from the Mapillary Vistas dataset. Each column shows the result of the SSCDNet trained with and without the data augmentation (DA) of the change mask in the test and the training phases.

| DA for test | - | ✓ | DA for training | - | ✓ |
|-------------|---|---|------------------|---|---|
| Animal      | 0.441 | 0.407 | 0.197 | 0.278 |
| Vehicle     | 0.746 | 0.733 | 0.499 | 0.659 |
| Barrier      | 0.482 | 0.509 | 0.225 | 0.402 |
| Area        | 0.800 | 0.873 | 0.646 | 0.803 |
| Structure   | 0.727 | 0.725 | 0.494 | 0.634 |
| Lane Marking | 0.544 | 0.467 | 0.252 | 0.377 |
| Vegetation  | 0.809 | 0.790 | 0.596 | 0.726 |
| Object (Traffic) | 0.445 | 0.341 | 0.181 | 0.290 |
| Object (Others) | 0.529 | 0.403 | 0.199 | 0.272 |
| Debris      | 0.294 | 0.140 | 0.248 | 0.201 |
| No Change   | 0.982 | 0.976 | 0.916 | 0.955 |
| mIoU        | 0.624 | 0.579 | 0.405 | 0.511 |

4.3.3 Semantic change detection for the PSCD dataset

Figure 6 shows examples of the semantic change detection results for the entire process of our proposed method. (See the supplementary material for other results.) Table 4 shows mIoU of each method for the PSCD dataset. Certainly, if the semantic change detection dataset, whose creation is labor-intensive, is available, the strategy of the end-to-end learning for semantic change detection can be applied, and the performance is almost the best (CSSCDNet). However, even if the dedicated dataset is unavailable, the SSCDNet can estimate semantic scene changes for each input image successfully depending on the change detection accuracy. For example, in the top rows of Fig. 7, the CSCDNet can accurately detect scene changes although there are some detections errors owing to reflections from window-glasses and advertisement boards on the buildings because of the lack of training data. (The CSCDNet trained with the PCD dataset can detect some changes of advertisement boards but not those of vegetations, and vice versa for the CSCDNet trained using the PSCD dataset.) In the lower part of Fig. 7, the lack of training data causes large errors. The better performance of the CSCDNet trained using the PSCD dataset than that using the PCD dataset indicates that even only the change detection dataset of the same domain as the target data should be used if it is available.

The SSCDNet trained using the augmentation estimates the change areas and the semantic labels considering their semantic contexts, thus, removes uncertain areas. Unlike the synthetic data (Sec. 4.3.2), the SSCDNet trained with the data augmentation for the change mask performs better than that trained without the augmentation even when using the ground-truth change mask. These results indicate that the augmentation is effective for not only change detection but also semantic labeling. Furthermore, the SSCDNet using the ground-truth change mask performs better than the CSSCDNet, which is trained using the full set of the semantic change detection dataset. Hence, the SSCDNet will exhibit higher performance when accurate change mask information is available by other methods [42, 43, 34] and sensors.

5. Conclusion

We proposed a novel semantic change detection scheme with only weak supervision. As far as we know, this is the first method studying on the semantic scene change detection task. The proposed method is composed of the two CNNs, the CSCDNet and the SSCDNet. The SSCDNet can deal with difference of camera viewpoints and achieves state of the art change detection performance for the PCD dataset. The SSCDNet can be trained with dataset synthe-
sized from semantic image segmentation datasets to avoid creating a new dataset for semantic change detection. To evaluate the effectiveness of the proposed method, we created the first publicly available dataset for semantic scene change detection, named as the PSCD dataset. Experimental results with this dataset verified the effectiveness of the proposed scheme in the semantic change detection task.

In the future, we intend to improve the estimation accuracy of the proposed method. Domain adaptation methods [46, 51] could bridge the gap between the networks trained using the PSCD dataset and others. Also, panoramic convolutional layers [7, 44] will improve the estimation accuracies for the PCD and the PSCD datasets, which are composed of panoramic images. Furthermore, the probability distribution of the semantic label sampling in the training dataset synthesis for the SSCDNet might be decided by prior distribution of the number of semantic labels to address class imbalance problems.
Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 18K18071.

References

[1] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi. Street-View Change Detection with Deconvolutional Networks. In Robotics: Science and Systems, 2016.

[2] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. arXiv preprint arXiv:1702.01105, 2017.

[3] N. Bourdis, D. Marraud, and H. Sahbi. Constrained optical flow for aerial image change detection. In IGARSS, 2011.

[4] M. Camplani, L. Maddalena, G. M. Alcover, A. Petrosino, and L. Salgado. A benchmarking framework for background subtraction in rgbd videos. In International Conference on Image Analysis and Processing, 2017.

[5] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang. A deep visual correspondence embedding model for stereo matching costs. In ICCV, 2015.

[6] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In CVPR, 2005.

[7] B. Coors, A. Paul Condurache, and A. Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In The European Conference on Computer Vision (ECCV), September 2018.

[8] D. Crispell, J. Mundy, and G. Taubin. A Variable-Resolution Probabilistic Three-Dimensional Model for Change Detection. IEEE Transactions on Geoscience and Remote Sensing, 50(2), 2012.

[9] R. C. Daudt, B. Le Saux, and A. Boulch. Fully convolutional siamese networks for change detection. In 2018 25th IEEE International Conference on Image Processing (ICIP), pages 4063–4067. IEEE, 2018.

[10] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau. High Resolution Semantic Change Detection. arXiv preprint arXiv:1810.08452, 2018.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR, 2009.

[12] J. Dong, J. G. Burnham, B. Boots, G. Rains, and F. Dellaert. 4D Crop Monitoring: Spatio-Temporal Reconstruction for Agriculture. In ICRA, 2017.

[13] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In ICCV, December 2015.

[14] D. Dwibedi, I. Misra, and M. Hebert. Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection. In ICCV, 2017.

[15] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. IJCV, 59(2), 2004.

[16] A. Fujita, K. Sakurada, T. Imaizumi, R. Ito, S. Hikosaka, and R. Nakamura. Damage Detection from Aerial Images via Convolutional Neural Networks. In MVA, 2017.

[17] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. In CVPR, 2016.

[18] R. Hamaguchi and S. Hikosaka. Building detection from satellite imagery using ensemble of size-specific detectors. In CVPR, June 2018.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016.

[20] A. Huertas and R. Nevatia. Detecting Changes in Aerial Views of Man-Made Structures. In ICCV, 1998.

[21] ISPRS. 2D Semantic Labeling Contest. http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html.

[22] H. Kataoka, S. Shirakabe, Y. Miyashita, A. Nakamura, K. Iwata, and Y. Satoh. Semantic Change Detection with Hypermaps. arXiv preprint arXiv:1604.07513, 2016.

[23] S. H. Khan, X. He, F. Porikli, and M. Bennamoun. Forest change detection in incomplete satellite images with deep neural networks. IEEE Transactions on Geoscience and Remote Sensing, 55(9), 2017.

[24] S. H. Khan, X. He, F. Porikli, M. Bennamoun, F. Sohel, and R. Togneri. Learning Deep Structured Network for Weakly Supervised Change Detection. In IJCAI, 2017.

[25] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays. Learning deep representations for ground-to-aerial geolocalization. In CVPR, 2015.

[26] K. Matzen and N. Snavely. Scene Chronology. In ECCV, 2014.

[27] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison. SceneNet RGB-D: Can 5M Synthetic Images Beat Generic ImageNet Pre-training on Indoor Segmentation? 2017.

[28] V. Mnih. Machine Learning for Aerial Image Labeling. PhD thesis, University of Toronto, 2015.

[29] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kontschieder. The Mapillary Vistas Dataset for
Semantic Understanding of Street Scenes. In *ICCV*, 2017.

[30] T. Pollard and J. L. Mundy. Change Detection in a 3-d World. In *CVPR*, 2007.

[31] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image Change Detection Algorithms: A Systematic Survey. *TIP*, 14(3), 2005.

[32] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *CVPR*, 2016.

[33] K. Sakurada and T. Okatani. Change Detection from a Street Image Pair using CNN Features and Superpixel Segmentation. In *BMVC*, 2015.

[34] K. Sakurada, T. Okatani, and K. Deguchi. Detecting Changes in 3D Structure of a Scene from Multi-view Images Captured by a Vehicle-Mounted Camera. In *CVPR*, 2013.

[35] K. Sakurada, T. Okatani, and K. M. Kitani. Massive City-scale Surface Condition Analysis using Ground and Aerial Imagery. In *ACCV*, 2014.

[36] K. Sakurada, W. Wang, N. Kawaguchi, and R. Nakamura. Dense Optical Flow based Change Detection Network Robust to Difference of Camera Viewpoints. *arXiv preprint arXiv:1712.02941*, 2017.

[37] G. Schindler and F. Dellaert. Probabilistic temporal inference on reconstructed 3D scenes. In *CVPR*, 2010.

[38] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017.

[39] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *ICCV*, 2015.

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

[41] S. Stent, R. Gherardi, B. Stenger, and R. Cipolla. Detecting Change for Multi-View, Long-Term Surface Inspection. In *BMVC*, 2015.

[42] A. Taneja, L. Ballan, and M. Pollefeys. Image based detection of geometric changes in urban environments. In *ICCV*, 2011.

[43] A. Taneja, L. Ballan, and M. Pollefeys. City-Scale Change Detection in Cadastral 3D Models Using Images. In *CVPR*, 2013.

[44] K. Tateno, N. Navah, and F. Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[45] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *ICCV*, 1999.

[46] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, June 2018.

[47] K. Wang, C. Gou, and F.-Y. Wang. M4CD: A Robust Change Detection Method for Intelligent Visual Surveillance. *arXiv preprint arXiv:1802.04979*, 2018.

[48] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar. CDnet 2014: An expanded change detection benchmark dataset. In *CVPRW*, 2014.

[49] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, 2015.

[50] J. Zhontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *JMLR*, 17(1), 2016.

[51] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, September 2018.
Supplementary Material

A. Details of network architectures

Figures 8-11 show the details of the network architectures for each block, the CSCDNet, the SSCDNet, and the CSSCDNet, respectively. The parentheses in the correlation layers [13] represent the channel dimension of each output feature map. The correlation between two patches centered at \( x_1 \) in the first map \( f_1 \) and \( x_2 \) in the second map \( f_2 \) is then defined as

\[
c(x_1, x_2) = \sum_{o \in [-k,k] \times [-k,k]} \langle f_1(x_1 + o), f_2(x_2 + o) \rangle \tag{3}
\]

for a square patch of size \( K := 2k + 1 \). The output channel dimension is calculated as \( K^2 \). We set \( k = 10 \) through all experiments in this paper.

B. Details of class integration

The Mapillary Vistas dataset [29] for research use contains 20,000 scene images and pixel-wise semantic labels with 66 semantic classes (including an unlabeled class). We integrate them into the following 11 classes: animal, vehicle, barrier, area, structure, lane marking, vegetation, traffic, others, debris, and no change. Table 5 shows the detail of the integration.

Table 5. Details of class integration. The Mapillary Vistas dataset does not contain debris in their images, hence we added images taken in tsunami-damaged areas.

| Integrated classes | Original classes of the Mapillary Vistas dataset |
|--------------------|--------------------------------------------------|
| Animal             | Bird, Ground Animal, Person, Bicyclist, Motorcyclist, Other Rider |
| Vehicle            | Bicycle, Boat, Bus, Car, Caravan, Motorcycle, On Rails, Other Vehicle, Trailer, Truck, Wheeled Slow |
| Barrier            | Curb, Fence, Guard Rail, Other Barrier, Wall |
| Area               | Bike Lane, crosswalk Plain, Curb Cut, Parking, Pedestrian Area, Rail Track, Road, Service Lane, Sidewalk |
| Structure          | Bridge, Building, Tunnel |
| Lane Marking       | Crosswalk Zebra, Marking General |
| Vegetation         | Vegetation |
| Traffic            | Traffic Light, Traffic Sign Back, Traffic Sign Front |
| Others             | Banner, Bench, Bike Rack, Billboard, Catch Basin, Cctv Camera, Fire Hydrant, Junction Box, Mailbox, Manhole, Phone Booth, Pothole, Street Light, Pole, Traffic Sign Frame, Utility Pole, Trash Can |
| Debris             | (From our additional images taken in tsunami-damaged areas) |
| No Change          | Mountain, Sand, Sky, Snow, Terrain, Water, Car Mount, Ego Vehicle, Unlabeled |

C. Additional results

Figures 12-15 show other semantic change detection results of our proposed method for the PSCD dataset. These qualitative results indicate that our proposed method can estimate semantic changes from an image pair of a scene successfully. Figure 16 shows examples of the failure cases. In the future, we intend to reduce these estimation errors due to lack of training data and camera projection models by introducing domain adaption methods [46, 51] and panoramic convolution layers [7, 44].

Figures 17-18 show examples of the semantic change detection results for TSUNAMI and GSV of the PCD dataset [33], which does not contain ground-truth of the semantic change labels.
Figure 8. The details of each block architecture.

Figure 9. The details of the CSCDNet architecture. The parentheses in the correlation layers represent the channel dimension of each output feature map.
Figure 10. The details of the SSCDNet architecture.

Figure 11. The details of the CSSCDNet architecture. The parentheses in the correlation layers represent the channel dimension of each output feature map.
Figure 12. Examples of semantic scene change detection for the PSCD dataset
Figure 13. Examples of semantic scene change detection for the PSCD dataset
Figure 14. Examples of semantic scene change detection for the PSCD dataset
| $t_0$ | $t_1$ | GT of change mask | CSSCDNet | GT change mask + SSCDNet | CSSCDNet (PCD) | CSSCDNet (PSCD) |
|-------|-------|-------------------|----------|--------------------------|----------------|----------------|
| $t_0$ | $t_1$ | GT of change mask | CSSCDNet | GT change mask + SSCDNet | CSSCDNet (PCD) | CSSCDNet (PSCD) |
| $t_0$ | $t_1$ | GT of change mask | CSSCDNet | GT change mask + SSCDNet | CSSCDNet (PCD) | CSSCDNet (PSCD) |
| $t_0$ | $t_1$ | GT of change mask | CSSCDNet | GT change mask + SSCDNet | CSSCDNet (PCD) | CSSCDNet (PSCD) |
| $t_0$ | $t_1$ | GT of change mask | CSSCDNet | GT change mask + SSCDNet | CSSCDNet (PCD) | CSSCDNet (PSCD) |
| $t_0$ | $t_1$ | GT of change mask | CSSCDNet | GT change mask + SSCDNet | CSSCDNet (PCD) | CSSCDNet (PSCD) |
| $t_0$ | $t_1$ | GT of change mask | CSSCDNet | GT change mask + SSCDNet | CSSCDNet (PCD) | CSSCDNet (PSCD) |

Figure 15. Examples of semantic scene change detection for the PSCD dataset.
GT of change mask
CSCDNet (PCD)
CSCDNet (PSCD)

GT label of \( t_0 \) and \( t_1 \)
CSSCDNet
GT change mask + SSCDNet
CSCDNet (PCD) + SSCDNet
CSCDNet (PSCD) + SSCDNet

Figure 16. Examples of failure cases of semantic scene change detection for the PSCD dataset
Figure 17. Examples of semantic scene change detection for the TSUNAMI of the PCD dataset
Figure 18. Examples of semantic scene change detection for the GSV of the PCD dataset