2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Self-supervised Simultaneous Alignment and Change Detection

Yukuko Furukawa, Kumiko Suzuki, Ryuhei Hamaguchi, Masaki Onishi, Ken Sakurada

***Abstract*— This study proposes a self-supervised method for detecting scene changes from an image pair. For mobile cameras such as drive recorders, to alleviate the camera viewpoints' difference, image alignment and change detection must be optimized simultaneously because they depend on each other. Moreover, lighting condition makes the scene change detection more difficult because it widely varies in images taken at different times. To solve these challenges, we propose a self-supervised simultaneous alignment and change detection network (SACD-Net). The proposed network is robust specifically in differences of camera viewpoints and lighting conditions to simultaneously estimate warping parameters and multi-scale change probability maps while change regions are not taken into account of calculation of the feature consistency and semantic losses. Based on comparative analysis between our self-supervised and the previous supervised models as well as ablation study of the losses of SACD-Net, the results show the effectiveness of the proposed method using a synthetic dataset and our new real dataset.**

I. INTRODUCTION

Today, busy cities experience an enormous amount of changes every day, such as urban innovation, heavy traffic, and vegetation changes. Monitoring of the scene changes has a wide range of applications, including urban planning and management, commercial marketing, smart technologies development, and environmental assessment. In computer vision and remote sensing, this topic has been widely studied [1]–[7], to automatically detect scene changes by comparing a pair of images captured at different times.

In most applications, change detection algorithms require pixel-level alignment of input images because the images captured by mobile cameras, such as drive recorders, usually have different camera viewpoints. In this case, the image appearance is significantly changed. This results in a poor performance on a naive approach of comparing pixel values or pixel features between the two images.

In previous methods, the process of image alignment means a data pre-processing task, separately from the change detection. The alignment is based on sparse corresponding points extracted from whole areas of input images, where the change regions are assumed to be small enough, and the impact can be ignored in the process [8]. However, in many real-world situations, a target scene often contains large or scattered changes. In these cases, the change regions should also be considered in the alignment process. Moreover, both of image alignment and change detection are to be accurately performed nearly at the same time.

*This work was partially supported by the New Energy and Industrial Technology Development Organization (NEDO) and JSPS KAKENHI Grant Number 20H04217.

The authors are associated with National Institute of Advanced Industrial Science and Technology, Koto-ku, Tokyo, 135-0064, Japan {furukawa-y, kumiko2.suzuki,ryuhei.hamaguchi, onishi-masaki, k.sakurada}@aist.go.jp

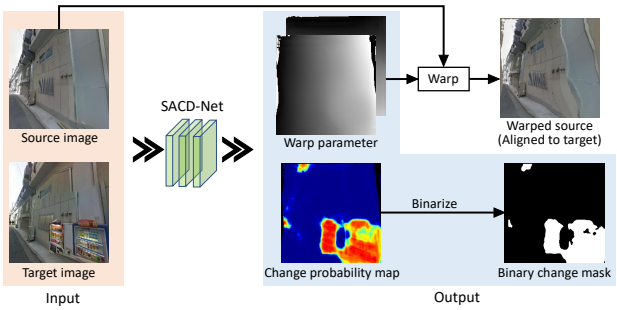


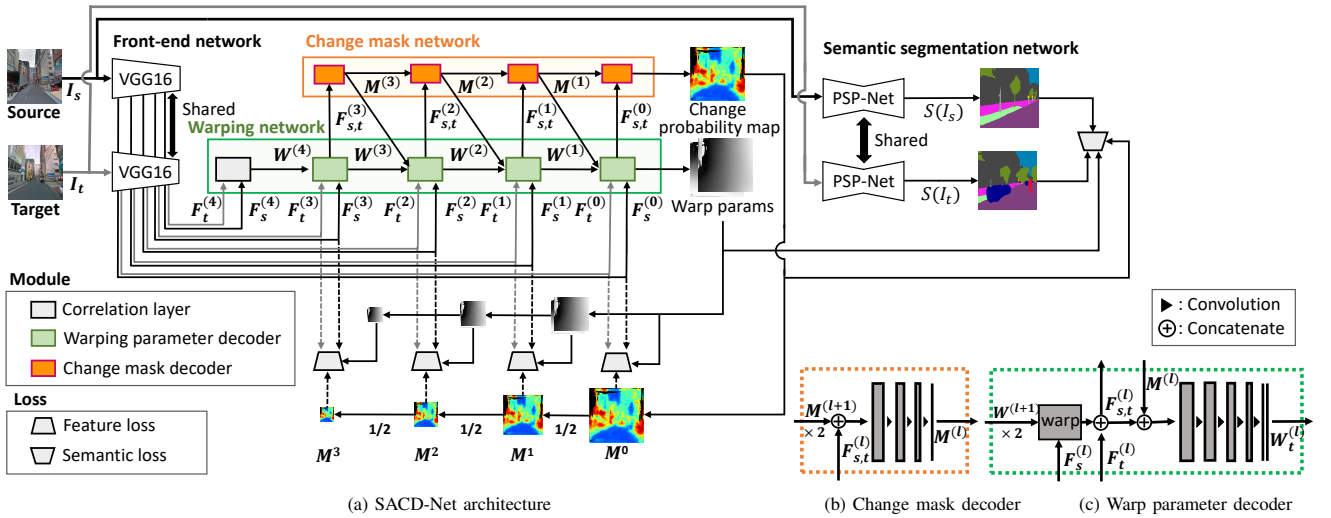
Fig. 1. Overview of SACD-Net. The network aims to simultaneously estimate a warping parameter for alignment and a change mask from a given pair of source and target images, which often appear differently depending on spatiotemporal changes.

In addition, the image pairs that need a change detection process generally involve different lightning conditions such as color changes of the sky and free shapes of shadows because they are taken at various times in a day. To avoid this adverse lighting effect, it is essential to absorb such a trivial difference occurring in a pair of given images.

To this end, we propose a Convolutional Neural Network (CNN) model, or a simultaneous alignment and change detection network (SACD-Net), to simultaneously estimate pixel-wise warping parameters and a change probability map from a given image pair (Fig. 1). The proposed model can be trained in an end-to-end and self-supervised manner by utilizing a consistency constraint between images before and after warping. Meanwhile, the change probability map is estimated as regions that cannot satisfy the consistency constraint, even with correct warping. We apply our model to real-world and simulation datasets to make comparative analysis between our model and previous fully-supervised models as well as ablation study to show the effectiveness our model. The main contributions of this work, we propose:

- a method that simultaneously performs both the image alignment and change detection, which have been separately handled in the previous change detection methods.
- a change detection CNN model that can be learned in an end-to-end and self-supervised manner without any ground truth for the change mask.

The results show the effectiveness of the proposed method that simultaneously estimates alignment and scene change between image pair to reduce the effects, such as differences in the lighting conditions or camera viewpoints.



(a) SACD-Net architecture (b) Change mask decoder (c) Warp parameter decoder

Fig. 2. Schematics of the proposed network (SACD-Net). The network consists of a front-end network (VGG-16), warping network, change mask network, and semantic segmentation network (PSP-Net). Based on the features from the front-end network, the warping network and change mask network estimates warp parameters and change probability map in a coarse-to-fine manner. Each change mask decoder and warp parameter decoder consists some convolution layers. The semantic segmentation network promotes training to use high-level features.

II. Related Work

In this section we review works that tackle the task of image alignment and change detection.

A. Image Alignment

One of the straightforward methods for image alignment is to match feature points between input images and estimate the warping parameters using the matched points. Scale-invariant feature transforms (SIFT) [9] and speeded up robust features (SURF) [10] are widely used as feature descriptors, although these hand-crafted descriptors have some limitations in terms of illumination changes and strong deformations. Feature descriptors based on CNNs can extract features that are robust against changes of illumination and camera viewpoint [11] [12]. However, these methods cannot cope with strong deformation nor non-rigid changes of a given image because the extracted features are represented as patches, meaning that the resulting correspondence points are too sparse to handle.

Rocco et al. proposed a method to apply to various types of image deformation to estimate pixel-wise correspondences between input images in the form of warping parameters consisting of 24 elements (6 parameters for affine and 18 parameters for thin-plate spline transformations) [13] [14]. Melekhov et al. proposed a Dense Geometric Correspondence Network (DGC-Net) to directly estimate the correspondence of each pixel, and introduced a matchability map to estimate common fields of view between images [15]. These approaches can learn correspondences in a self-supervised way, even if images were strongly deformed. The approach benefits from synthetic datasets where arbitrary warping is added. However, they focus on image pairs without object changes, appearing or disappearing, nor illumination; therefore, they cannot be directly applied to change detection problems.

B. Change Detection

Some methods work well to recognize scene changes by calculating image subtraction when camera positions are fixed [16]–[19]; however, mobile cameras require alignment before change detection.

Mukojima et al. proposed a method to detect obstacles on rails using a camera attached to the front of trains [8]. The method uses a DeepFlow architecture to align images using a flow field estimated from sparse correspondences [20]. As mentioned above, these alignment methods based on sparse features are more difficult to apply to images that have strong deformation or non-rigid changes. Moreover, the alignment is executed ignoring changes.

Another approach takes input of an image pair from their different viewpoints to directly predict the change areas [21]–[27]. Guo et al. proposed a convolutional siamese metric network, which introduces a threshold in the change distance loss to mitigate the difference in camera views [28]. Sakurada et al. proposed CSCDNet, which utilizes a correlation layer to detect scene changes in vehicular imagery [6]. However, these methods require the ground truth of a change mask as a supervision signal, which in practice is laborious to collect.

Upon deep consideration of these previous studies, image alignment and acquisition of ground truth datasets are yet difficult problems for detecting changes between images. Thus, in this work, we propose a method for simultaneous image alignment and change detection. Moreover, we introduce feature loss and semantic loss to reduce the effect of illumination changes.

III. Method

The proposed CNN model, Simultaneous Alignment and Change Detection Network (SACD-Net), is to simultaneously estimate both the pixel-wise warping parameter $\boldsymbol{W}$ and the change mask $\boldsymbol{M}$ between the images, featuring an

end-to-end and self-supervised training manner. This section describes the overview of the network architecture and the loss function for self-supervised training.

A. Network architecture

As shown in Fig 2, the SACD-Net architecture consists of four parts: a front-end network, a warping network, a change mask network, and a semantic segmentation network.

Front-end network: to extract multi-level feature pyramids $\boldsymbol{F}_s = \{\boldsymbol{F}_s^{(0)}, \ldots, \boldsymbol{F}_s^{(L)}\}$ and $\boldsymbol{F}_t = \{\boldsymbol{F}_t^{(0)}, \ldots, \boldsymbol{F}_t^{(L)}\}$ from the source and the target image to calculate a consistency loss. The feature pyramid is composed of intermediate feature maps $\boldsymbol{F}^{(l)}$ from five different layers of the front-end network (i.e., L= 4). For this network, VGG-16 architecture [29] is previously trained on the ImageNet dataset [30].

Warping network: to estimate a pixel-wise warping parameter $\boldsymbol{W}_{s\to t} = \{\boldsymbol{W}_{s\to t}^{(0)}, \ldots, \boldsymbol{W}_{s\to t}^{(L)}\}$. As proposed in [15], the parameter is estimated in a coarse-to-fine manner using the extracted feature pyramid $\boldsymbol{F}_s$ and $\boldsymbol{F}_t$. First, the top-level features $\boldsymbol{F}_s^{(4)}$ and $\boldsymbol{F}_t^{(4)}$ of the pyramids are fed into a correlation layer [31] that estimates a coarse warping parameter $\boldsymbol{W}^{(4)}$ which is then fed into the sequence of warping parameter decoders, where the parameter is refined stage by stage. Fig 2 shows the architecture of the warping parameter decoders. Inside each decoder, a warping parameter $\boldsymbol{W}^{(l+1)}$ from the previous stage is first upsampled. Then the source feature $\boldsymbol{F}_s^{(l)} \in \boldsymbol{F}_s$ is warped by the upsampled parameter, and concatenated with the target feature $\boldsymbol{F}_t^{(l)} \in \boldsymbol{F}_t$. The concatenated feature $\boldsymbol{F}_{s,t}^{(l)}$ is fed into the decoder with four or five convolutional layers that outputs a finer-scale warping parameter. $\boldsymbol{F}_{s,t}^{(l)}$ is also used to estimate a change probability map in the change mask network. Additionally, a change probability mask simultaneously learned in the change mask network is fed into the warping network in each stage, which serves as weighting for a warping parameter, namely a change area is less trained. The opposite warping parameter $\boldsymbol{W}_{t\to s}$ is simply calculated by converting an input source image and a target image.

Change mask network: to estimate a change probability map $\boldsymbol{M} \in \{\boldsymbol{M}_s, \boldsymbol{M}_t\}$ between input images, where the subscript represents a viewpoint. As the warping network, the network also takes a coarse-to-fine strategy. It consists of a sequence of change mask decoders (Fig. 2). In each decoder, the estimated probability mask $\boldsymbol{M}^{(l+1)}$ from the previous stage is first upsampled, and then concatenated with the feature $\boldsymbol{F}_{s,t}^{(l)}$ coming from the warping network. The concatenated feature is fed into the decoder with three or four convolution layers that output a finer-scale change probability map, while the change map is also fed to the warping network as mentioned earlier. The initial change probability map at the first stage is estimated using the feature from the warping network ($\boldsymbol{F}_{s,t}^{(3)}$), as there is no change probability map from the previous stage. The final estimation $\boldsymbol{M}$ is acquired by applying a sigmoid function to the output of the last change mask decoder. Hence, $\boldsymbol{M}$ has a value range of $[0, 1]$, 0 represents change, and 1 represents unchanged.

Segmentation network: a unique feature of our network architecture to improve the consistency in semantic labels before and after being warped. This network uses a Pyramid scene parsing network (PSPNet) architecture proposed in [32], which is one of the robust and well-performing segmentation models.

B. Loss function

The loss function is used for the self-supervised training. This section describes three key factors to achieve the function in our SACD-Net: feature consistency loss, mask regularization loss, and semantic consistency loss.

Feature consistency loss: ideally, a warped source image should match the target image excluding change regions. The key idea of the proposed method is warping correction to be learned in a self-supervised manner under the constraint that the target image excludes change regions. Here, RGB image space is not held because pixel values are significantly affected by lighting conditions. Therefore, our method is designed to maintain consistency in *feature space* where robustness to lighting conditions can be achieved by high-level semantic features. Specifically, the proposed model penalizes the Euclidean distance between the warped source features and the target features. Batch normalization is applied to features before calculating the loss so that all datasets have the same value range.

Given the estimated warping parameter $\boldsymbol{W}_{s\to t}$ and change probability map $\boldsymbol{M}$, we define warped source feature $\boldsymbol{W}_{\boldsymbol{s\to t}}^{(l)}(\boldsymbol{F}_s^{(l)})$ and down-sampled change probability map $\boldsymbol{M}^{(l)}$ at l-th stage. Thus, the feature consistency loss is:

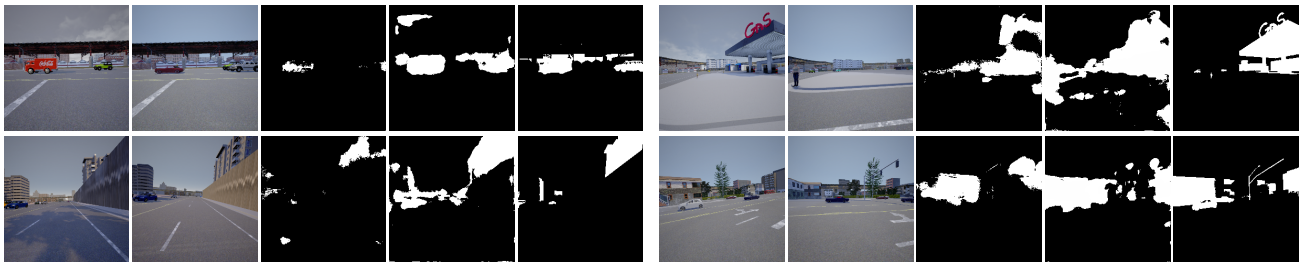
$$L_f = \sum_l \lambda^{(l)} \frac{1}{h^{(l)} w^{(l)}} \sum_{i,j} \left[\boldsymbol{M}_t^{(l)} \odot f_{MSE}(\boldsymbol{W}_{\boldsymbol{s\to t}}^{(l)}(\boldsymbol{F}_s^{(l)}), \boldsymbol{F}_t^{(l)}) \right]_{i,j} \tag{1}$$

Here, $h^{(l)}, w^{(l)}$ are the height and width of the l-th feature map. $f_{MSE}(\cdot,\cdot)$ represents an the Euclidean distance map between the source and the target feature. The loss is calculated bidirectionally between a target and a source, meaning from a target to a source and vice verse.

By minimizing Eq. (1), the warping network can learn the correct warping parameter while the change mask network learns how to recognize change regions as those that cannot satisfy the consistency constraint, even with correct warping.

Mask regularization loss: minimization of the feature consistency loss can easily find a trivial solution where the whole scene is estimated as a change region, and all the pixels are eliminated from the consistency loss. To prevent training from falling into such a trivial solution, a regularization term is introduced for the estimated change probability map.

$$L_m = \frac{1}{hw} \sum_{i,j} [-\log \boldsymbol{M}*]_{i,j} \tag{2}$$



(a) (b) (c) (d) (e) (a) (b) (c) (d) (e)

Fig. 3. Pers-SYM images for change detection. Four target images (a) are respectively followed by (b) source, (c) PWC-Net+FC-Siam-diff viewpoint, (d) SACD-Net viewpoint (ours), and (e) ground truth with change mask.

(a) (b) (c) (d) (e) (a) (b) (c) (d) (e)

Fig. 4. Pers-GSV images for change detection. Four target images (a) are respectively followed by (b) source, (c) PWC-Net+FC-Siam-diff viewpoint, (d) SACD-Net viewpoint (ours), and (e) ground truth with change mask.

where $\boldsymbol{M*}$ represents the intersection of change masks of both directions as shown below.

$$\boldsymbol{M*} = \boldsymbol{M_t} \odot \boldsymbol{W_{s\to t}}(\boldsymbol{M_s}) \tag{3}$$

As mentioned above, Eq. (2) is mainly a constraint to prevent all $\boldsymbol{M*}$ to be labeled as 0, i.e. "changed". Eq. (2), additionally, imposes a consistency between the change mask $\boldsymbol{M_t}$ and warped change mask $\boldsymbol{W_{s\to t}}(\boldsymbol{M_s})$ because minimizing the intersection of these two masks $\boldsymbol{M*}$ is equal to minimizing the gap between two change masks ($\boldsymbol{M_t}$, $\boldsymbol{W_{s\to t}}(\boldsymbol{M_s})$), which are supposed to be the same.

Semantic consistency loss: to attend to high-level semantic change rather than low-level change, such as illumination change, semantic-level information can be helpful for better training. Specifically, PSPNet [32] pre-trained on the Cityscapes dataset [33] is applied to acquire class probability maps $\boldsymbol{p}_s$ and $\boldsymbol{p}_t$ from source image $\boldsymbol{I}_s$ and target image $\boldsymbol{I_t}$. From the class probability maps, semantic consistency loss is calculated in the same way as feature consistency loss.

$$L_{sem} = \frac{1}{hw} \sum_{i,j} [\boldsymbol{M}_t \odot d_{KL}(\boldsymbol{W}_{s\to t}(\boldsymbol{p}_s), \boldsymbol{p}_t)]_{i,j} \tag{4}$$

Here, $d_{KL}(\cdot,\cdot)$ represents the KL-distance between the two class probability maps. Let $\boldsymbol{p}_{s\to t}^{(i,j)}$ and $\boldsymbol{p}_t^{(i,j)}$ represent the (i,j) element of $\boldsymbol{W}_{s\to t}(\boldsymbol{p}_s)$ and $\boldsymbol{p}_t$, respectively, i.e., the estimated class distribution at pixel (i,j) of the warped source and target image, then the KL-distance becomes:

$$d_{KL}^{(i,j)}(\boldsymbol{p}_{s\to t}, \boldsymbol{p}_t) = \sum_c p_{s\to t}^{(i,j)}(c) \left[\frac{\log p_{s\to t}^{(i,j)}(c)}{\log p_t^{(i,j)}(c)} \right], \tag{5}$$

where c represents the class index. The semantic consistency loss is also calculated in another direction (from target to source). PSPNet is also trained with this loss using a small learning rate. This is required because of the domain gap between the target dataset and Cityscapes dataset (the semantic segmentation model is slightly optimized and modified to adopt the target dataset via supervision of the consistency).

Total loss: letting α, β, and γ be the weights for each loss, the final loss becomes as follows.

$$L = \alpha L_f + \beta L_m + \gamma L_{sem} \tag{6}$$

The initial weight for each loss is set equivalently and consistently throughout the entire experiment.

IV. Experiments

We performed experiments to evaluate the effectiveness of our model in two types of approaches: comparative analysis between our self-supervised method and the previous supervised ones, and ablation study of the losses of SACD-Net. This section presents the details of the experiments such as the datasets, baseline and network training.

A. Experimental settings

Dataset: we selected a dataset along the roadside due to many images taken at different times via a camera mounted on a car. All dataset were resized to 240×240. We performed the experiments on two types of image datasets for real world and simulation: Pers-GSV, Pers-SYM, respectively. Additionaly, we prepared the Pers-SL dataset to increase the number of the real world's training data samples.

TABLE I

ACCURACY OF CHANGE DETECTION. F1 SCORE AND MEAN INTERSECTION-OVER-UNION (MIOU) OVER THE PERS-SYM AND PERS-GSV DATASET.

Method	Ground truth	Pers-SYM		Pers-GSV	
		F_1	mIoU	F_1	mIoU
FC-Siam-conc. [1]	✓	0.467 (0.545)	0.616 (0.654)	0.544	0.535
FC-Siam-diff. [1]	✓	0.471 (0.548)	0.617 (0.655)	0.544	0.534
PWC-Net [34] + FC-Siam-conc. [1]	✓	0.492 (0.612)	0.624 (0.689)	0.579	0.526
PWC-Net [34] + FC-Siam-diff. [1]	✓	**0.510** (0.610)	**0.632** (0.687)	0.586	0.528
CosimNet [28]	✓	0.494 (0.541)	0.624 (0.645)	0.561	0.536
CSCDNet [6]	✓	0.437 (0.508)	0.605 (0.636)	0.600	0.557
DASNet [35]	✓	0.400 (0.419)	0.534 (0.545)	0.489	0.304
DGC-Net [15] + consistency loss		0.245	0.452	0.402	0.328
SACD-Net w/o $L_{sem}(ours)$		0.383	0.512	0.531	0.558
SACD-Net w/ $L_{sem}(ours)$		0.497	0.611	**0.616**	**0.591**

TABLE II

COMPARISON OF FEATURE CONSISTENCY LOSS OVER PERS-GSV

L_f layers	$L_f^{(0)}$	$L_f^{(1)}$	$L_f^{(2)}$	$L_f^{(3)}$	F_1	mIoU
(a)	✓	✓	✓	✓	0.595	0.554
(b)		✓	✓	✓	0.604	0.576
(c)			✓	✓	**0.616**	0.587
(d)				✓	**0.616**	**0.591**
(e)	✓	✓	✓		0.580	0.527
(f)	✓	✓			0.572	0.527
(g)	✓				0.568	0.501

The Pers-GSV dataset is a new perspective google street-view change detection dataset created from Panoramic Semantic Change Detection dataset [6], which contains 1,536 image pairs taken at different time points (train: 1056, test: 480) with change masks as ground truth. The perspective image pairs were cropped from the panoramic image pairs captured with 90-degree-angular field of view. Among the image pairs, valid pairs with a common field of view were selected by feature matching with the SURF descriptor.

Additionally, we went through the same process as Pers-GSV creation to prepare our own street-level perspective image pair dataset (Pers-SL), which contains 6,144 image pairs (train: 4,928, validation: 1,216) without a ground truth of their change masks. This aims at making additional amount of training data available for unsupervised method.

The Pers-SYM dataset, proposed in [36], is built by the CARLA simulator [37], where perspective images are captured at different simulator settings with 90-degree-angular field of view. It also offers three more perspective images at each scene, of which yaw is shifted by approximately 10-degree intervals, the row and pitch are randomly shifted within 5 degree, and the horizontal translation is randomly shifted within 1 m. The dataset contains 15,000 scenes (train: 8,000, valid: 2,000, test: 5,000), and provides each scene with a change mask, a semantic label, and a depth image.

Baseline: to validate the effectiveness of the proposed method, we compared our method to both supervised and unsupervised baselines. For the supervised baselines, three types of methods were evaluated: (i) a method that ignores a viewpoint differences (FC-Siam-concat, -diff [1] and DASNet [35]); (ii) a two-stage method that first aligns image pairs to detect their changes (PWC-Net [34] +FC-Siam-cocat and -diff [1]; and (iii) a one-stage method that implicitly deals with viewpoint differences inside the network (CSCDNet [6] and CosimNet [28]). To train the Pers-SYM dataset, we also experimented with the pre-difined number of samples, which is also the same setting as Pers-GSV and we believe it's quantitatively reasonable to fairly evaluate the proposed model. For the unsupervised baseline, we built a naive version of the proposed method.

We simply added a consistency loss in *image space* to DGC-Net [15] architecture, and used its matchability mask decoder as a change mask decoder to validate the design of the proposed model, namely, multi-stage mask refinement and *feature space* consistency loss.

Training: the proposed model was trained unsupervisedly for the whole process, using Adam [38] optimizer with weight decay coefficient of 10^{-5}. The Front-end network weights were frozen. The initial learning rates of the warping network was set to 10^{-6}, those for the change mask network on Pers-SL and Pers-SYM was 10^{-4}, and those for the change mask network on Pers-GSV was 10^{-6}. PSPNet for the segmentation network was trained using a smaller learning rate of 10^{-8}. The weights in the total loss (α, β, and γ in Eq. 6) were set to 4, 1, and 1, respectively, which were adjusted to give each loss have an equivalent value. We trained on the Pers-SL before training the Pers-GSV dataset.

B. Comparative analysis

Table I shows the accuracy of change detection (F_1 score) and mean intersection-over-union (mIoU) over the Pers-SYM and Pers-GSV datasets. The values in brackets indicate the results where the supervised models were trained with all the labeled data of 8,000 image samples. The proposed SACD-Net performed competitively with the supervised baselines. Compared to DGC-Net [15], our model considerably improved the performance showing the effectiveness of the multi-stage mask refinement and feature space consistency loss. Although supervised methods achieved high scores when many labelled data were available in the Pers-SYM dataset, our self-supervised model still outperforms some of the baselines when the number of samples was restricted, which we believe is a practical situation. For all the models, we used the threshold of 0.5 for binarizing the output probability maps with no prior knowledge.

TABLE III

COMPARISON OF DIFFERENT VIEWPOINT OVER PRES-SYM

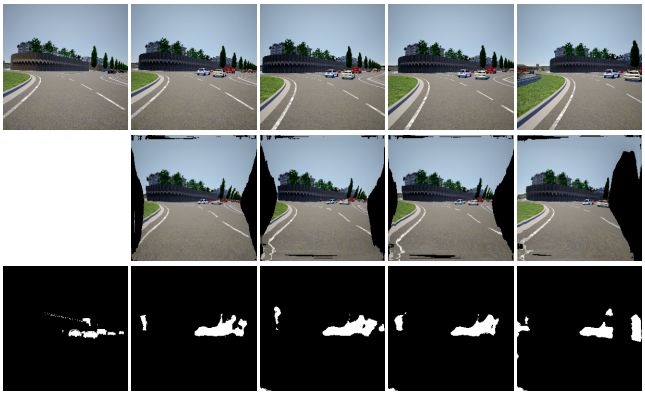
Method	Ground truth	$\Delta yaw = 0°$		0° - 10°		10° - 20°		20° - 30°	
		F_1	mIoU	F_1	mIoU	F_1	mIoU	F_1	mIoU
FC-Siam-conc. [1]	✓	0.514	0.651	0.464	0.614	0.466	0.615	0.444	0.594
FC-Siam-diff. [1]	✓	0.570	0.679	0.457	0.609	0.458	0.609	0.440	0.591
PWC-Net [34] + FC-Siam-conc. [1]	✓	0.648	0.718	0.493	0.625	0.493	0.625	0.439	0.590
PWC-Net [34] + FC-Siam-diff. [1]	✓	0.671	0.733	0.511	0.633	0.514	0.635	0.452	0.596
CosimNet [28]	✓	0.538	0.659	0.492	0.622	0.492	0.622	0.472	0.602
CSCDNet [6]	✓	0.580	0.683	0.409	0.592	0.404	0.589	0.393	0.577
DASNet [35]	✓	0.517	0.632	0.389	0.524	0.391	0.526	0.364	0.482
DGC-Net [15] + consistency loss		0.305	0.533	0.234	0.439	0.235	0.440	0.240	0.409
SACD-Net w/o $L_{sem}(ours)$		0.433	0.575	0.356	0.499	0.355	0.499	0.361	0.497
SACD-Net w/ $L_{sem}(ours)$		0.569	0.666	0.483	0.599	0.482	0.599	0.473	0.587

Fig 3 shows examples of the estimated change masks of the Pers-SYM dataset. From the upper left and lower right scenes, cars are correctly captured. Additionally, the top right scene indicates that images are aligned properly, and accordingly the change of a gas station is predicted. This change dataset, however, contains small and complicated change areas as shown in the left two scenes, resulting in relatively low performance for all models as a simulation dataset. Fig 4 shows examples of the estimated change masks of the Pers-GSV dataset. The warp is overall correct. The two scenes on the left show that the texture changes are captured correctly. In the upper right scene, the changes are well captured while the weather and shadows are different between source and target images.

C. Ablation study

Table II presents how the feature consistency losses in different layers (a) to (g) affect the performance on the Pers-GSV, where $L_f^{(3)}$ is at the highest level and $L_f^{(0)}$ at the lowest level. Lower-level features are known to be more sensitive to colors, edges, and high-level features to shapes and objects. In this experiment, the losses were eliminated one-by-one in the lower-to-higher order, from (b), (c) to (d), and higher-to-lower order, from (e), (f) to (g). The result shows that Pers-GSV performs best with the highest layer because the highest layer has the ability to learn better abstract and conceptual information such as change.

Table III shows the robustness of the proposed model against the viewpoint difference using the Pers-SYM dataset, where the yaw angles between source and target images are different. The results show that supervised models implicitly learned the difference in the viewpoint and achieved high performance when the viewing angle was set the same degree; meanwhile, F_1 score meaning the accuracy of change detection decreases because the viewpoint gap increases. The proposed model degrades but moderately as the difference of view angles arises. Given that our F_1 score sometimes outperforms baselines when the yaw angle is shifted, our method is reasonably robust for view point change because it has learned the warping parameters explicitly. As shown in Fig 5, the change areas are defined as the difference in view angle increases while source images are reasonably warped.



(a) (b) (c) (d) (e)

Fig. 5. Examples of change detection from different camera angles over Pers-SYM dataset, where (a) target, (b) source of which camera angle is the same as target, (c) source of which yaw is shifted in the range of 0° to 10°, (d) 10° to 20°, and (e) 20° to 30°. (a) is ground truth and (b)-(e) are predicted. The first row, second row, third row represents input images, warped images, and change masks, respectively (The second row in (a) blank). Here, black areas in warped images represent out-of-view areas, where are not considered as a corresponding area in this study.

V. CONCLUSION

We proposed a self-supervised simultaneous alignment and change detection network (SACD-Net). The SACD-Net architecture can simultaneously optimize alignment and change detection between images captured from different viewpoints even with large or scattered changes. Moreover, the SACD-Net can be trained without the ground truth. To evaluate the effectiveness of the proposed method, we used two perspective change detection datasets, Pers-GSV and Pers-SYM. In the experiments, the SACD-Net showed competitive, or rather better performance compared to the supervised methods. The ablation study with viewpoint difference shows that our model performs competitively as supervised methods even in the case when the camera angle is quite different. In the future, for a deeper understanding of changes, we will focus on identifying which image detected changes comes from.

REFERENCES

[1] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch, "Fully convolutional siamese networks for change detection," in *ICIP*, 2018, pp. 4063–4067.

[2] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau, "High Resolution Semantic Change Detection," *arXiv preprint arXiv:1810.08452*, 2018.

[3] Harsh Jhamtani and Taylor Berg-Kirkpatrick, "Learning to describe differences between pairs of similar images," in *EMNLP*, 2018.

[4] Dong Huk Park, Trevor Darrell, and Anna Rohrbach, "Viewpoint invariant change captioning," *arXiv preprint arXiv:1901.02527*, 2019.

[5] Tomoyuki Suzuki et al., "Semantic Change Detection," in *ICARCV*, 2018.

[6] Ken Sakurada, Mikiya Shibuya, and Weimin Wang, "Weakly supervised silhouette-based semantic scene change detection," *ICRA*, 2020.

[7] Ryuhei Hamaguchi, Ken Sakurada, and Ryosuke Nakamura, "Rare event detection using disentangled representation learning," in *CVPR*, 2019, pp. 9327–9335.

[8] Y. Kawanishi I. Ide H. Murase M. Ukai N. Nagamine H. Mukojima, D. Deguchi and R. Nakasone, "Moving camera background-subtraction for obstacle detection on railway track," in *ICIP*, 2016.

[9] D.G.Lowe, "Distinctive image features from scale-invariant keypoints," in *IJCV*, November 2004.

[10] T. Tuytelaars H. Bay and L. Van Goo, "Surf: Speeded up robust features," in *In European Conference on Computer Vision*, May 2006.

[11] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *CVPR Workshops*, June 2018.

[12] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua, "Lift: Learned invariant feature transform," in *ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, Eds., 2016, pp. 467–483.

[13] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic, "Convolutional neural network architecture for geometric matching," in *CVPR*, 2017, pp. 6148–6157.

[14] Ignacio Rocco, Relja Arandjelović, and Josef Sivic, "End-to-end weakly-supervised semantic alignment," in *CVPR*, June 2018.

[15] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala, "DGC-Net: Dense Geometric Correspondence Network," in *WACV*, 2019, pp. 1034–1042.

[16] Daniel Crispell, Joseph Mundy, and Gabriel Taubin, "A Variable-Resolution Probabilistic Three-Dimensional Model for Change Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, 2012.

[17] Andres Huertas and Ramakant Nevatia, "Detecting Changes in Aerial Views of Man-Made Structures," in *ICCV*, 1998.

[18] Richard J Radke, Srinivas Andra, Omar Al-Kofahi, and Badrinath Roysam, "Image Change Detection Algorithms: A Systematic Survey," *TIP*, vol. 14, no. 3, 2005.

[19] Kunfeng Wang, Chao Gou, and Fei-Yue Wang, "M4CD: A Robust Change Detection Method for Intelligent Visual Surveillance," *arXiv preprint arXiv:1802.04979*, 2018.

[20] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid, "Deepflow: Large displacement optical flow with deep matching," in *ICCV*, 2013, pp. 1385–1392.

[21] Simon Stent, Riccardo Gherardi, Bjorn Stenger, and Roberto Cipolla, "Detecting Change for Multi-View, Long-Term Surface Inspection," in *BMVC*, 2015.

[22] Pablo F Alcantarilla, Simon Stent, German Ros, Roberto Arroyo, and Riccardo Gherardi, "Street-View Change Detection with Deconvolutional Networks," in *Robotics: Science and Systems*, 2016.

[23] Ken Sakurada and Takayuki Okatani, "Change Detection from a Street Image Pair using CNN Features and Superpixel Segmentation," in *BMVC*, 2015.

[24] Aito Fujita, Ken Sakurada, Tomoyuki Imaizumi, Riho Ito, Shuhei Hikosaka, and Ryosuke Nakamura, "Damage Detection from Aerial Umages via Convolutional Neural Networks," in *MVA*, 2017.

[25] Salman H Khan, Xuming He, Fatih Porikli, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri, "Learning Deep Structured Network for Weakly Supervised Change Detection," in *IJCAI*, 2017.

[26] Massimo Camplani, Lucia Maddalena, Gabriel Moyá Alcover, Alfredo Petrosino, and Luis Salgado, "A benchmarking framework for background subtraction in rgbd videos," in *International Conference on Image Analysis and Processing*, 2017.

[27] Kentaro Toyama, John Krumm, Barry Brumitt, and Brian Meyers, "Wallflower: Principles and practice of background maintenance," in *ICCV*, 1999.

[28] Enqiang Guo, Xinsha Fu, Jiawei Zhu, Min Deng, Yu Liu, Qing Zhu, and Haifeng Li, "Learning to measure change: Fully convolutional siamese metric networks for scene change detection," *arXiv preprint arXiv:1810.09111*, 2018.

[29] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, 2015.

[31] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox, "Flownet: Learning optical flow with convolutional networks," in *ICCV*, December 2015.

[32] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," *CVPR*, 2017.

[33] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," *CVPR*, 2016.

[34] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *CVPR*, 2018.

[35] Jie Chen, Ziyang Yuan, Jian Peng, Li Chen, Haozhe Huang, Jiawei Zhu, Tao Lin, and Haifeng Li, "Dasnet: Dual attentive fully convolutional siamese networks for change detection of high resolution satellite images," *arXiv:2003.03608*, 2020.

[36] Kento Doi, Ryuhei Hamaguchi, Shun Iwase, Rio Yokota, Yutaka Matsuo, and Ken Sakurada, "Epipolar-guided deep object matching for scene change detection," *arXiv preprint arXiv:2007.15540*, 2020.

[37] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[38] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.