Semi-Supervised Scene Change Detection by Distillation from Feature-metric Alignment

Seonhoon Lee
KAIST
shlee@rit.kaist.ac.kr

Jong-Hwan Kim
KAIST
johkim@rit.kaist.ac.kr

Abstract

Scene change detection (SCD) is a critical task for various applications, such as visual surveillance, anomaly detection, and mobile robotics. Recently, supervised methods for SCD have been developed for urban and indoor environments where input image pairs are typically unaligned due to differences in camera viewpoints. However, supervised SCD methods require pixel-wise change labels and alignment labels for the target domain, which can be both time-consuming and expensive to collect. To tackle this issue, we design an unsupervised loss with regularization methods based on the feature-metric alignment of input image pairs. The proposed unsupervised loss enables the SCD model to jointly learn the flow and the change maps on the target domain. In addition, we propose a semi-supervised learning method based on a distillation loss for the robustness of the SCD model. The proposed learning method is based on the student-teacher structure and incorporates the unsupervised loss of the unlabeled target data and the supervised loss of the labeled synthetic data. Our method achieves considerable performance improvement on the target domain through the proposed unsupervised and distillation loss, using only 10% of the target training dataset without using any labels of the target data.

1. Introduction

Scene change detection (SCD) has been attracting increasing attention as numerous emerging applications utilize SCD as the core task [24], *e.g.*, visual surveillance [12], anomaly detection [9], mobile robotics [8], remote sensing [19], and AR [30]. The goal of SCD is to localize changes in a given scene compared to the same scene at a different time. When the scene is provided as an image, the objective of SCD is to segment the changed objects or regions between a reference and a query image, which are captured at past and current times, respectively. Over the last few decades, researchers have developed various SCD methodologies for analyzing satellite images in remote sensing. More recently, there has been a growing interest in developing SCD methodologies for images captured in urban streets and indoor environments where autonomous vehicles or mobile robots are deployed.

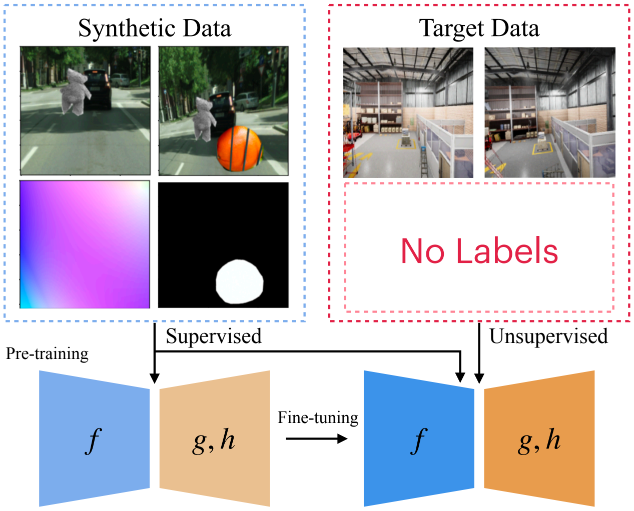

Figure 1. Our method is a semi-supervised SCD learning that uses readily available labeled synthetic data and unlabeled data from the target domain to enhance model performance on the target domain. The proposed method is practical since it does not require the expensive pixel-wise labels of the target domain. f represents the encoder and g, h are the decoders of the SCD model.

For the vehicular or mobile robotic applications of SCD, the input image pairs usually are not perfectly aligned due to different camera viewpoints and imperfect matching of two images [28]. Therefore, developing SCD methods robust to image pairs with unaligned viewpoints is crucial for practical use. Recent methods address the viewpoint variance problem by adopting correlation layers to consider viewpoint difference implicitly [33], utilizing a pre-trained optical flow model [3] or jointly learning optical flow and change detection using a synthetically generated dataset that has the ground truth labels for flow and change map estimation [28].

These methods are based on the supervised learning framework, which requires a dataset with pixel-wise change labels for the target domain of SCD. Gathering the correct pixel-wise change labels is a significant task that has a crucial impact on the performance of SCD. However, annotating precise pixel-wise change labels is labor-intensive, consuming considerable costs and human hours [33, 48]. Furthermore, supervised image alignment also requires labor-intensive annotations like flow maps, making the supervised SCD learning methods more challenging to apply to various target domains in the real world. Although an unsupervised loss has been proposed for performing warping parameter estimation and change detection for unaligned images based on image features [13], the reported performance indicates that learning solely from the proposed loss is challenging and does not match the SCD performance of supervised learning methods.

To address the issues mentioned above, we present two primary contributions. First, we propose unsupervised loss with regularization methods based on the feature-metric alignment for jointly learning the flow and change map. The proposed loss is designed based on the occlusion-aware photometric loss with two modifications: replacing the RGB-based photometric error with the multi-level feature dissimilarity and weighting by estimated change probability. Subsequently, we design a practical learning scheme that combines the unsupervised loss with the supervised loss of labeled synthetic data, which can be obtained inexpensively. Specifically, we propose a semi-supervised learning method based on a distillation loss between a teacher and a student network. This approach is applicable when labels are absent in the target domain of SCD, like in Fig. 1. Our learning method utilizes the labeled synthetic dataset presented in [28] and incorporates two key features. Firstly, we use the proposed unsupervised loss only for the teacher network. Secondly, we separate the parameters of the teacher decoder from the parameters of the student network. By employing this training strategy, we can train the teacher network to perform better on the target domain while training the student network with data augmentation to detect changes robustly.

In summary, our research makes the following contributions:

1. **Unsupervised Loss for SCD**: Inspired by unsupervised optical flow estimation, we propose a feature-metric loss and regularization methods that encourage the change-aware feature-metric alignment to simultaneously learn the flow and change maps to deal with unaligned image pairs in the target domain of SCD.

2. **Semi-Supervised SCD Learning**: We propose a semi-supervised SCD learning method based on a distillation loss using data augmentation, consolidating unsupervised loss for unlabeled target data and supervised loss for labeled synthetic data. Our proposed learning method enables the SCD model to learn more precise estimates of the target domain using the unsupervised loss while also learning feature representations that are more robust to illumination changes through data augmentation.

2. Related Work

2.1. Change Detection

Recent change detection methods leverage CNN models and have outperformed classical methods: Chen *et al*. [7] developed an attention ConvLSTM architecture that localizes changed region pixel-wise; Nguyen *et al*. [26] suggested utilizing pertinent features extracted from triplet CNN architecture for change detection; Lei *et al*. [20] introduced hierarchical paired channel fusion network that detects changed region based on hierarchical features via spatial attention; Sakurada *et al*. [33] presented CSCD-Net that estimates change pixels from the correlation between query and reference feature maps, and Park *et al*. [28] designed SimSaC architecture which jointly learns optical flow and change detection on the synthetically generated SCD dataset to deal with unaligned image pairs.

Those methods are all supervised learning frameworks that require costly human-annotated pixel-wise labels. An unsupervised SCD method for unaligned image pairs based on image feature differences was proposed in [13]. Still, its performance heavily relies on a semantic segmentation model. Sachdeva *et al*. [31] proposed a method utilizing synthetic change detection datasets generated by leveraging the pre-trained image inpainting model, but it performs a bounding-box-based change detection task. There are several unsupervised learning frameworks for satellite image change detection [2, 4, 5, 34, 35, 46, 47], but they are not proper for our scope of SCD since they assume almost perfect image alignment. In addition, recent research on semi-supervised learning for change detection in satellite imagery [1, 14, 21] has not yet addressed the lack of labeled data in the target domain.

2.2. Image Alignment

Image alignment is finding the correspondence between two images, allowing one image to be warped to align with the other. Among various techniques for correspondence problems, methods for optical flow and geometric image matching are most related to image alignment regarding the task of SCD. The goal of optical flow estimation is to find a flow field, a dense field of displacement vectors that represent the pixel movement required to align the matching pixels between two images. Optical flow estimation usually treats consecutive frames of a video. Dosovitskiy *et al*. developed a CNN-based model, FlowNet [10], that estimates a flow field based on the correlation tensor between two feature maps of an input image pair. Since the advent of FlowNet, most of the recently proposed methods have taken advantage of the correlation tensor [15, 16, 36, 37, 39, 40, 45]. These models utilize multiple correlation tensors computed from feature maps of different resolutions and focus on the flow map refinement method. Geometric image matching also aims for the same goal of optical flow estimation but mainly treats image pairs with massive geometric displacement. Melekhov *et al*. designed DGC-Net [25], which leverages a coarse-to-fine framework for computing correlation tensors considering large pixel displacements. Inspired by DGC-Net, Truong *et al*. developed GLU-Net [41] capable of handling any resolution of the image pairs with either small or large geometric displacement. Recently, PDC-Net [43], PDC-Net+ [42], and DKM [11] were proposed to solve the geometric image matching in a probabilistic manner to estimate the certainty of correspondences.

3. Methodology

Overview. Our goal is to develop a semi-supervised learning method for the SCD model that can effectively perform on the target domain, which utilizes unlabeled target data as well as labeled synthetic data that is inexpensive to collect. Specifically, our focus is on the scenario where we have ground truth labels of alignment and change for the synthetic dataset generated from public image datasets but have only reference-query image pairs without labels for the target domain to perform SCD. The scenario is practical in applying the SCD model to various target domains in the real world since it does not require alignment and change labels in each target domain, which are significantly expensive to obtain.

To address the scenario, we propose an unsupervised feature-metric loss that uses multi-level image features and regularization methods to train SimSaC [28], the state-of-the-art supervised SCD model, with unlabeled target data. In addition, we propose a distillation-based training method to increase the robustness of the SCD model to environmental changes, such as illumination changes.

3.1. Network Architecture and Supervised Loss

Architecture. To detect changes in unaligned image pairs, we employ the SimSaC architecture, which is designed to estimate both flow and change probability maps through coarse-to-fine refinement using image features from feature pyramid networks (FPN). SimSaC consists of three parts: the FPN encoder, denoted as f, and two decoders, denoted as g and h. g and h estimate the flow map $\hat{V} \in \mathbb{R}^{H\times W\times 2}$ and the binary change probability map $\hat{M} \in \mathbb{R}^{H\times W\times 1}$, respectively. These predictions are based on the multi-level features F_r and F_q extracted by f for a reference image $I_r \in \mathbb{R}^{H\times W\times 3}$ and a query image $I_q \in \mathbb{R}^{H\times W\times 3}$, respectively. SimSaC utilizes the estimated flow map to detect changes robustly to viewpoint differences between the reference and query images. The formulations are as follows:

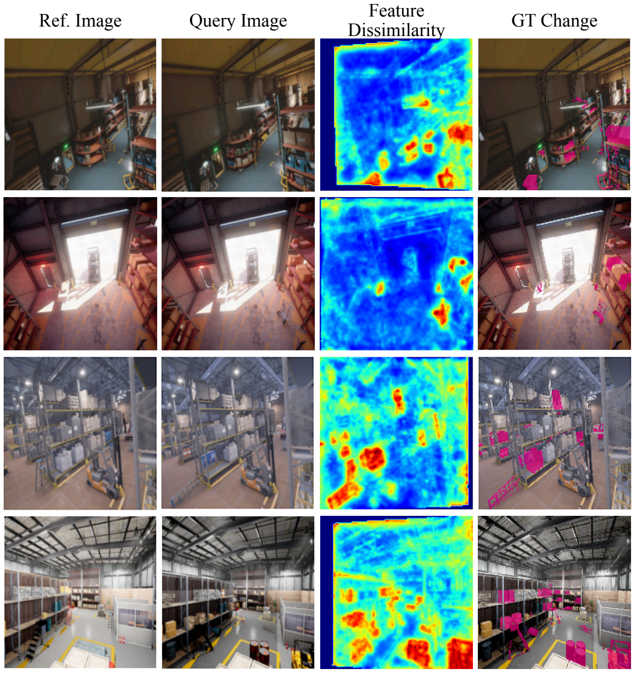

Figure 2. Feature dissimilarity maps between the warped reference image and the query image. Pixels with red color represent higher dissimilarity. The feature dissimilarity map contains valuable information for accurately estimating real changes.

$$\hat{V}_\theta = g_\theta(f_\theta(I_r), f_\theta(I_q)), \quad \hat{M}_\theta = h_\theta(f_\theta(I_r), f_\theta(I_q)), \quad (1)$$

where θ represents the parameters of the SCD model.

Supervised loss. For training the model with the labeled data, we use the hierarchical end-point error ℓ_V [28] and the hierarchical focal loss ℓ_M [28] (refer to the supplementary material), which are used for SimSaC, as follows:

$$\mathcal{L}_{sup} = \mathbb{E}_{(I_{r,q},V,M)\sim\mathcal{D}_l}\left[\ell_V(\hat{V}, V) + \ell_M(\hat{M}, M)\right], \quad (2)$$

where $I_{r,q}$ is an image pair, $\mathcal{D}_l$ is the labeled dataset, V is the ground truth flow map, and M is the ground truth change mask. Unlike [28], which used both synthetic dataset and target dataset for $\mathcal{L}_{sup}$, we compute $\mathcal{L}_{sup}$ only for the synthetic dataset.

3.2. Feature-metric Loss and Regularization

We propose unsupervised loss that consists of the feature-metric loss inspired by unsupervised optical flow estimation and two regularization methods to complement the feature-metric loss. The key aspect of unsupervised optical flow learning is minimizing photometric loss between two images, which encourages the flow map to align the images with a similar appearance. The photometric loss is often calculated as the Charbonnier [38], SSIM [44], or Census loss [22] of the RGB values between the pixels of the same location in the query image and the reference image warped by the estimated flow map. For computing the photometric loss, the occluded pixels are excluded by the occlusion mask, estimated with heuristic methods, since they do not have a correspondence in the image pairs.

Feature-metric loss. We propose the feature-metric loss based on the formulation of the photometric loss. The feature-metric loss encourages the change-aware feature-metric alignment of input image pairs. In this loss, the RGB-based error and the occlusion mask of the photometric loss are replaced by the multi-level feature dissimilarity and the change probability map, respectively, as follows:

$$\begin{gathered}\ell_{feat}(\hat{M},\hat{V}) = \frac{\sum_i (1-\hat{M}_i)(1-s(\omega(\bar{F}_r,\hat{V}),\bar{F}_q)_i)}{\sum_i (1-\hat{M}_i)},\\ s(F_r,F_q)_i = \prod_l \frac{1+\cos(F^l_{r,i},F^l_{q,i})}{2},\end{gathered} \tag{3}$$

where i is the index of each pixel location, $\omega(\cdot)$ is the differentiable warping function, *e.g.* bilinear sampling [17], l is the layer level of the multi-level feature maps F_r and F_q, $\cos(\cdot)$ is the cosine similarity, and the bar above the letter represents gradient stopping. Note that we match the size of $\hat{M}$, $\hat{V}$ and F^l to the size of F^1 by downsampling $\hat{M}$, $\hat{V}$, and upsampling F^l for $1 < l \leq L$.

In contrast to optical flow estimation, which handles consecutive frames of video, there may be a considerable difference in time between the reference and query images in SCD. This difference in time can result in appearance variation of the query image caused by environmental changes, *e.g.* changes in day/night cycles, weather, or other illumination conditions. Therefore, instead of RGB-based errors, we use an error based on image features robust to environmental changes. For the feature-based error, we utilize the multi-level feature dissimilarity $1-s$, where s is computed by multiplying all feature similarity maps from each FPN level to consider both low-level and high-level correspondence simultaneously.

The proposed feature-metric loss ℓ_{feat} is the weighted average of the feature dissimilarity between the F^r warped by $\hat{V}$ and F^q, weighted by the estimated probability of 'not changed' $(1-\hat{M})$. The weighted average is to prevent the trivial solution $\hat{M}=1$. Minimizing ℓ_{feat} encourages $\hat{V}$ to align the reference-query pair with viewpoint difference and to increase $\hat{M}$ for regions with high dissimilarity between the aligned reference and query features. Fig. 2 illustrates how the multi-level feature dissimilarity map is effective in estimating genuine changes. Note that we do not update F_r and F_q by the gradient of ℓ_{feat} to prevent the encoder from overfitting.

Change regularization. Training the SCD model solely based on ℓ_{feat} may result in overconfident change probability since ℓ_{feat} does not encourage low change probability on pixels with low feature dissimilarity. To address this, we impose a regularization on the estimated change map by taking the negative log, which prevents the overconfident change probability, as follows:

$$\begin{gathered}\ell_{cr}(\hat{M}) = -\frac{1}{HW}\sum_i \bar{s}_i \log(1-\hat{M}_i),\\ \bar{s} = \texttt{stop\_grad}(s(\omega(F_r,\hat{V}),F_q)).\end{gathered} \tag{4}$$

ℓ_{cr} imposes relatively weak regularization on the change probability of pixels with low feature similarity while regulating the change probability to be low for pixels with high feature similarity. Therefore, it effectively complements the blind spot of ℓ_{feat}.

Edge-aware smoothness. Edge-aware smoothness regularization [18, 22] is a widely used technique in unsupervised optical flow estimation. This aims to enhance the model's ability to learn object motion boundaries with finer detail by regulating the flow map to adapt to edges in the image. Since the change probability map should also be estimated to fit the boundaries of the changed objects, we propose an edge-aware smoothness regularization for the change probability map as follows:

$$\ell_{sm}(\hat{M}) = \frac{1}{HW}\sum_i (1-\bar{s}_i)\left(\sum_n \ell_{bce}(\frac{\partial \hat{M}_i}{\partial x_n}, E_i)\right), \tag{5}$$

where ℓ_{bce} is the binary cross-entropy loss, and E is an edge mask of an image whose elements are 1 for the edge pixels and 0 for the others. We generate the edge mask using the Canny edge detector.

The conventional smoothness regularization for optical flow only penalizes non-edge pixels with a high flow gradient. Unlike the conventional method, we impose an additional penalty on pixels at edges if the change gradients at those pixels are not close to 1 by applying the binary cross-entropy between the change gradient and the edge mask. Since this additional penalty should only be applied to pixels predicted to be part of changed regions, we weight the regularization with the feature dissimilarity.

Unsupervised loss. The overall unsupervised loss for the unlabeled target data is defined as follows:

$$\mathcal{L}_{unsup} = \mathbb{E}_{I_{r,q}\sim\mathcal{D}_u}\left[\ell_{feat} + \lambda_1\ell_{cr} + \lambda_2\ell_{sm}\right], \tag{6}$$

where $\mathcal{D}_u$ is the unlabeled dataset, λ_1 and λ_2 are weight coefficients for each regularization.

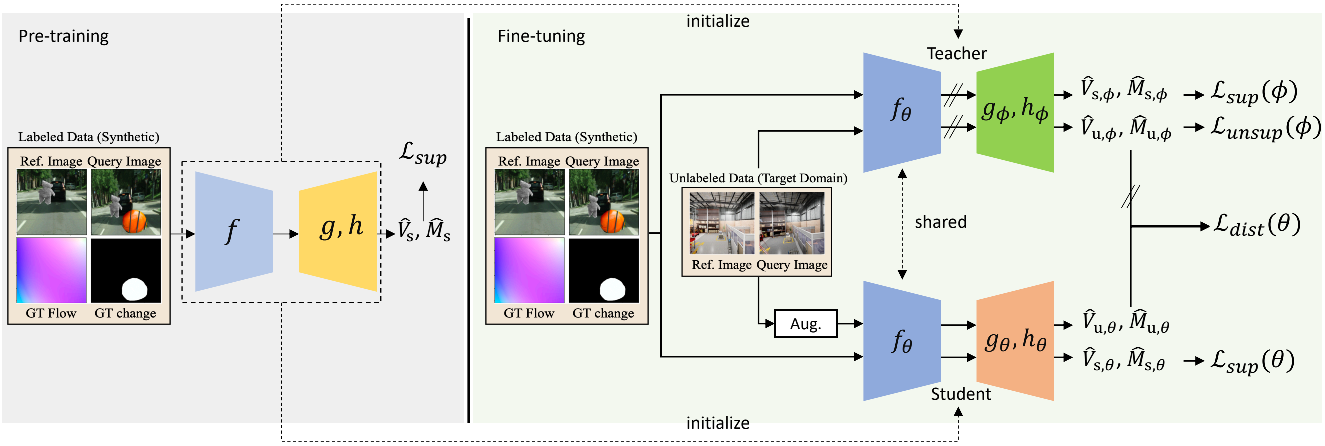

Figure 3. The overview of the proposed semi-supervised learning method. The teacher network utilizes unlabeled target data for unsupervised learning to learn more accurate estimates for the target domain. The student network becomes a robust model by training on augmented unlabeled data using the teacher's estimates as pseudo-labels.

3.3. Semi-Supervised SCD Learning

Distillation loss. When using the proposed unsupervised loss, it is important that the features extracted from the image are robust to environmental changes, such as changes in illumination. However, the unsupervised loss is only used to learn flow and change maps, not features, so it relies on the robustness of the pre-trained features. Therefore, to further improve the robustness of the feature against environmental changes, we use a distillation loss as an additional training objective based on the student-teacher structure. We define the distillation loss using ℓ_V and ℓ_M as follows:

$$\mathcal{L}_{dist}(\theta) = \mathbb{E}_{(I_{r,q},\tilde{I}_{r,q})\sim\mathcal{D}_u}\left[\ell_V(\hat{V}_\theta,\hat{V}_\phi) + \ell_M(\hat{M}_\theta,\hat{M}_\phi)\right], \tag{7}$$

where $\tilde{I}_{r,q}$ is the photometrically augmented $I_{r,q}$, θ and ϕ represent the student network parameters and the teacher decoder parameters, respectively. During the training, the teacher network generates estimates of the original unlabeled data, and the student network learns to predict correct outputs for the augmented unlabeled data by leveraging the teacher's estimation as pseudo-labels.

Training scheme. We propose a semi-supervised SCD learning method for the robust model based on the student-teacher structure, which jointly minimizes $\mathcal{L}_{sup}$, $\mathcal{L}_{unsup}$, and $\mathcal{L}_{dist}$. The proposed training method has two characteristics to utilize the unsupervised loss and the distillation loss effectively. First, the supervised loss is used for training both the student and teacher networks, while the unsupervised loss is used only for the teacher network. This asymmetry is to prevent the student decoder from being trained incorrectly by feature-metric alignment based on image features that are not robust to augmentation. Second, the decoder parameters of the student and the teacher are separated. By separating the decoder parameters, we encourage the teacher to learn accurate flow and change estimation on the target domain based on the unsupervised change-aware feature-metric alignment and the student to learn flow and change estimation robust to the photometric augmentation based on the distillation loss. We share the parameters of the encoder so that the teacher also utilizes the robust feature representation learned by the distillation loss. Fig. 3 represents the proposed learning scheme.

With indicating which specific loss term contributes to the learning of individual parameters, the total loss is defined as follows:

$$\mathcal{L}(\theta,\phi) = \mathcal{L}_{sup}(\theta)+\mathcal{L}_{dist}(\theta)+\mathcal{L}_{sup}(\phi)+\mathcal{L}_{unsup}(\phi). \tag{8}$$

The proposed learning method enables the teacher network to estimate precise change maps by the unsupervised loss. At the same time, the student network learns robust change estimation by training on strongly augmented target data, distilling the knowledge from the teacher's estimates based on the feature-metric alignment.

4. Experiments

4.1. Datasets and Experimental Setup

Datasets. To evaluate the effectiveness of our proposed method, we utilized two target datasets: ChangeSim [27] and PCD [32]. ChangeSim consists of image sequences collected from 10 different photo-realistic simulated warehouse environments, 6 of which are used for training and 4 for testing. The data for each warehouse contains *normal*, which is the base data, and *dusty-air* and *low-illumination*, which have visual variations. Following the convention in [27], we only use the training split of *normal* (13,221 pairs) for model learning and evaluate the model on the test splits of *normal*, *dusty-air*, and *low-illumination*, each

Labeled Data	Unlabeled Data	Method	ChangeSim			PCD	
			Normal	Dusty-air	Low-illum.	Tsunami	GSV
Synthetic	-	SimSaC [28]	57.8	29.4	26.4	75.1	31.2
		SimSaC†	57.8	29.2	32.8	64.5	57.4
Synthetic & Target (10%)	-	SimSaC†	66.0	42.3	39.3	73.5	62.5
Synthetic	Target (10%)	Ours	**66.6**	**52.0**	**43.9**	**85.2**	**78.2**
Target (100%)	-	CDNet++ [29]	-	-	-	86.0	68.0
		HPCFNet [20]	-	-	-	86.8	77.6
		CSCDNet [33]	32.6	15.5	13.5	75.7	69.1
		DR-TANet [6]	40.2	22.0	17.7	74.1	68.6
Synthetic & Target (100%)	-	SimSaC [28]	66.5	**56.8**	42.7	**90.4**	**78.4**
		SimSaC†	**70.2**	50.5	**44.6**	87.6	75.7
Synthetic	Target (100%)	Ours	68.3	54.1	44.5	86.4	78.4

Table 1. Quantitative results of the comparative study on the target datasets. The target datasets are ChangeSim and PCD, respectively. The baselines are the supervised model trained with only the Synthetic dataset or both the Synthetic and the labeled target datasets. For evaluating our method, we do not use the labels of the target datasets. The percentage value in parentheses means the ratio of the number of the used image pairs for training to the size of the entire target training set. SimSac† is our implementation of SimSaC. The best two results are marked in bold and underlined, respectively.

of them has about 4,200 pairs of images. PCD contains 100 panoramic image pairs of street views (GSV) and another 100 pairs of post-tsunami environments (TSUNAMI). Based on the setting in [33], we collect 60 cropped images per panoramic image by sliding and plane rotation, resulting in 9,600 image pairs for training and 2,400 pairs for evaluation.

We utilized the synthetic dataset introduced in [28] as the labeled dataset for the proposed semi-supervised learning method. The change and flow labels of the Synthetic dataset are generated by randomly compositing foreground object images to background images using the cut-paste method and by applying random geometric transformations to the composited images, respectively. Following the instruction in [28], we randomly sampled and exploited the same number of data samples in the synthetic dataset as for the number of the target dataset.

Evaluation. We use the F1-score, the harmonic mean of precision and recall, of change detection as the evaluation metric. We set the baseline as the supervised SCD models trained with the Synthetic dataset alone or both the Synthetic dataset and the target dataset with GT labels. When using both the Synthetic and the labeled target datasets, we follow the training schedule used in SimSaC. For SimSaC, which is the state-of-the-art supervised SCD model, we present both the performance reported in [28] and the performance of our implementation of SimSaC, SimSaC†.

Implementation. We followed the training protocol of SimSaC and pre-trained the network on the synthetic dataset for 25 epochs before fine-tuning the network using the proposed semi-supervised learning approach. We optimize the model for 15 epochs by the AdamW optimizer [23] with a learning rate decay of 4×10^{-4}. We employed a batch size of 8 for both the labeled synthetic and unlabeled target datasets. The learning rate is 2×10^{-5} and halves at epoch 3, epoch 6, and epoch 9. Every image of the datasets is resized to 520×520 for training. λ_1 is 0.2 for ChangeSim and 0.01 for PCD. λ_2 is set to 1 for all datasets. For data augmentations, we use random color jittering, random gray scaling, random shadowing, random brightness contrast, and random sun-flare, which are all used for the original SimSaC. We implement our method using Pytorch on a 4.2GHz i7 CPU desktop with a single NVIDIA RTX 3090 GPU.

4.2. Comparative Study

Table 1 shows the quantitative comparison results on the ChangeSim and PCD datasets.

Baselines trained with Synthetic. We first compared the performance of our proposed semi-supervised learning method with baseline supervised models trained on the synthetic dataset. This experiment covers the scenario of performing SCD on target domains where GT labels do not exist, which is our main concern. In this experiment, to demonstrate the efficiency and effectiveness of the proposed method, we randomly sampled only 10% of training samples from each unlabeled target training dataset and used them for our method, assuming the scarcity of the unlabeled target dataset. The results show that our learning method achieves considerable performance improvement compared to the baselines on all benchmarks. Compared to the baseline on ChangeSim, our method improves the F1-score by 15.2%, 76.8%, and 33.8% for *normal*, *dusty-air*, and *low-illum.*, respectively. Our method also improves performance by 13.4% and 36.2% for TSUNAMI and GSV, respectively. In addition, we compared our method with SimSaC† trained on both the synthetic dataset and the 10% of the target training dataset with GT labels. Interestingly, our method performs better than the baseline in such a challenging condition. This suggests that our learning scheme is highly effective in addressing scenarios where the target dataset is scarce and the ground truth labels are unavailable.

Baselines trained with the target labels. To investigate the extent to which the proposed learning method can replace GT target labels, we compared the performance of our method to the performance of supervised baseline models trained fully exploiting the labels of the target training dataset. As same as the previous experiment, we also did not use any labels of the target datasets for our method. Instead, we used all the unlabeled training data, not just 10% of the dataset. This comparison setting is highly challenging since the baselines fully utilize the labels of all the data. Our method performs slightly worse on ChangeSim *normal*, ChangeSim *dusty-air*, and TSUNAMI of PCD by 2.7%, 4.7%, and 4.4% compared to the state-of-the-art supervised model, respectively, and achieves almost same performance on ChangeSim *low-illum.* and GSV of PCD. These results again demonstrate the effectiveness of the proposed semi-supervised SCD learning method. Fig. 4 presents the qualitative results of our method compared with the supervised baseline.

4.3. Ablation Study

In all ablation experiments, we used 10% of the target training dataset on each domain. We configured λ_2 as 1 and λ_1 as 0.2 for ChangeSim and 0.01 for PCD.

Configuration	ChangeSim			PCD	
	Normal	Dusty-air	Low-illum.	TSUNAMI	GSV
ℓ_{feat}	52.5	25.9	20.2	83.1	75.1
+ ℓ_{cr}	65.2	**37.2**	33.8	84.0	76.1
+ ℓ_{sm}	64.0	29.5	26.6	84.8	77.0
+ ℓ_{cr}, ℓ_{sm}	**65.4**	35.2	**37.5**	**85.1**	**77.5**

Table 2. Ablation study on the regularization methods of unsupervised loss

Regularization methods for unsupervised loss. Table 2 presents the performance of four different unsupervised learning settings without $\mathcal{L}_{dits}$ to exclude the effect of the distillation. Both regularization methods are designed to suppress excessive change estimation, with ℓ_{cr} serving to lower the change estimation threshold and ℓ_{sm} ensuring that the estimated change region equals the object area inside the estimated change region. Both for the ChangeSim dataset and the PCD dataset, using both the change regularization and the smoothness regularization methods shows the best performance except for *dusty-air*, and not using both regularization methods shows the worst performance. The results show that the proposed regularization methods effectively compensate for the blind spot of ℓ_{feat}.

Configuration	ChangeSim			PCD	
	Normal	Dusty-air	Low-illum.	TSUNAMI	GSV
$\mathcal{L}_{sup}$	57.8	29.2	32.8	64.5	57.4
+ $\mathcal{L}_{dist}$	60.1	36.9	35.5	50.8	47.6
+ $\mathcal{L}_{unsup}$	65.4	35.2	37.5	85.1	77.5
+ $\mathcal{L}_{unsup}, \mathcal{L}_{dist}$	**66.6**	**52.0**	**43.9**	**85.2**	**78.2**

Table 3. Ablation study on the loss terms of the proposed semi-supervised learning method

Loss terms of semi-supervised learning. Table 3 demonstrates the effectiveness of each loss term of the proposed semi-supervised learning method. By utilizing both $\mathcal{L}_{unsup}$ and $\mathcal{L}_{con}$, the model accomplishes the best performance both on ChangeSim and PCD. Training $\mathcal{L}_{con}$ alone shows improvement on ChangeSim but records the worst performance on PCD, which implies the distillation loss alone can not deal with the domain gap between Synthetic and the target. By using $\mathcal{L}_{unsup}$ alone for training, the model achieves fairly high performance, but there is no significant performance improvement when visual variations occur due to factors such as illumination and air turbidity.

Configuration	ChangeSim		
	Normal	Dusty-air	Low-illum.
Shared	65.6	47.1	41.7
Separated	66.6	52.0	43.9

Table 4. Ablation study on the decoder parameter settings

Parameter separation. Table 4 represents the effectiveness of the parameter separation for robustness to environmental changes of the model. The results show that the teacher and student networks perform better when the decoder parameters are separated per our design intent.

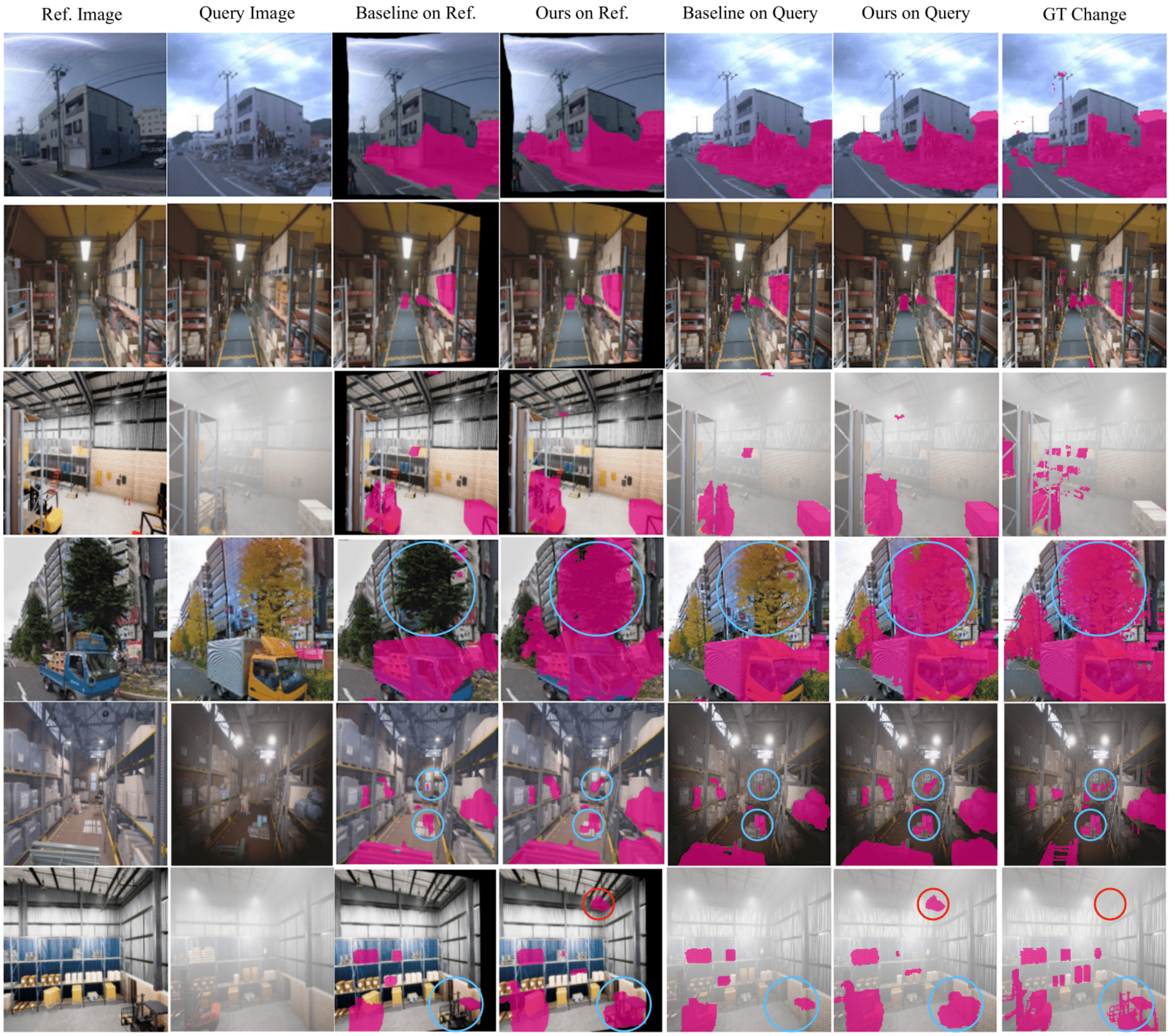

Figure 4. Qualitative results of the proposed method. The baseline is the supervised SimSaC trained with the target data fully using the ground truth labels. Our method effectively learns to estimate actual changes without using any target labels. Red circles show the incorrectly estimated change. Blue circles represent the actual areas of change that the baseline does not predict, while our method does.

5. Conclusion

In this paper, we proposed a practical and effective semi-supervised SCD method that incorporates the supervised loss of the labeled synthetic data and the unsupervised loss of the unlabeled target data. To enable training of the SCD model without using target labels, we made the following key contributions: (1) developing a feature-metric loss that allows for joint learning of change detection and alignment by leveraging change-aware feature-metric alignment, (2) designing two regularization methods that compensate for the feature-metric loss, and (3) proposing a semi-supervised learning approach that trains a teacher and a student network with separate parameters, with the teacher network trained using the unsupervised loss. Our method demonstrated its effectiveness by achieving comparable performance to the supervised model on the target domain, all without relying on any target data labels.

Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by MSIT (South Korea) (No.2020-0-00440).

References

[1] Wele Gedara Chaminda Bandara and Vishal M Patel. Revisiting consistency regularization for semi-supervised change detection in remote sensing images. *arXiv preprint arXiv:2204.08454*, 2022. 2

[2] Luca Bergamasco, Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised change detection using convolutional-autoencoder multiresolution features. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 2

[3] Shuhui Bu, Qing Li, Pengcheng Han, Pengyu Leng, and Ke Li. Mask-cdnet: A mask based pixel change detection network. *Neurocomputing*, 378:166–178, 2020. 1

[4] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021. 2

[5] Jie Chen, Ziyang Yuan, Jian Peng, Li Chen, Haozhe Huang, Jiawei Zhu, Yu Liu, and Haifeng Li. Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1194–1206, 2020. 2

[6] Shuo Chen, Kailun Yang, and Rainer Stiefelhagen. Dr-tanet: Dynamic receptive temporal attention network for street scene change detection. In *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021. 6

[7] Yingying Chen, Jinqiao Wang, Bingke Zhu, Ming Tang, and Hanqing Lu. Pixelwise deep sequence learning for moving object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9):2567–2579, 2017. 2

[8] Erik Derner, Clara Gomez, Alejandra C Hernandez, Ramon Barber, and Robert Babuška. Change detection using weighted features for image-based localization. *Robotics and Autonomous Systems*, 135:103676, 2021. 1

[9] Keval Doshi and Yasin Yilmaz. Any-shot sequential anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 934–935, 2020. 1

[10] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015. 3

[11] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 3

[12] Zhihang Fu, Yaowu Chen, Hongwei Yong, Rongxin Jiang, Lei Zhang, and Xian-Sheng Hua. Foreground gating and background refining network for surveillance object detection. *IEEE Transactions on Image Processing*, 28(12):6077–6090, 2019. 1

[13] Yukuko Furukawa, Kumiko Suzuki, Ryuhei Hamaguchi, Masaki Onishi, and Ken Sakurada. Self-supervised simultaneous alignment and change detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6025–6031. IEEE, 2020. 2

[14] Fan Hao, Zong-Fang Ma, Hong-Peng Tian, Hao Wang, and Di Wu. Semi-supervised label propagation for multi-source remote sensing image change detection. *Computers & Geosciences*, 170:105249, 2023. 2

[15] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8981–8989, 2018. 3

[16] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2462–2470, 2017. 3

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 4

[18] Rico Jonschkowski, Austin Stone, Jonathan T Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 557–572. Springer, 2020. 4

[19] Lazhar Khelifi and Max Mignotte. Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis. *Ieee Access*, 8:126385–126400, 2020. 1

[20] Yinjie Lei, Duo Peng, Pingping Zhang, Qiuhong Ke, and Haifeng Li. Hierarchical paired channel fusion network for street scene change detection. *IEEE Transactions on Image Processing*, 30:55–67, 2020. 2, 6

[21] Lian Liu, Danfeng Hong, Li Ni, and Lianru Gao. Multilayer cascade screening strategy for semi-supervised change detection in hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1926–1940, 2022. 2

[22] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Ddflow: Learning optical flow with unlabeled data distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8770–8777, 2019. 4

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[24] Murari Mandal and Santosh Kumar Vipparthi. An empirical review of deep learning frameworks for change detection: Model design, experimental frameworks, challenges and research needs. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 1

[25] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala. Dgc-net: Dense geometric correspondence network. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1034–1042. IEEE, 2019. 3

[26] Tien Phuoc Nguyen, Cuong Cao Pham, Synh Viet-Uyen Ha, and Jae Wook Jeon. Change detection by training a triplet network for motion feature extraction. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(2):433–446, 2018. 2

[27] Jin-Man Park, Jae-Hyuk Jang, Sahng-Min Yoo, Sun-Kyung Lee, Ue-Hwan Kim, and Jong-Hwan Kim. Changesim: towards end-to-end online scene change detection in industrial indoor environments. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8578–8585. IEEE, 2021. 5

[28] Jin-Man Park, Ue-Hwan Kim, Seon-Hoon Lee, and Jong-Hwan Kim. Dual task learning by leveraging both dense correspondence and mis-correspondence for robust change detection with imperfect matches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13749–13759, 2022. 1, 2, 3, 6

[29] K Ram Prabhakar, Akshaya Ramaswamy, Suvaansh Bhambri, Jayavardhana Gubbi, R Venkatesh Babu, and Balamuralidhar Purushothaman. Cdnet++: Improved change detection with deep neural network feature correlation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 6

[30] Christopher Reardon, Jason Gregory, Carlos Nieto-Granda, and John G Rogers III. Designing a mixed reality interface for autonomous robot-based change detection. In *Virtual, Augmented, and Mixed Reality (XR) Technology for Multi-Domain Operations II*, volume 11759, pages 136–143. SPIE, 2021. 1

[31] Ragav Sachdeva and Andrew Zisserman. The change you want to see. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3993–4002, 2023. 2

[32] 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)*, volume 61, pages 1–12, 2015. 5

[33] Ken Sakurada, Mikiya Shibuya, and Weimin Wang. Weakly supervised silhouette-based semantic scene change detection. In *2020 IEEE International conference on robotics and automation (ICRA)*, pages 6861–6867. IEEE, 2020. 1, 2, 6

[34] Minseok Seo, Hakjin Lee, Yongjin Jeon, and Junghoon Seo. Self-pair: Synthesizing changes from single source for object change detection in remote sensing imagery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6374–6383, 2023. 2

[35] Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE transactions on geoscience and remote sensing*, 60:1–16, 2021. 2

[36] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1599–1610, 2023. 3

[37] Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. Smurf: Self-teaching multi-frame unsupervised raft with full-image warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3887–3896, 2021. 3

[38] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2432–2439. IEEE, 2010. 4

[39] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8934–8943, 2018. 3

[40] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 3

[41] Prune Truong, Martin Danelljan, and Radu Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6258–6268, 2020. 3

[42] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[43] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2021. 3

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[45] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022. 3

[46] Bin Yang, Le Qin, Jianqiang Liu, and Xinxin Liu. Utrnet: An unsupervised time-distance-guided convolutional recurrent network for change detection in irregularly collected images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. 2

[47] Wenhua Zhang, Licheng Jiao, Fang Liu, Shuyuan Yang, Wei Song, and Jia Liu. Sparse feature clustering network for unsupervised sar image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 2

[48] Aleksandar Zlateski, Ronnachai Jaroensri, Prafull Sharma, and Frédo Durand. On the importance of label quality for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1479–1487, 2018. 2