Letter

Object Helps U-Net Based Change Detectors

Lan Yan, Qiang Li, and Kenli Li

Dear Editor,

This letter focuses on leveraging the object information in images to improve the performance of the U-Net based change detector. Change detection is fundamental to many computer vision tasks. Although existing solutions based on deep neural networks are able to achieve impressive results. However, these methods ignore the extraction and utilization of the inherent object information within the image. To this end, we propose a simple but effective method that employs an excellent object detector to extract object information such as locations and categories. This information is combined with the original image and then fed into the U-Net based change detection network. The successful application of our method on MU-Net and the experimental results on CDnet2014 dataset show the effectiveness of the proposed method, and the correct object information is helpful in change detection.

Change detection is a crucial task in visual perception [1], [2] and video analytics [3], serving as a foundational bedrock for a diverse array of computer vision applications including but not limited to action recognition, traffic monitoring, and object tracking. It is usually employed as a pre-processing step to provide focus of attention for classification, tracking, and behavior analysis, etc. Change detection is challenging due to intricate factors such as the presence of cluttered backgrounds, perturbations introduced by camera motion during video acquisition, fluctuations in weather conditions, and variations in illumination settings. A common approach for change detection is to perform background subtractions, i.e., to consider the changing region of interest as foreground pixels and the non-changing parts as background pixels. This binary classification problem has garnered substantial attention from the research community over the years, and some effective solutions are proposed. Especially in recent years, owes to the progress of deep learning, the availability of large-scale change detection datasets [4], and the development of hardware computing capabilities and parallel processing technologies [5], the change detectors based on deep neural network have achieved impressive results.

The majority of extant change detection methods [6]–[11] based on deep neural networks predominantly leverage a single image as input, and then use elaborate network to directly segment the foreground objects. However, they neglect the mining and utilization of the target information in the video frame. Considering that change detection requires the precision segmentation of foreground objects, and the video frame inherently contains the category and location information of objects, we think that these information are helpful to improve the performance of change detector.

Currently, the change detectors based on deep neural networks mainly include three categories. The first category is multiple network models [6], [11], where FgSegNet [11] is a representative approach that trains a model for each video on the dataset and performs testing separately. The second is dual/two networks models [7], [8], which are mostly based on generative adversarial networks (GAN) to realize change detection through adversarial learning of generator and discriminator. The third is single network models [9], [10], which mainly design network structures combining advanced convolutional neural networks such as U-Net and ResNet and train one single model only on the dataset. U-Net contains a large number of long connections, so that the features can better relate to the original information of the input image, which helps to restore the information loss caused by down-sampling. At this point, we think that U-Net is essentially similar to the residual connection. Therefore, in the single network model, U-Net becomes an excellent method for change detection. Besides, dual/two networks models (e.g., BSPVGAN [7]) and multiple network models (e.g., FgSegNet [11]) have shown commendable performance achievements. Particularly, our prior work BSPVGAN [7] converges the principle of Bayesian networks with generative adversarial networks, conceiving change detection as a classification problem under probability, and it is a state-of-the-art. Thus, in this letter, we mainly focus on U-Net based change detectors and discuss how to improve these models with the help of object's information within the given image.

To this end, we propose a new change detection method, which capitalizes on object information such as the spatial location and semantic class to heighten model performance. Considering that the goal of object detection is to locate and identify the objects in an image, this is exactly consistent with the object information we need. Therefore, we leverage the object detector to extract the object information. The acquisition of these information is easy because it is automatically generated by a pre-trained object detector, without the need for dedicated training on the change detection dataset. Notably, MU-Net2 [9] also extracts additional information from video frames without supervising, but these information are pixel-level cues derived from optical flow motion and classical background reduction algorithms. In contrast, our approach focuses on the object information and employs the boundingbox-level cues.

MU-Net [9] is currently the best U-Net based change detection model, which contains two versions. Except MU-Net2 which combines spatiotemporal cues, MU-Net1 only takes images as input. Thus, MU-Net1 becomes a natural choice for incorporating object information to verify the effectiveness of our method. It is worth noting that to streamline exposition, the subsequent reference to MU-Net1 will be denoted as MU-Net for conciseness. Experimental results on the CDnet2014 dataset show that our method can improve the performance of U-Net based change detection model. In particular, as shown in Fig. 1, the object information can make the U-Net based change detection model pay more attention to the foreground object and reduce the negative effects of cluttered background, thus contributing to the improvement of change detection performance.

Proposed approach: Our method aims at fully mining and utilizing the object information in the given image to improve the performance of the U-Net based change detector. Fig. 2 displays the comparison of our approach with the popular U-Net based change detection approach. As shown in the Fig. 2, different from popular methods that only use RGB images as input, our method introduces masks with object information in addition to the original images. Benefiting from the active research on object detection in the field of computer vision, we adopt off-the-shelf object detectors to obtain the position and category information of the foreground object and generate a mask. Specifically, we first use the object detector to detect all images in a video. After the detection results are obtained, we remove the wrongly detected bounding boxes and judge the foreground objects based on the temporal cues of the video frames. According to the bounding boxes of the foreground objects, we obtain the masks required for subsequent processing.

For each input image, its corresponding foreground target mask can be generated by an excellent pre-trained object detector. After that, we concatenate the RGB images and their corresponding masks, then feed the results into an U-Net based change detection network. This coupling stands as one of the most straightforward way to lever-

Corresponding author: Lan Yan.

Citation: L. Yan, Q. Li, and K. Li, "Object helps U-Net based change detectors," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 2, pp. 548–550, Feb. 2024.

L. Yan and K. Li are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: ylan@hnu.edu.cn; lkl@hnu.edu.cn).

Q. Li is with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: qiangli.casia@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2023.124032

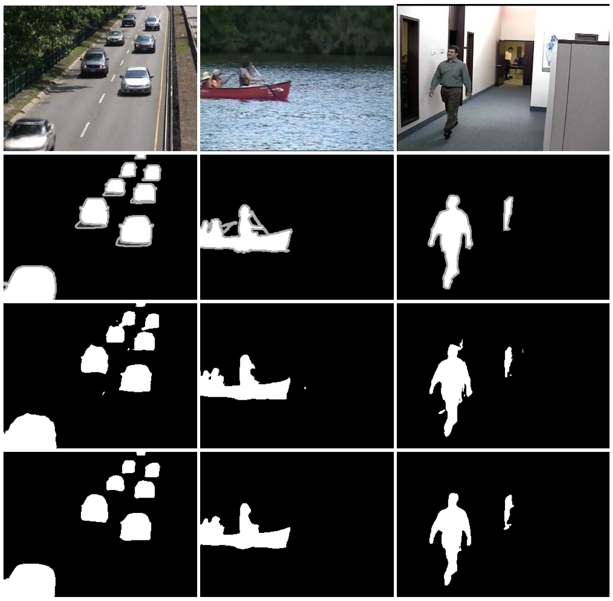
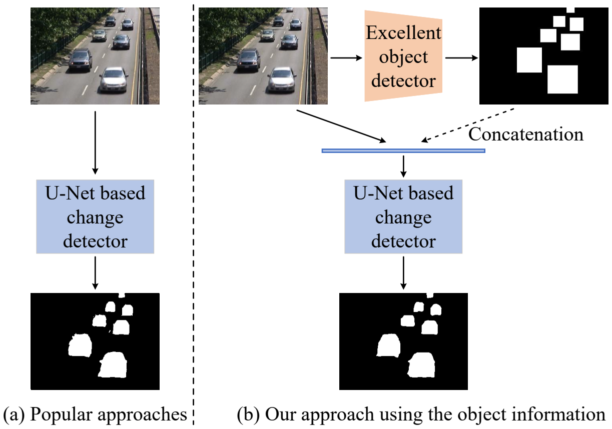


Fig. 1. Examples of change detection results. From top to bottom are the input images, ground truth, the results of original MU-Net, and results of the improved MU-Net based on our method, sequentially.


Excellent object detector
Concatenation
U-Net based change detector
U-Net based change detector

(a) Popular approaches

(b) Our approach using the object information

Fig. 2. Comparison of our approach with the popular U-Net based change detection approaches. (a) The popular approaches takes only RGB images as input; (b) Our approach takes advantage of the object information within the given image, which is provided by an excellent object detector.

age object information. There are two main reasons why we choose this option. First, using this scheme only requires a small adjustment to the structure of the input side of the network, so that our method can be flexibly applied to a variety of U-Net based change detectors. Second, due to the existence of long connections in U-Net, the original input information can be well utilized in the network. This, in turn, obviates the need for redundant reiteration of input information within the network's intermediate layers.

Our approach leverages the object information gleaned from the object detector to make the network pay more attention to foreground objects. It is simple but effective. Nonetheless, the potency of our approach relies on the target detector’s precision in localizing and identifying objects, as well as accurately categorizing them as foreground entities. Any inaccuracies in these aspects can potentially cast an influence over the change detection performance. To obviate the deleterious impact of erroneous object information and its potential to misguide the change detector, we institute a bounding box selection strategy. In adherence to the temporal cues ingrained within video frames, we selectively eschew bounding boxes displaying tenuous correlations. This process culminates in the retention solely of bounding boxes that manifest consistent judgments across consecutive frames. This concerted effort is poised to minimize the propagation of misguided foreground object cues, thereby bolstering the performance of our change detection method.

Apply our approach to boost MU-Net: Our method is easy to deploy into existing U-Net based change detection networks. An application example of our method is provided in this section. We apply our method to the advanced MU-Net as shown in Fig. 3. Considering the advantage of one-stage object detection methods (e.g., YOLO [12]) over two-stage ones (e.g., Faster R-CNN [13]) in terms of inference speed, we adopt the state-of-the-art YOLOv8 [12] detector in the improved version of MU-Net to offer object information. The adopted object detector is pre-trained on the large-scale object detection dataset.

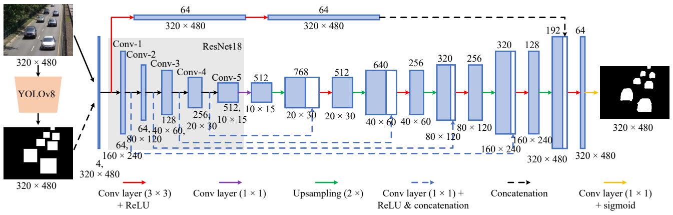


Fig. 3. The application of our method to MU-Net. Compared to the original MU-Net, this improved version introduces object information provided by the state-of-the-art object detector YOLOv8 [12].

In order to mitigate the negative impact caused by false detection, we apply the proposed bounding box selection strategy to the improved version of MU-Net. Specifically, after completing the extraction of object information for all video frames, for each image frame, we select five frames before it and five frames after it, and a total of ten frames are used as reference images. If the current frame is the first frame or the last frame, the number of reference images is five. Other cases in turn, finally for any image frame, the number of reference images will not be less than five. A bounding box in the current frame is kept if it has associated cues in no less than two reference images, otherwise it is discarded. For the existence of an associated cue in the reference frame, it mainly depends on intersection over union (IoU). Concretely, if the reference frame contains a bounding box with an IoU exceeding 0.1 when compared to the bounding box in question, it indicates the existence of an associated cue. In addition, it is necessary to determine whether the retained bounding box is the foreground or not, and retain the foreground object while discarding the background object. If a bounding box in the current frame has other bounding boxes with IoU greater than 0.95 in five or more reference frames, the bounding box belongs to the background, otherwise it belongs to the foreground.

After the judgment of adjudication of foreground object bounding boxes, we set the foreground region to pure white and the background region to pure black, thus generating a single-channel foreground mask. We concatenate it with the original RGB image to form a new four-channel input. Since the original MU-Net takes RGB images with three channels as input, the number of input channels of the first convolutional layer of the improved version network, namely Conv-1, is modified to 4, as shown in Fig. 4.

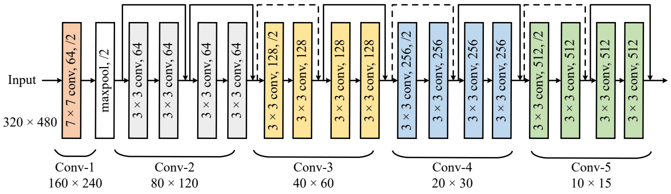


Fig. 4. ResNet-18 encoder backbone adopted in MU-Net. The solid shortcuts keep dimensions constant, while the dotted shortcuts increase dimensions. In the improved version of MU-Net, the input to Conv-1 has 4 channels instead of original 3.

Experiments: To compare the performance between the improved version and the original MU-Net[1], we conduct experiments on the CDnet2014 dataset. As the most comprehensive change detection dataset, CDnet2014 comprises 11 categories and 53 different video

[1] https://github.com/CIVA-Lab/Motion-U-Net

Table 1. Evaluation Results of the Original MU-Net and the Improved Version (ours) on CDnet 2014. ↑ Represents Higher is Better, While ↓ Means Lower is Better

Category	Recall ↑		FPR ↓		FNR ↓		PWC ↓		Precision ↑		F-measure ↑	
	MU-Net	Ours	MU-Net	Ours	MU-Net	Ours	MU-Net	Ours	MU-Net	Ours	MU-Net	Ours
PTZ	0.9551	0.9493	0.0002	0.0002	0.0449	0.0507	0.0342	0.0358	0.9562	0.9014	0.9549	0.9146
BadWeather	0.9800	0.9799	0.0007	0.0007	0.0200	0.0201	0.0865	0.0847	0.9606	0.9578	0.9701	0.9687
Baseline	0.9905	0.9901	0.0005	0.0005	0.0095	0.0099	0.0709	0.0643	0.9827	0.9854	0.9865	0.9877
CameraJit	0.9798	0.9811	0.0011	0.0011	0.0202	0.0189	0.1714	0.1655	0.9684	0.9694	0.9738	0.9750
DynamicBg	0.9823	0.9825	0.0002	0.0001	0.0177	0.0175	0.0281	0.0208	0.9727	0.9802	0.9774	0.9813
Intermitt	0.9802	0.9815	0.0004	0.0003	0.0198	0.0185	0.1749	0.1498	0.9937	0.9949	0.9868	0.9881
LowFrameR	0.7319	0.8293	0.0006	0.0005	0.2681	0.1707	0.1078	0.1116	0.9836	0.9297	0.7339	0.8154
NightVid	0.9609	0.9595	0.0008	0.0008	0.0391	0.0405	0.1381	0.1457	0.9552	0.9487	0.9579	0.9538
Shadow	0.9911	0.9915	0.0013	0.0013	0.0089	0.0085	0.1512	0.1478	0.9706	0.9715	0.9807	0.9814
Thermal	0.9829	0.9828	0.0009	0.0009	0.0171	0.0172	0.1365	0.1330	0.9787	0.9792	0.9808	0.9810
Turbulence	0.9735	0.9750	0.0002	0.0002	0.0265	0.0250	0.0315	0.0294	0.9712	0.9710	0.9721	0.9729
Overall	0.9553	**0.9639**	0.0006	**0.0006**	0.0447	**0.0361**	0.1028	**0.0990**	**0.9721**	0.9627	0.9523	**0.9563**

sequences, where spatial resolutions of the video vary from 320 × 240 to 720 × 576.

According to the experimental setup of MU-Net, we randomly selected 200 annotated frames from each video to form a total of 10 600 frame for training and the rest for testing. During training, 90% of 10 600 frame are used as training set and 10% as validation set. We adopt Adam optimizer with a learning rate of 0.0001 and the learning rate decreases by a factor of 10 every 20 epochs. Each model is trained for 40 epochs.

We choose six metrics commonly used in change detection for performance evaluation, including recall, false positive rate (FPR), false negative rate (FNR), percentage of wrong classification (PWC), precision and F-measure. The lower the FPR, FNR and PWC, the better model performance. The specific calculation formula of these indicators can be found at "changedetection.net".

We train two models, the original MU-Net and the improved version of MU-Net, on the CDnet2014 dataset. The quantitative evaluation results are reported in Table 1, where the original MU-Net is denoted as MU-Net, and the modified version of MU-Net is represented as ours. As listed in Table 1, compared to the original MU-Net, the improved version achieves better performance in the four metrics of overall recall rate, FNR, PWC and F-measure, that is, 0.9639, 0.0361, 0.999 and 0.9563. The performance difference between the improved version and the original MU-Net in FPR is very small. However, in the terms of precision metric, the overall performance of the improved version which introduces additional object information is worse than that of the original MU-Net which only uses images. This is because the improved version has lower precision for change detection in four categories of video sequences, including PTZ, badWeather, lowFramerate and nightVideos. The images constituting these videos exhibit marked dissimilarity from the images typically encountered by the pre-trained object detector. Consequently, it is difficult for the object detector to provide the correct object information of these images. While our meticulously devised bounding box selection and foreground object discrimination strategies abate the prevalence of erroneous detections, we still cannot completely eliminate the false detection results that cause a negative impact on the U-Net based change detector. According to the results in Table 1 and the above analysis, it is evident that our method is effectiveness and the use of accurate target information is beneficial for boosting the U-Net based change detector.

Conclusion: In this letter, we propose to exploit object information in images to boost the performance of U-Net based change detection models. We design a simple but effective method to realize the extraction and utilization of the object information. The proposed method is applied to the advanced MU-Net to achieve performance improvement. The experimental results verify the effectiveness of our method and that the object information is helpful for the U-Net based change detector.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China (62302161, 6230 3361) and the Postdoctoral Innovation Talent Support Program (BX20230114).

References

[1] L. Yan, W. Zheng, and F.-Y. Wang, "Heterogeneous image knowledge driven visual perception," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 1, pp. 255–257, 2024.

[2] W. Zheng, L. Yan, C. Gou, and F.-Y. Wang, "Computational knowledge vision: Paradigmatic knowledge based prescriptive learning and reasoning for perception and vision," *Artificial Intelligence Review*, vol. 55, no. 8, pp. 5917–5952, 2022.

[3] Z. Qin, X. Lu, X. Nie, D. Liu, Y. Yin, and W. Wang, "Coarse-to-fine video instance segmentation with factorized conditional appearance flows," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 5, pp. 1192–1208, 2023.

[4] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An expanded change detection benchmark dataset," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2014, pp. 387–394.

[5] M. Duan, K. Li, X. Liao, and K. Li, "A parallel multiclassification algorithm for big data using an extreme learning machine," *IEEE Trans. Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2337–2351, 2018.

[6] L. A. Lim and H. Y. Keles, "Learning multi-scale features for foreground segmentation," *Pattern Analysis and Applications*, vol. 23, pp. 1369–1380, 2020.

[7] W. Zheng, K. Wang, and F.-Y. Wang, "A novel background subtraction algorithm based on parallel vision and bayesian GANS," *Neurocomputing*, vol. 394, pp. 178–200, 2020.

[8] W. Zheng, K. Wang, and F. Wang, "Background subtraction algorithm based on bayesian generative adversarial networks," *ACTA Automatica Sinica*, vol. 44, no. 5, pp. 878–890, 2018.

[9] G. Rahmon, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Motion U-Net: Multi-cue encoder-decoder network for motion segmentation," in *Proc. Int. Conf. Pattern Recognition*, 2021, pp. 8125–8132.

[10] O. Tezcan, P. Ishwar, and J. Konrad, "BSUV-Net: A fully-convolutional neural network for background subtraction of unseen videos," in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision*, 2020, pp. 2774–2783.

[11] L. A. Lim and H. Y. Keles, "Foreground segmentation using convolutional neural networks for multiscale feature encoding," *Pattern Recognition Letters*, vol. 112, pp. 256–262, 2018.

[12] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE TPAMI*, vol. 39, no. 6, pp. 1137–1149, 2017.