Paper Title: AN ENCODER-DECODER NETWORK WITH MULTI-SCALE PULLING FOR LOCAL CHANGE DETECTION

Reviewer: 073F

We are thankful to the reviewer for providing thorough and constructive comments on our manuscript. We have revised the manuscript, as suggested by the reviewer and will submit the final manuscript.

Comment1: Encoder/Decoder approach seems to yield good results, and it is a general framework that might be worth exploring alternative architectures just so that we understand the technique.

Reply: In this work, we have conducted an ablation study to choose the most appropriate deep learning network for the proposed background subtraction algorithm. From Table 1, it may be observed that the use of existing deep neural networks such as VGG-16, GoogLeNet, and ResNet-50 for local change detection cannot accurately preserve the fine and coarse-scale features. Thus, these techniques produce a lesser value of the average F-measure for the challenging CDNet-2014 dataset. So, in the proposed scheme, a VGG-19 network is adhered to as an encoder that provides better accuracy than the existing deep neural networks.

Table1: Ablation study of average F-measure comparison on CDNet-2014 database with different Encoder

Category	VGG-16	GoogLeNet	ResNet-50	Proposed
BadWeather	0.9594	0.8557	0.9294	**0.9905**
Baseline	0.8949	0.7961	0.9461	**0.9981**
Camera Jitter	0.9422	0.8864	0.9518	**0.9970**
Dynamic Background	0.7356	0.6588	0.8220	**0.9962**
Intermittent Object Motion	0.7538	0.6488	0.8453	**0.9926**
Low Framerate	0.6175	0.5947	**0.8080**	0.7480

Night Videos	0.7526	0.6003	0.8585	**0.9852**
PTZ	0.7816	0.7136	0.7776	**0.9924**
Shadow	0.9084	0.8049	0.9647	**0.9971**
Thermal	0.8546	0.7725	0.9444	**0.9954**
Turbulence	**0.9207**	0.7637	0.8011	0.9148
Overall	0.8292	0.7360	0.8772	**0.9643**

Also, we have conducted an ablation study to check the effectiveness of the proposed algorithm with the decoder network consisting of stacked transposed convolutional layers and without transposed convolutional layers. From Table 2, it may be observed that the proposed model with a decoder network composed of stacked transposed convolutional layers attained better accuracy and effectively projected the feature level into pixel level than without transposed convolutional layers.

Table2: Ablation study of average F-measure comparison on CDNet-2014 database with different Decoder

Category	Proposed model with decoder network without transposed convolutional layers	Proposed model with decoder network consisting of transposed convolutional layers
BadWeather	0.9781	**0.9905**
Baseline	0.9927	**0.9981**
Camera Jitter	0.9785	**0.9970**
Dynamic Background	0.9796	**0.9962**
Intermittent Object Motion	0.9727	**0.9926**
Low Framerate	**0.7698**	0.7480
Night Videos	0.9340	**0.9852**
PTZ	0.9495	**0.9924**
Shadow	0.9866	**0.9971**

Thermal	0.9689	**0.9954**
Turbulence	**0.9597**	0.9148
Overall	0.9518	**0.9643**

Comment2: I'd zoom in on the images a bit more so that we can see the comparisons a bit better.

Reply: We are extremely sorry for our mistakes, and we have revised the manuscript, as suggested by the reviewer.