Computer Vision and Image Understanding

journal homepage: www.elsevier.com

ChangeNet-v2: Semantic Change detection with Convolutional Neural Networks

K. Ram Prabhakar[a], Akshaya Ramasamy[b], Suvaansh Bhambri[a], Jayavardhana Gubbi[b,**], R. Venkatesh Babu[a], Balamuralidhar Purushothaman[b]

[a]*Department of Computational and Data Sciences, Indian Institute of Science, INDIA - 560012*
[b]*Innovation labs, Tata Consultancy Services, INDIA*

ABSTRACT

Identifying areas of change between multiple images of the same scene captured at different times is a fundamental processing step in many image processing algorithms. Common challenges involved are illumination variation, camera jitter, shadows, etc. The task becomes even more challenging when the test and reference images are captured at a different time of the day, season, and viewpoint. In different applications, change can manifest in three types: insertion of an object, removal of an object and change in its state (including movement, shape, pose or form). The human mind interprets the change by comparing the current status with historical data at the intelligence level rather than using only visual information. In this paper, we present a deep architecture called ChangeNet-v2 for detecting changes between pairs of images and express the same semantically (label the change). The main objective is to detect changes at the semantic level rather than detecting all the changes in the background, which are irrelevant to the application. A parallel deep Convolutional Neural Network (CNN) architecture for localizing and identifying the changes between image pair has been proposed in this paper. We start with computing convolutional feature pyramid for both images. Later, the changes are segmented by finding a correlation between them. We benchmark our method on three different datasets: VL-CMU-CD, GSV, and TSUNAMI. Compared with several traditional and other deep learning-based change detection methods, our proposed method achieves higher overall accuracy and f-score in our experiments on all three datasets.

1. Introduction

Change detection in video analysis is often used as a stepping stone for high-level scene understanding. In its conventional form, the methods are used for identifying changes in the background by comparing any two consecutive frames or limited to short term temporal analysis (St-Charles et al. (2015)). In remote sensing literature, change detection is referred to as surface component alteration that is very useful in automatic land use analysis (Hussain et al. (2013)). The fact that the satellite images are registered helps in pixel-level change detection tasks that have been successfully extended to object-level change analysis. Some of the key challenges for visual change detection between any two images include variations in lighting or illumination, contrast, quality, resolution, noise, scale, pose and occlusion. The first five attributes are experienced in any change detection scenario but the last three attributes are either not experienced in short term temporal analysis or it can be easily handled by dropping frames.

Most of the methods in the literature that model background pixels to detect change address the first five attributes. In

**An earlier brief version of this work was published in Varghese et al. (2018).

**Corresponding author:
e-mail: j.gubbi@tcs.com (Jayavardhana Gubbi)

the case of remote sensing, where change detection is widely used, change in scale, pose and occlusion are rarely seen and the above methods can be easily deployed with suitable preprocessing. Although these approaches are a part of decision making, it involves low-level image analytics such as foreground-background segmentation. In more complex inferencing using visual input, particularly in pattern recognition and category formation, higher-level cognition is essential. For instance, when two images are being compared that have variations in pose, illumination, color information, and occlusion, the methods in literature often fail due to unregistered images, pose and scale variations as well as occlusions.

The traditional approach to this problem is to detect moving objects from the difference between the current frame and reference frame, often called *background image*, or *background model*. To do so, several foreground extraction methods consists of two stages: background modeling and foreground extraction. The first stage involves developing a model for the static or background pixels. In the second stage, the developed background model is used to detect pixels (*foreground* pixels) that deviate from the estimate. The success of such methods relies on the accuracy of the estimated background model. They need to be updated for new and challenging scenes.

Recently, Convolutional Neural Networks (CNN) are used to learn problem-dependent features that outperform traditional methods in foreground extraction. Lim and Keles (2018) proposed a method to intelligently fuse multiscale CNN features with feature pooling, to learn class-specific foreground extractor. More recently, Varghese et al. (2018) have shown that the learned CNN features are robust enough to handle challenging problems and their method is the current state-of-the-art in VL-CMU-CD dataset. However, they have shown to perform lower than VL-CMU-CD in GSV and TSUNAMI dataset. As a whole, there is no single method that outperforms in all the three datasets. Such a method would be able to detect both pixel level and semantic level changes irrespective of the challenges posed in real-life conditions.

Figure 1 shows an example from VL-CMU-CD change detection dataset (Alcantarilla et al. (2018)), where higher-level inferencing is required to detect the rubbish dumping on the pavement and the appearance changes are spread throughout the images. Motivated by the above problem, we develop a universal semantic change detection approach that can be used in various environmental conditions. In this paper, a novel deep learning architecture is proposed for change detection that targets higher-level inferencing. The new network architecture involves extracting features using CNN and combining filter outputs at different levels to localize the change. Finally, detected changes are identified using the same network, and output is an object-level change detection with the label. The proposed architecture is compared with the state-of-the-art using three different modern change detection datasets: VL-CMU-CD (Alcantarilla et al. (2018)), TSUNAMI (Sakurada and Okatani (2015)), and GSV (Sakurada and Okatani (2015)) datasets.

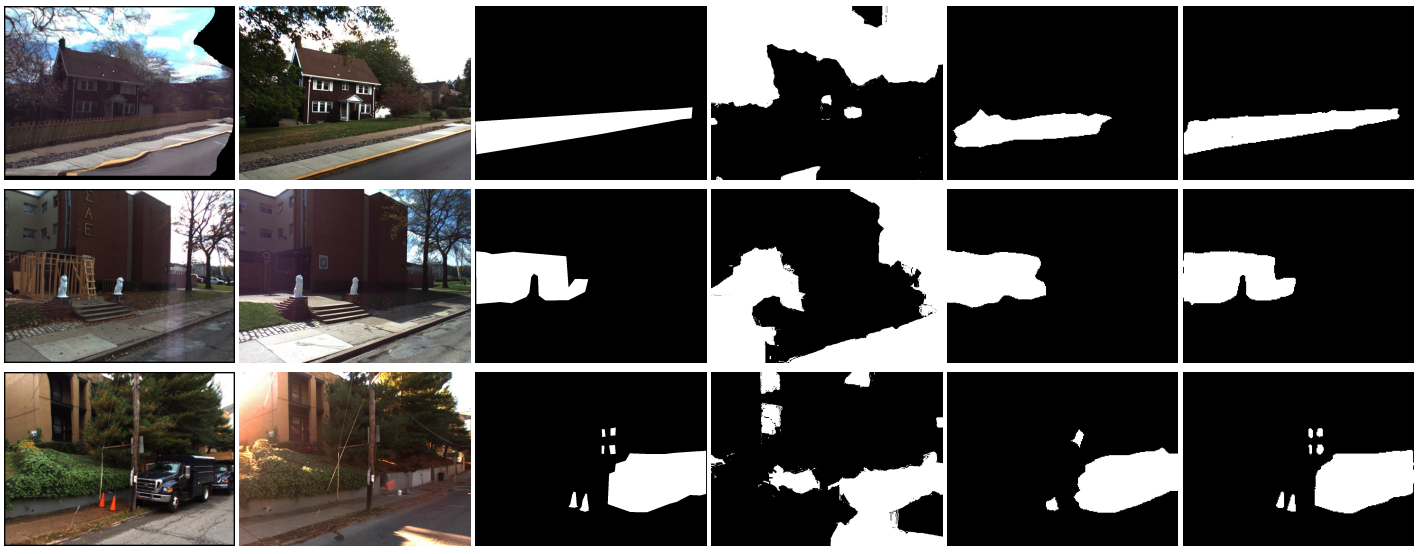
In summary, our contributions are:

- We propose a universal change detection method that is robust to camera motion and various environmental challenges.
- Through experimental evaluation, we show the efficacy of the proposed method in VL-CMU-CD, GSV and TSUNAMI datasets.

The rest of the paper is organized as follows. In Section 2, we discuss the related works. The proposed method is described in Section 3. The implementation details, evaluation protocols and results are discussed in Section 4. Finally, we conclude the paper in Section 5.

2. Related Works

Change detection has a very rich history of related works. It will be exhaustive to describe them all in this paper. Thus, we only focus on popular and recent best-performing change detection algorithms. To model the background model, statistical (or parametric) methods using Gaussians are proposed [Wren et al. (1997), Stauffer and Grimson (1999), Allili et al. (2007)] to model each pixel as a background or foreground pixel. However, parametric methods are computationally inefficient; to al-



(a) Test (b) Reference (c) Ground truth (d) SuperPixel (e) ChangeNet (f) ChangeNet-v2

Fig. 1: Our novel neural network architecture, *ChangeNet-v2*, can accurately segment the semantic changes between test and reference image. It can not only detect the changes with better boundary precision, but can also label them. Qualitative comparison for binary segmentation with SuperPixel (Gubbi et al. (2017)), ChangeNet (Varghese et al. (2018)) and proposed method for images from VL-CMU-CD dataset.

leviate this problem, various non-parametric methods [Barnich and Van Droogenbroeck (2011), Hofmann et al. (2012), St-Charles et al. (2015)] are proposed. "Vibe" proposed by Barnich and Van Droogenbroeck (2011) makes use of randomly selected pixels over time to compute background model. Additionally, Vibe updates neighbouring pixel background as well whenever a pixel is updated. This iterative process makes Vibe robust to motion artifacts. St-Charles et al. (2015) proposed an improved version of Vibe called "SubSCENE". They use color and local binary pattern information to increase spatial coherency.

One interesting development in semantic change detection was reported by Gressin et al. (2013) on satellite image processing. Although the work is on simulated data, for the first time, they have reported the perspective of change detection at different inference levels such as object, theme, and database akin to our work. As far as we are aware, this is the first work alluding to different levels in change detection. In a similar work, Kataoka et al. (2016) talk about semantic change detection by adding semantic meaning to the changing area. First, they find changed area using hyper maps, and then add semantic meaning to that changed area. Since last few years, after deep learning has become the main approach in computer vision, there have been some efforts in the creation of the dataset as well as in building change detection procedures.

Sakurada and Okatani (2015) was the first such attempt and they built two data sets with 100 image pairs known as TSUNAMI dataset and Google Street View (GSV) dataset. These are panoramic images created using street view separated temporally by several days or months. In addition to the creation of a dataset, they proposed a complex super pixel-based approach that uses the Convolutional Neural Network (CNN) for feature extraction. The low-resolution feature map generated from CNN network is combined with super-pixel segmentation to get precise segmentation boundaries of the changes. Although deep learning is used in the pipeline, there are many other hyperparameters in the procedure that needs finetuning for different scenarios.

Going one step further, Alcantarilla et al. (2018) propose a network called CDnet for finding structural changes in street view video. They create a new dataset of 152 categories with 11 unique object classes called VL-CMU-CD dataset. To create nearly registered images, they use visual simultaneous localization and mapping (SLAM) to get the 3D point cloud and then

project the points onto a 2D reference plane after determining the reference pose. It is a pixel-level change detection approach and uses contraction and expansion layers for pixel-level classication. The contraction block creates data representation. In this process, it stores max-pooling output for later use in the expansion network. The expansion block has been used for improving change localization.

The proposed ChangeNet-v2 architecture is different from ChangeNet approach. Our network determines the category of change in addition to change localization. We use parallel weight tied networks for feature extraction. It ensures both the network learn the same features from the two images. Therefore, the features from both the images can be compared easily. In addition to this, we compute the correlation at different levels of convolution layers so that the model captures the sparse and finer details of the object. Bilinear interpolation is used in ChangeNet for upsampling the data and the filter parameters are learned in the network itself. Another feature of ChangeNet-v2 is that it combines predictions from different levels of convolution layer. Such an approach helps the model to capture both coarse and fine details of the object. Apart from deep learning approaches, a multi-scale superpixel approach for drone image analysis has been proposed by Gubbi et al. (2017) with limited success on VL-CMU-CD dataset. The focus of this paper is to implement the change detection system in a computationally challenging environment. In the recent past, there has been quite a good amount of success in pixel-level image analysis using deep architecture. Bansal et al. (2016) proposed a new architecture for predicting surface normal that is useful in 2D-3D alignment. They use the pre-trained VGG-16 network for feature extraction followed by three layers of fully connected layers for predicting surface normal for every pixel. Bansal et al. (2017) generalized their earlier work and created PixelNet architecture and demonstrated semantic segmentation and edge detection in addition to surface normal estimation using an extended VGG-16 network. Such work has demonstrated that CNN can learn pixel-level information in addition to its success in image categorization. Similar work has been extended for regions of interest where they propose a new network for simultaneously predicting human eye fixations and segmenting salient objects. In addition to single image pixel analysis, there has been some recent work including the similarity between two images or signal pairs. Du et al. (2017) propose a Siamese CNN network for checking whether two handwritten texts are written by the same person or not. Both the inputs are encoded with the same network and then concatenated output is fed to a two-class classifier to determine whether handwriting is the same or not.

With the developments in change detection and pixel-level analysis using deep learning, we are motivated to solve the hard problem of change detection using a deep network. VL-CMU-CD dataset is our target as the scene pairs are complex and taken at different view angle, illumination and seasons as well. It has 11 different class of structural changes like construction-maintenance, bin on the pavement, new signboards, traffic cone on the road, vehicles, etc., including the background. To the best of our knowledge, it is a novel architecture for visual change detection particularly resulting in scene labels that can be viewed as semantic change detection. We further train the network to determine the category of change in addition to the changing area. Both the tasks happen within the network and involve single training. Most of the background information is irrelevant in our case since those changes could be due to season, illumination or viewpoint variation. It mainly looks for changes at the object level as compared to Alcantarilla et al. (2018). The model input is the test and reference images. The output is detection, localization, and categorization of the changed region. It mainly answers the following three questions in the presence of seven variations, which have been discussed earlier: is there any change? if yes, what is the change? and where is the change in the image?

3. ChangeNet-v2

In this section, we introduce our proposed CNN based foreground extraction method and its variation. The input to the model are two images: dynamic *test* frame and a static *reference* frame. Both the frames are passed onto a Siamese CNN

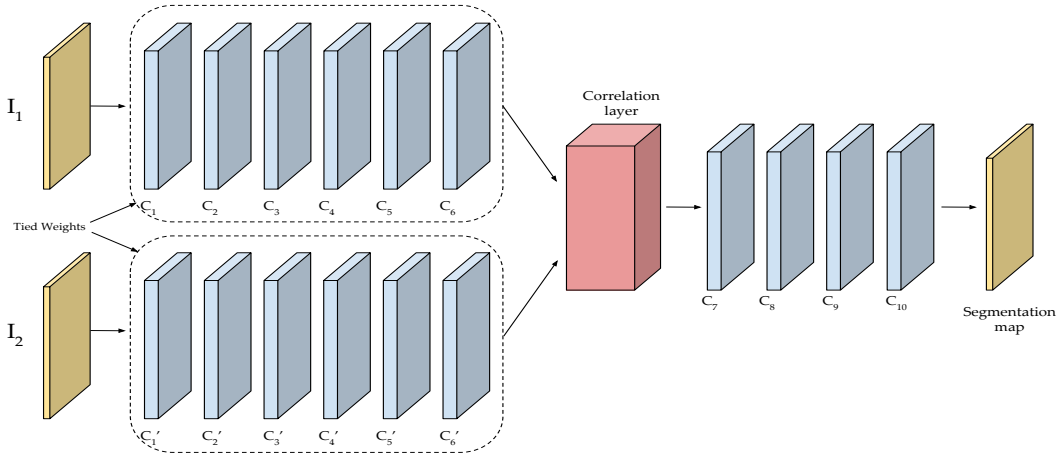


Fig. 2: Overview of proposed Correlation model. Test image (I_1) and reference image (I_2) are passed as input to Siamese architecture with six convolutional layers. Correlation between C_6 and C_6' estimates the likelihood of finding similar feature in a fixed neighbourhood. Convolutional layers C_7 to C_{10} convert the correlation map into binary segmentation output.

architecture to extract similar features from both of them. The pixel level similarity between features of test and reference images is computed using a correlation layer. The similarity features are passed onto further set of convolution layers to provide a probability score of belonging to background or foreground. In contrast to Scene-specific models (Babaee et al. (2018)), our method is a scene-agnostic one, meaning that we train a single CNN model for all test videos.

3.1. Correlation model

In Figure 2, we show our simple correlation model (named *Corr-model* further). Let I_1 be test frame with moving foreground object and I_2 be reference frame static background. The objective is to segment the moving foreground object from test frame. I_1 and I_2 are passed as input to a Siamese network, which consists of six convolutional layers with shared/tied weights. These features contain semantic information about both test and reference frame. One simplistic way to segment out the moving object is to subtract both features. However, that holds only for the case where both test and reference frames are registered. To solve this problem, we make use of correlation layer to compute pixel similarity.

A correlation layer computes patch comparison between features maps f_1 and f_2. The correlation between two patches p_1 (from f_1) and p_2 (from f_2) centered at a (x, y) is defined as follows:

$$C(p_1, p_2) = \sum_{o \in [-s,s] \times [-s,s]} \langle f_1(p_1 + o), f_2(p_2 + o) \rangle \tag{1}$$

where s is the size of the support window sampled around the pixel. Bigger the value of s, higher the robustness towards false matching. However, that also leads to more computation steps. To find the relative displacement of p_1, the correlation operation is applied to all pixels in a search area $T \times T$ of f_2 centered at (x, y). This results in an output of T^2 correlation values for every pixel in f_1. The computed correlation map is passed onto set of convolutional layers to obtain binary segmentation mask.

To find the right value for s, we did ablation with two settings: pixel correlation (termed as $corr - model(pixel)$ in Table 1) and patch correlation (termed as $corr - model(patch)$ in Table 1) with $s = 5$. From Table 1, we find that patch correlation does not offer enough improvement over pixel correlation for the extra computation.

Further, we also did ablation on the suitable feature maps to apply correlation as shown in Table 1. Along with correlation computed at sixth convolutional layer, we append correlation computed at first and second convolutional layer as well. However, appending such correlation maps offer very little improvement, suggesting the fact that sixth convolutional layer has enough semantic information for foreground extraction.

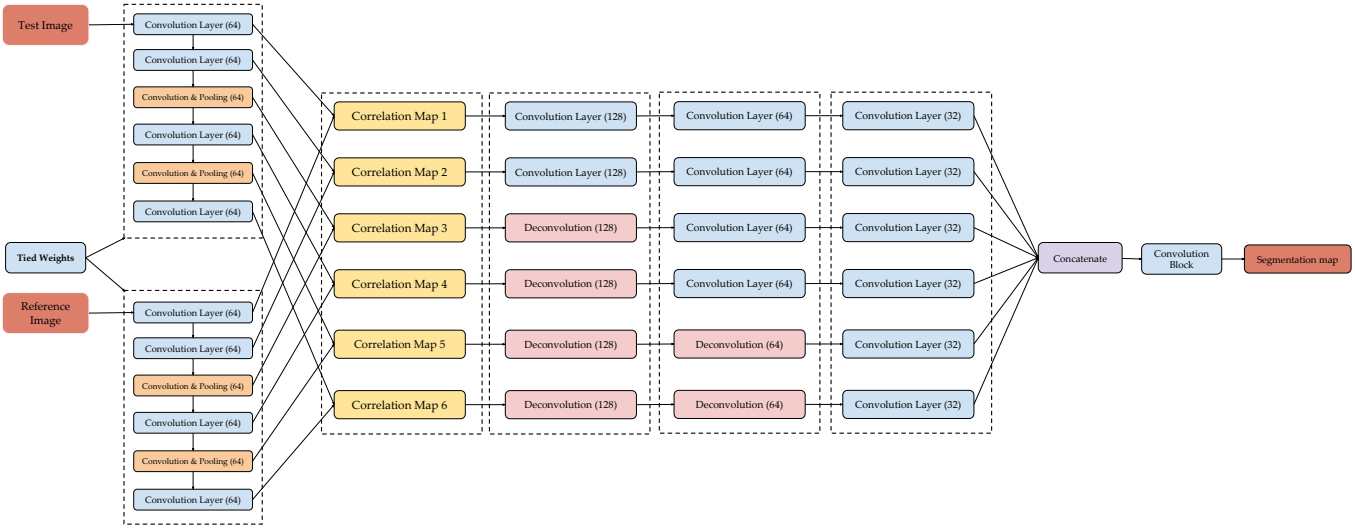


Fig. 3: Overview of proposed ChangeNet-v2 model. In contrast to model in Fig. 2, we compute correlation between all feature maps in Siamese architecture. Also, as the siamese architecture has max-pooling in later layers, their feature maps are deconvolved to obtain feature map with same size.

3.2. ChangeNet-v2 model

Motivated by success of pyramidal feature pooling in CNNs, we modify previous architecture to extract pyramidal convolutional features (Figure 3). The Siamese network from previous model is modified to include two max pooling layers to condense maximum spatial features. Also, we compute correlation for all feature maps in the Siamese architecture. However, as the features of later layers in Siamese network have different resolutions, we upsample them to same resolution with the help of transposed convolution layers.

4. Experiments

4.1. Network Implementation

The Siamese network consist of six convolutional Layers and two Max-pool layers (with stride=2 for corr-deconv model). Each convolutional layer consist of a convolution with kernel size = 3 with Batch Normalization and Leaky ReLU activation (with *alpha* = 0.01). For corr-deconv model, we use transpose-convolution layer (with kernel size=3, stride =2) to get the feature maps of size equivalent to the largest correlation map. The concatenated correlation maps are passed through four convolutional layers with kernel size = 5 and Leaky ReLU. Finally, the output segmentation map is obtained with Sigmoid activation function. The network is trained in a end-to-end fashion with binary cross entropy loss between predicted and ground truth segmentation map. We use Adam optimizer with learning rate of 0.0001 and $\beta_1 = 0.9$ and $\beta_1 = 0.999$ to optimize the loss function.

4.2. Datasets and protocols

We train and evaluate our method in three publicly available datasets: VL-CMU-CD (Alcantarilla et al. (2018)), GSV, and TSUNAMI (Sakurada and Okatani (2015)). VL-CMU-CD dataset consists of 1187 image pairs in total. We follow the dataset split used in ChangeNet (Varghese et al. (2018)) for our experiments as well. The dataset is split into ratio of 70:15:15% for training, validation and testing. Similartly, GSV and TSUNAMI datasets consists of 100 image pairs each, out of which, 70 image pairs were used for training and 15 image pairs each for validation and testing. We perform quantitative evaluation on all three datasets using five fold cross validation. For training the proposed model, we computed cross entropy loss between ground truth and predicted output with class re-balancing weights. We implemented our model in Tensorflow (Abadi et al. (2016)) installed on a workstation with Intel Xeon at 3.50 GHz CPU with 32GB RAM and a NVIDIA Titan X GPU card.

The three datasets and related methods in literature focus on binary classification, that is the final output is to detect change

Table 1: Comparison with State-of-art methods in VL-CMU-CD dataset.

Model	f-score
SuperPixel	0.1567
CDNet	0.5802
ChangeNet	0.8005
Corr-Model (Pixel)	0.9222
Corr-Model (Patch)	0.9270
Corr-Model (2^{nd} and 6^{th} layer)	0.9277
Corr-Model (1^{st} and 6^{th} layer)	0.9041
ChangeNet-v2 Model	**0.9385**

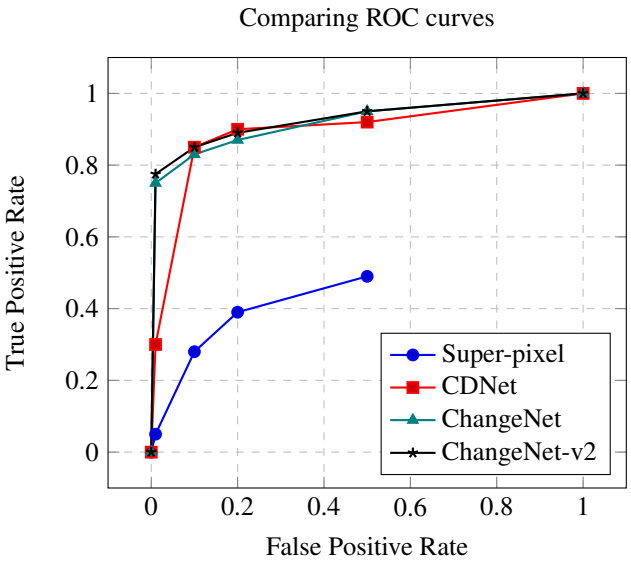


Fig. 4: ROC and TPR-FPR curve for binary class.

or no-change. We refer to this scenario as binary for the rest of the paper. We are also interested in labeling the object after the change is detected. We call this scenario multi-class. Currently, the system is built for multiclass classification of 10 commonly appearing objects in VL-CMU-CD dataset: barrier, bin, construction, person/bicycle, rubbish bin, sign board, traffic cone, and vehicle.

4.3. Results

4.3.1. Evaluation metrics

We evaluate the performance of the proposed method with state-of-the-art methods using standard F1 evaluation metric. Also, we evaluated proposed method on three datasets in Table 7 using following standard metrics: Accuracy, Precision, Recall, F1 score, mean Intersection over Union (mIoU), Matthew's correlation coefficient (MCC), Sensitivity, Percentage of Wrong Classifications (PWC), Specificity, False Positive Rate (FPR) and False Negative Rate (FNR).

4.3.2. Quantitative comparison

We compare our proposed ChangeNet-v2 method with three state-of-the-art methods in VL-CMU-CD dataset: SuperPixel (Gubbi et al. (2017)), CDNet (Alcantarilla et al. (2018)) and ChangeNet (Varghese et al. (2018)). The results are shown in Table 1. Compared with ChangeNet, proposed method offers over 14% improvement in f-score. The improvement in accuracy is attributed to the fact that proposed model is robust enough for image misalignment. Also, it is observed that corr-deconv model performs better than corr-model. Thus, for further evaluation in other datasets, we report only the performance of corr-deconv model.

The ChangeNet-v2 architecture was specifically designed keeping VL-CMU-CD dataset in mind due to its complexity. In order to validate the architecture, a 5 fold cross validation was conducted. The results are as shown in Table 3 and a healthy average f-score of 88.8% was obtained for binary classification and 73.4% for multi-class classification. To further confirm its performance, multi-scale SuperPixel Gubbi et al. (2017), CDnet Alcantarilla et al. (2018) and ChangeNet Varghese et al. (2018) were compared to the proposed architecture. In order to make a fair comparison, the results of binary classification (change or no change) of all the methods are compared by converting our class based output into binary form. The predicted change map of baseline approaches and our method are shown in Figure 1. Each sample exhibits different lighting and seasonal condition. The first column is the test image, which is compared against reference image in the second column. The ground truth change map is shown in the third column. The fourth and the fifth columns are the change detection results of SuperPixel and ChangeNet methods. The results of ChangeNet-v2 is shown

Table 2: Analysis of ChangeNet-v2 results at class level on VL-CMU-CD data set. Miscellaneous class has been excluded from the table as all the values were 0.

Classification	Class(→) Metric(↓)	Barrier	Bin	Construction	Other objects	Person/ Bicycle	Rubbish bin	Sign board	Traffic cone	Vehicle
Pixel-based	Precision	0.74	0.76	0.90	0.67	0.84	0.56	0.78	0.67	0.92
	Recall	0.70	0.72	0.85	0.65	0.79	0.50	0.69	0.60	0.88
	f-score	0.72	0.74	0.87	0.66	0.81	0.53	0.73	0.63	0.90
Object-based	Precision	1.00	0.97	0.88	1.00	1.00	0.96	1.00	1.00	1.00
	Recall	0.78	1.00	1.00	0.63	1.00	1.00	0.87	0.58	0.97
	f-score	0.87	0.98	0.94	0.78	1.00	0.97	0.93	0.73	0.98

Table 3: Average results of 5-fold cross validation for binary and multi-class categories in VL-CMU-CD dataset.

	Accuracy	Precision	Recall	f-score
Binary	99.2	93.7	93.9	93.8
Multi-class	78.5	76.0	71.3	73.4

Table 4: Performance metrics for ChangeNet-v2.

Metrics(→) Method(↓)	Pixel accuracy	Mean pixel accuracy	Mean IOU	Frequency weighted IOU
ChangeNet-v2	99.2	84.77	88.3	97.8

Table 5: The quantitative comparison of our method with other approaches for FPR = 0.1 and FPR = 0.01. The best scores are highlighed in **bold** and the second best in blue color.

	FPR = 0.1			FPR = 0.01		
Metrics (→) Methods (↓)	Precision	Recall	f-score	Precision	Recall	f-score
Super-pixel	0.17	0.35	0.23	0.23	0.12	0.15
CDnet	0.40	0.85	0.55	0.79	0.46	0.58
ChangeNet	0.79	0.80	0.79	0.80	0.79	0.79
ChangeNet-v2	**0.93**	**0.93**	**0.93**	**0.90**	**0.94**	**0.93**

Table 6: Comparison with state-of-the-art methods in VL-CMU-CD, GSV and TSUNAMI dataset with f-score.

Datasets (→) Methods (↓)	VL-CMU-CD	GSV	TSUNAMI
Super-pixel	0.15	0.26	0.38
CDNet	0.58	0.61	0.77
ChangeNet	0.80	0.45	0.73
ChangeNet-v2	**0.93**	**0.68**	**0.86**

in the last column. As shown in Figure 5, ChangeNet-v2 performs better than other approaches both in terms of the output as well as in terms of change class labeling. It gives a better performance in terms of accuracy and precision. Compared to other approaches, it gives additional information like what is the structural changes in the scene. In other words, our approach is able to tell where is the change in the scene as well as what the change is. ChangeNet-v2 performs well even though the background between image pair is different due to seasonal alterations and lighting conditions. For example, image pair in row 2 are taken at different lighting condition and ChangeNet-v2 was able to detect the changed area. An example of multiple changes in the same scene is depicted in row 4 where vehicle and a sign board are depicted as change. ChangeNet-v2 is able to identify both of them accurately. Results in row 5 shows performance of ChangeNet-v2 when images are captured at different seasonal condition. Reference image is taken in and its

Table 7: Performance metrics of ChangeNet-v2 for binary classification (change or no-change) on different datasets.

Metric (→) Dataset (↓)	Accuracy	Precision	Recall	F1	mIoU	MCC	Sensitivity (TPR)	PWC	Specificity (TNR)	FPR	FNR
TSUNAMI	0.921	0.845	0.881	0.863	0.827	0.808	0.881	0.078	0.938	0.062	0.119
GSV	0.840	0.629	0.759	0.688	0.665	0.587	0.759	0.159	0.865	0.135	0.241
VL-CMU-CD	0.992	0.937	0.939	0.938	0.883	0.934	0.939	0.765	0.995	0.004	0.060

having snow in the background. The same case applies to row 6 as well. Model performed well in this case, and it could detect and locate the rubbish bin. Since we approached the change detection problem at semantic level, we could mitigate irrelevant background information and reduce false alarms, if any.

Quantitative performance of our method is evaluated in two aspects. First aspect is how accurately it localized the change. Once it localized the change, what is the pixel labeling accuracy. Mainly, Intersection over Union (IoU) and pixel accuracy metrics are used for evaluating the performance. We considered 11 classes including background for this performance measurement. Model is evaluated with 177 image pairs and the results are generated. The performance metric for ChangeNet is given in Table 4. We achieved 98.4% pixel level accuracy and 83.93% mean pixel accuracy. In other words, 98.4% pixels are classified as change correctly. In that, 83.93% pixels are classied correctly per class basis. Also, we achieved 76.35% IoU. It compares the ground truth and predicted changed area on a per class basis. IoU is changed to 96.3% once we assigned the weights to class IoU based on their appearance frequency.

Table 2 shows the results of ChangeNet-v2 in identification of class-based change. Other than traffic cone and other objects, all other classes resulted in a f-score of over 0.8 for object level change detection. At pixel level, small objects including traffic cone, rubbish bin and signboard resulted in lower f-scores. Table 5 shows quantitative comparison of ChangeNet-v2 with ChangeNet, CDnet and Super-pixel methods for two different false positive rates of 0.1 and 0.01. As it can be seen, ChangeNet-v2 outperforms both the methods with impressive f-scores.

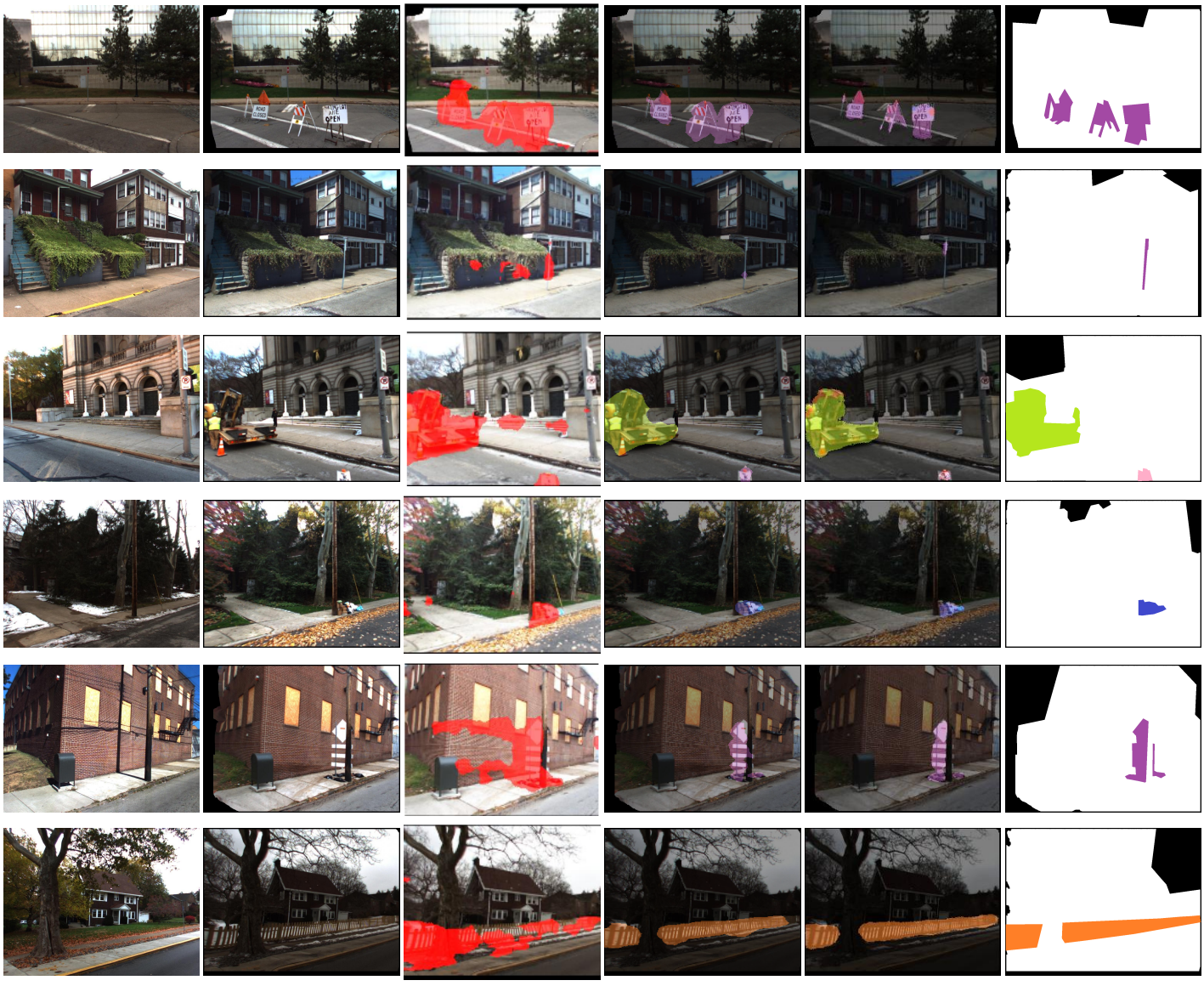
Figure 4 shows the Receiver Operator Characteristic (ROC) curve for binary classication. All the classes except background is considered as logical one. ChangeNet-v2 resulted in steep ROC curve with maximum true positive rate and minimum false positive rate. The area under ROC curve, i.e., AUC is 99.4%.

Finally, the performance of the three different methods on three different datasets is presented in Table 6. The proposed Changenet-v2 model performs better than compared methods in all three datasets in terms of f-score. It should be noted that the definition of change detection is evolving. The new datasets and techniques that can detect change at higher levels of inference are becoming possible. This paper is one of the early works in the direction, and hence there are very few methods in the literature that can be compared, which is presented in Table 6.

Detailed results of ChangeNet-v2 on the three datasets with eleven metrics is presented in Table 7. As it can be seen, the results on VL-CMU-CD dataset are very high with good performance on TSUNAMI dataset. There is a drop in GSV performance. The drop in performance on GSV dataset is attributed to the way the ground truth is created in these datasets. ChangeNet-v2 focuses on structural changes but GSV ground truth represents cars on the road as change. Hence, the overall performance seems to dip. The results of ChangeNet-v2 on TSUNAMI and GSV dataset are shown in Figures 6 and 7.

4.3.3. Qualitative comparison

We show the results generated by SuperPixel, ChangeNet and our method for images from GSV, TSUNAMI and VL-CMU-CD datasets in Fig. 1, 5, 6, and 7.



(a) Reference image (b) Test image (c) CDnet (d) ChangeNet (e) ChangeNet-v2 (f) Ground truth

Fig. 5: Qualitative comparison for multi-class segmentation with CDnet, ChangeNet and proposed method for images from VL-CMU-CD dataset.

5. Conclusion

In this paper, we introduced a simple class agnostic CNN model for change detection. Our CNN model has encoder-decoder architecture with correlation layer, which is trained in an end-to-end supervised fashion. We improve upon existing methods by computing pyramidal convolution features, which is matched using correlation layers; thus making our method robust against camera motion. Our method neither needs finetuning on test set like class specific models nor it requires any post-processing. The experimental evaluation reveals that we outperform existing methods in three different datasets with wide array of challenges.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: a system for large-scale machine learning., in: OSDI, pp. 265–283.

Alcantarilla, P.F., Stent, S., Ros, G., Arroyo, R., Gherardi, R., 2018. Street-view change detection with deconvolutional networks. Autonomous Robots 42, 1301–1322.

Allili, M.S., Bouguila, N., Ziou, D., 2007. A robust video foreground segmen-

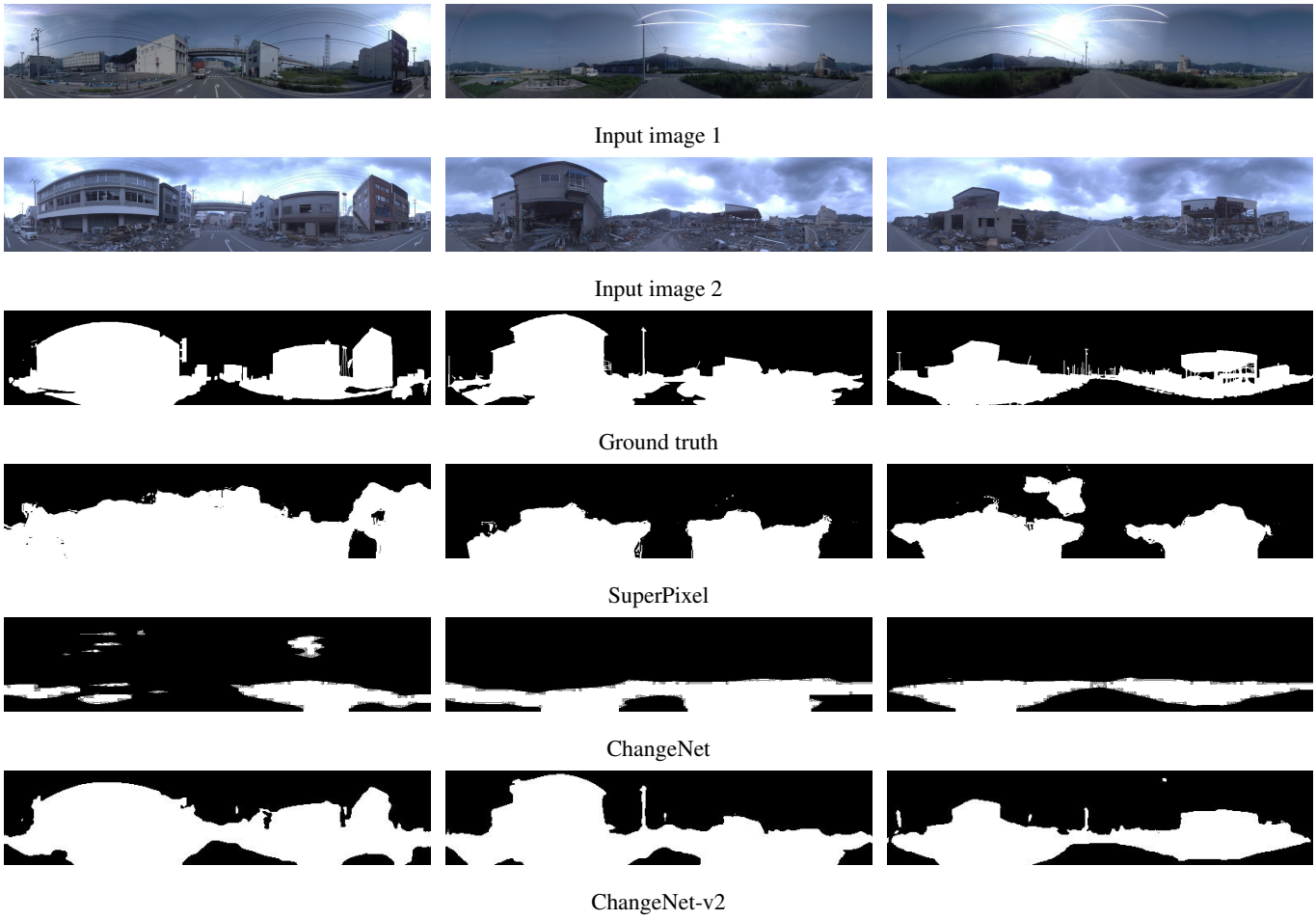


Fig. 6: Qualitative comparison with SuperPixel (Gubbi et al. (2017)), ChangeNet (Varghese et al. (2018)) and proposed method for images from TSUNAMI dataset.

tation by using generalized gaussian mixture modeling, in: Fourth Canadian Conference on Computer and Robot Vision, IEEE. pp. 503–509.

Babaee, M., Dinh, D.T., Rigoll, G., 2018. A deep convolutional neural network for video sequence background subtraction. Pattern Recognition 76, 635–649.

Bansal, A., Chen, X., Russell, B., Gupta, A., Ramanan, D., 2017. Pixelnet: Representation of the pixels, by the pixels, and for the pixels. arXiv preprint arXiv:1702.06506 .

Bansal, A., Russell, B., Gupta, A., 2016. Marr revisited: 2d-3d alignment via surface normal prediction, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5965–5974.

Barnich, O., Van Droogenbroeck, M., 2011. Vibe: A universal background subtraction algorithm for video sequences. IEEE Transactions on Image processing 20, 1709–1724.

Du, W., Fang, M., Shen, M., 2017. Siamese convolutional neural networks for authorship verification.

Gressin, A., Vincent, N., Mallet, C., Paparoditis, N., 2013. Semantic approach in image change detection, in: International Conference on Advanced Concepts for Intelligent Vision Systems, Springer. pp. 450–459.

Gubbi, J., Ramaswamy, A., Sandeep, N., Varghese, A., Balamuralidhar, P., 2017. Visual change detection using multiscale super pixel, in: Digital Image Computing: Techniques and Applications (DICTA), 2017 International Conference on, IEEE. pp. 1–6.

Hofmann, M., Tiefenbacher, P., Rigoll, G., 2012. Background segmentation with feedback: The pixel-based adaptive segmenter, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, IEEE. pp. 38–43.

Hussain, M., Chen, D., Cheng, A., Wei, H., Stanley, D., 2013. Change detection from remotely sensed images: From pixel-based to object-based approaches. ISPRS Journal of photogrammetry and remote sensing 80, 91–106.

Kataoka, H., Shirakabe, S., Miyashita, Y., Nakamura, A., Iwata, K., Satoh, Y., 2016. Semantic change detection with hypermaps. arXiv preprint arXiv:1604.07513 2.

Lim, L.A., Keles, H.Y., 2018. Learning multi-scale features for foreground segmentation. arXiv preprint arXiv:1808.01477 .

Sakurada, K., Okatani, T., 2015. Change detection from a street image pair using cnn features and superpixel segmentation., in: BMVC, pp. 61–1.

St-Charles, P.L., Bilodeau, G.A., Bergevin, R., 2015. Subsense: A universal change detection method with local adaptive sensitivity. IEEE Transactions on Image Processing 24, 359–373.

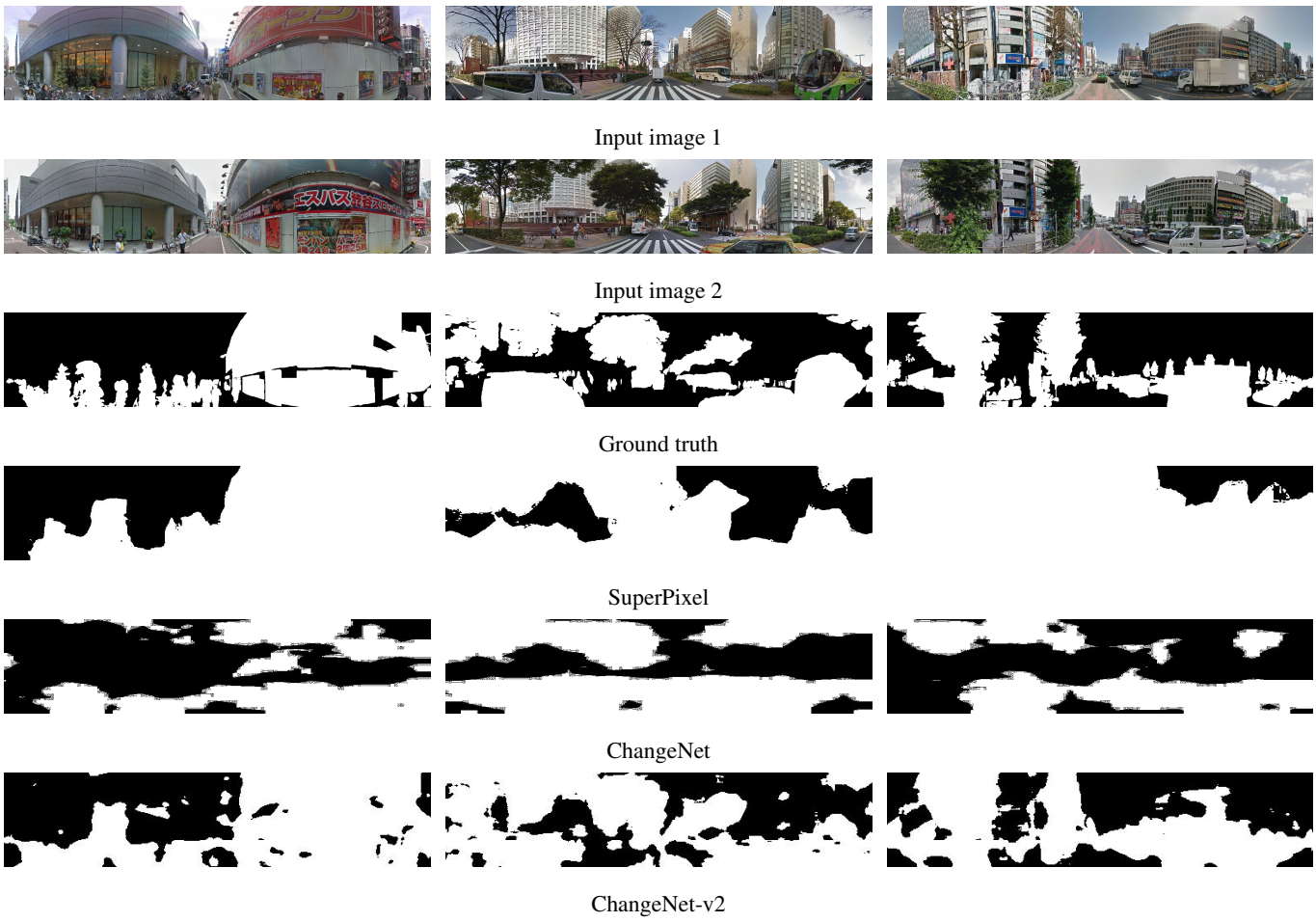


Fig. 7: Qualitative comparison with SuperPixel (Gubbi et al. (2017)), ChangeNet (Varghese et al. (2018)) and proposed method for images from GSV dataset.

Stauffer, C., Grimson, W.E.L., 1999. Adaptive background mixture models for real-time tracking, in: Computer Vision and Pattern Recognition, IEEE. p. 2246.

Varghese, A., Jayavardhana, G., Akshaya, R., Balamuralidhar, P., 2018. Changenet: A deep learning architecture for visual change detection, in: European Conference on Computer Vision Workshops (ECCVW), IEEE.

Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.P., 1997. Pfinder: Real-time tracking of the human body. IEEE Transactions on Pattern Analysis & Machine Intelligence , 780–785.