BgSub: a background subtraction model for effective moving object detection

Islam Osman[1 [0000−0001−9935−2417]] and Mohamed S. Shehata[1 [1111−2222−3333−4444]]

University of British Columbia, 3333 University Way, Kelowna, BC
islam.osman@ubc.ca

Abstract. Moving object detection is a core task in computer vision. However, existing deep learning-based moving object detection methods require a large number of labeled frames to achieve good generalization and performance. In moving object detection tasks, there is no such a large-scale labeled dataset because the labeling process requires a lot of time and effort. In this paper, we compiled a large-scale dataset by 1) combining existing moving object detection datasets. 2) using an inpainting deep learning model to transform datasets from video object segmentation task to moving object detection tasks. 3) generating synthetic datasets by combining random backgrounds with random foreground objects. Additionally, we propose a novel deep-learning model that performs background subtraction on the object level. This model is trained on the compiled dataset and shows superior performance in the moving object detection task. The model is evaluated using the CDNet dataset and results are compared with current state-of-the-art models. The results show that our model outperforms the best-reported state-of-the-art model by 1.6%.

Keywords: Moving object detection · Background subtraction · Deep learning.

1 Introduction

Moving object detection (MOD) is a critical task in various computer vision applications, including object tracking, autonomous driving, and surveillance [1], [19], [39]. As a result, it has garnered significant attention from researchers in recent years. The task of moving object detection is to output a binary mask where the moving objects are represented as white pixels and the background is represented as black pixels. There are many challenges in this task such as dynamic background (e.g., trees and sea movement), shadows, illumination changes, and camera motion.

Techniques for moving object detection are generally classified into three categories: 1) unsupervised learning methods, 2) semi-supervised learning methods, and 3) supervised learning methods. Unsupervised learning methods use holistic approaches to differentiate moving objects from the background without the

need for labeled data [8], [44]. These methods often require careful tuning of hyperparameters to perform effectively across different datasets, making them task-specific. However, they typically do not perform as well as supervised methods in terms of accuracy. Semi-supervised learning methods, the second category, aim to achieve high performance by using a combination of labeled and unlabeled data [17], [18]. While these methods can reduce the amount of labeled data needed, they usually require substantial training time to converge and may still face challenges in generalization. Finally, supervised learning methods [28], [9] tend to deliver the highest segmentation accuracy among the three categories but at the cost of requiring large amounts of labeled data. Since there is no large-scale labeled dataset in moving object detection tasks, these models have limited performance. We define a large-scale labeled dataset as a dataset that has hundreds of thousands of labeled images such as ImageNet [14] in image classification, and COCO [27] in object detection.

In this paper, we proposed a novel deep learning model for moving object detection called BgSub as an abbreviation for Background Subtraction. This model is a mixture of a ViT transformer and a convolutional neural network. To overcome the problem of the limited labeled dataset, we compiled a large-scale dataset of more than $300,000$ labeled images. This dataset has three parts. 1) Real dataset: a combination of multiple moving object detection datasets. 2) Transformed dataset: transforming datasets of video object segmentation to moving object detection. 3) Synthetic dataset: a dataset generated by selecting from a random background and combining it with random objects at random locations. Finally, we trained the proposed model on the compiled dataset and evaluated the model using a benchmark dataset for moving object detection called CDNet [46].

The paper is organized as follows: Section 2 provides a literature review. Section 3 provides a detailed explanation of the proposed work. Section 4 depicts the results of the experiments that were conducted. Finally, Section 5 concludes and summarizes the paper and discusses the future directions.

2 Related work

Moving object detection has evolved significantly, with methods generally categorized into statistical and CNN-based approaches. Statistical methods build a background model using intensity values and motion vectors (e.g., optical flow) to detect moving objects as pixels that deviate from the model. These include techniques like ego-motion compensation[3], feature-based methods[38], motion-based approaches[4], image-cue-based models[40], and subspace-based approaches[15], [16]. While effective in certain scenarios, these methods often struggle with high computational costs and are prone to performance degradation in environments with dynamic backgrounds or significant camera movement[29], [37].

CNN-based methods have shown improved accuracy in detecting moving objects, especially in complex environments with changing lighting or background

motion[46]. Early models like ConvNets[10] were trained on datasets like CDNet2014 to perform simple subtraction between current frames and static background images. Over time, more sophisticated architectures emerged. For instance, Babaee et al.[5] incorporated pre-processing steps to adapt to dynamic backgrounds and post-processing like spatial-median filtering[41] to accurately detect camouflaged objects. End-to-end models, such as Cascaded CNN[45], introduced multi-resolution networks to detect objects at varying scales while ensuring spatial coherence. Advanced models, such as Johnander et al.[21], Zhang et al.[48], and FgSegNet[25], [24], leveraged multi-scale feature extraction, optical flow, and temporal information to refine moving object detection. Furthermore, motion-guided attention models[23] integrated appearance and motion saliency. MODY-Net [35] uses an ensemble of multi-scale outputs to minimize the model uncertainty. Hence, reduces the false-positives. REFNet-TBPI [32] uses a multi-scale fusion network to capture objects' regions and boundaries. Hence, the model detects the object's fine details. Additionally, the model is trained using a continual learning technique called task-based parameter isolation (TBPI) to prevent forgetting across video sequences which improves the overall performance. TransBlast [31] is an attempt to use transformers in the MOD task with an augment loss function that maximizes the separation of foreground and background using a subspace learning module. This model has good generalization. However, its performance was limited because transformers require a massive amount of labeled data to achieve good performance. Hence, in this paper, we compiled a large-scale dataset to train a transformer-based model mixed with CNN layers to produce accurate masks achieving state-of-the-art performance in the MOD task.

3 Proposed work

In this paper, we propose a novel deep-learning model for segmenting moving objects from frame sequences given the background frame. Hence, the model learns to perform background subtraction on high-level features instead of the traditional pixel-to-pixel background subtraction. This model is a combination of vision transformer (ViT) and convolutional neural networks (CNN). The transformer is used to extract high-level features from the current frame and background frame. On the other hand, CNN is used to learn background subtraction of the high-level features extracted by the ViT and produce the output mask of the moving objects. This model is trained using a large-scale dataset that we compiled from various of different sources. In this section, we show the model architecture, the process of compiling the large-scale dataset, and the training procedure.

3.1 BgSub Architecture

BgSub consists of four major components as shown in Fig. 1, 1) Frame encoder, 2) Feature pyramid module, 3) Multi-scale fusion, and 4) Segmentation head.

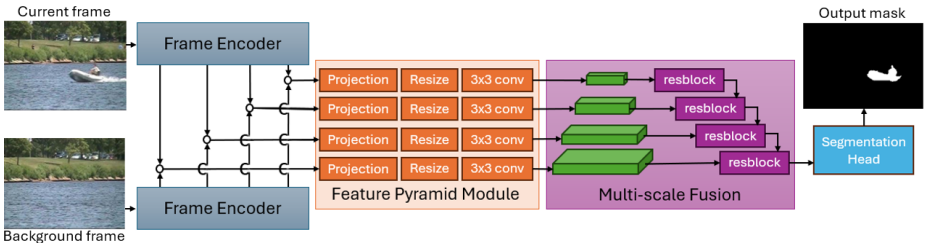

Fig. 1. The architecture of BgSub.

Frame encoder is used to extract feature maps from the current frame and the background frame. The encoder is based on a DinoV2 [30] transformer, which is a form of vision transformer trained using self-supervised learning on ImageNet. The model consists of 12 transformer blocks with embedding sizes that vary based on the model size (e.g., ViT-s embedding size is 384, while ViT-b embedding size is 768). The patch size is 16 (i.e., the input image is split into non-overlapping 16×16 patches). The input frame size is 224×224. We use the output of the last 4 blocks of the encoder to generate the feature pyramid. The output size of each block is $C \times H/16 \times W/16$, where C is the embedding size, H is the frame height, and W is the frame width.

Feature pyramid module (FPM) is used to extract multi-scale features from the frame encoder. Feature pyramid has 4 parallel paths. Each path consists of 4 sequential layers. The first layer is a concatenation layer to combines both features of the current frame and the background frame. The second layer is a projection layer (i.e., 1×1 convolutional layer) to produce different numbers of channels for different scales, such that small scales have a larger number of features than large scales to keep the inference time of the model reasonable (as the time needed to performing convolution operation on a large scale is longer than that on a small scale). The third layer is a resizing layer to get different scales. The last layer is a 3×3 convolutional layer to produce the feature map of each scale. The output is a feature pyramid of 4 scales, where each level of the pyramid has a scale of $2\times$ the next level. The lowest level has a feature map of size $C_1 \times H \times W$, while the top level has a feature map of size $C_4 \times H/8 \times W/8$

Multi-scale fusion is used to merge the feature pyramid levels into a single feature map. The fusion of pyramid levels is performed sequentially. The pyramid level i level is merged with level $i-1$ through a residual block (i.e., three convolutional layers with a skip connection from the first layer to the last layer). Hence, the output of the residual block at level 1 (i.e., lowest level) has information on all feature pyramid levels.

Segmentation head is used to produce the final output, which is a binary mask that represents the moving objects in white pixels and the background in black pixels. The segmentation head uses the output feature map of the multi-scale fusion process using three convolutional layers, two of which are 3×3

convolutional and the last one is a 1×1 convolutional layer followed by a sigmoid activation function.

3.2 MOD dataset

The compiled moving object detection dataset is a large-scale dataset of more than $300,000$ labeled images. This dataset consists of three types of datasets. 1) Real datasets: a combination of existing moving object detection datasets. This dataset includes AAU [6], LASIESTA [13], SBM [12], CDNet [46], and urbantracker [20]. These datasets have frame sequences with their corresponding segmentation masks. We select the background frame for each frame sequence as the frame that has the least amount of white pixels in its segmentation mask (i.e., the frame with almost no moving objects in it). Sample frames of this dataset are shown in Figure 2. 2) Transformed datasets: video object segmentation datasets focus on tracking and segmenting a specific object along a video sequence. The problem is that the object always exists in all frames. Hence, we can not select the background frame for any of the frame sequences. To overcome this issue, we used an in-painting model called ProPainter [49]. This model can remove an object from an image and replace it with a predicted background. For each frame sequence, we use the mask of the first frame and feed it to the ProPainter to generate the background frame. The datasets used are DAVIS16 [36] and YT-VOS18 [47]. An example of background generation is shown in Figure 3. 3) Synthetic dataset: we randomly select a background frame from all frame sequences, and then we add foreground objects from a random number of randomly selected frame sequences. An example of the generation is shown in Figure 4. The ratios of all datasets are shown in Figure 5.

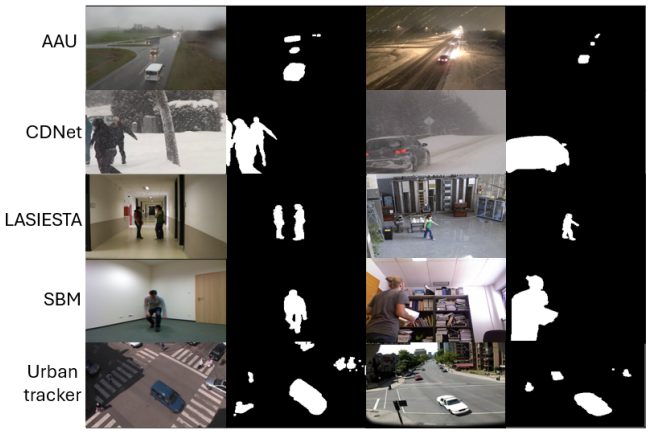

Fig. 2. Sample frames from existing MOD datasets.

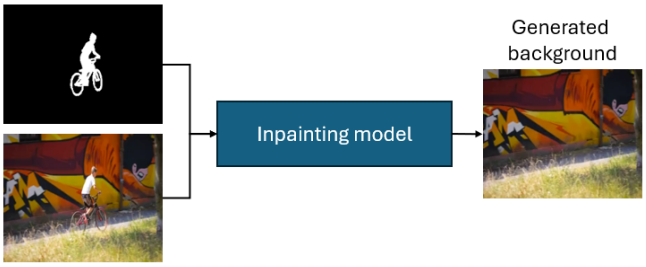

Fig. 3. Generating background frame to transform VOS dataset to MOD dataset.

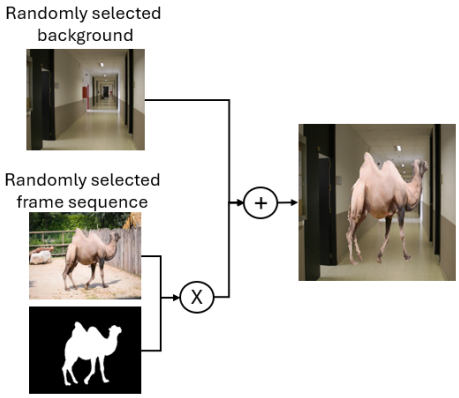

Fig. 4. An example of generating synthetic frame.

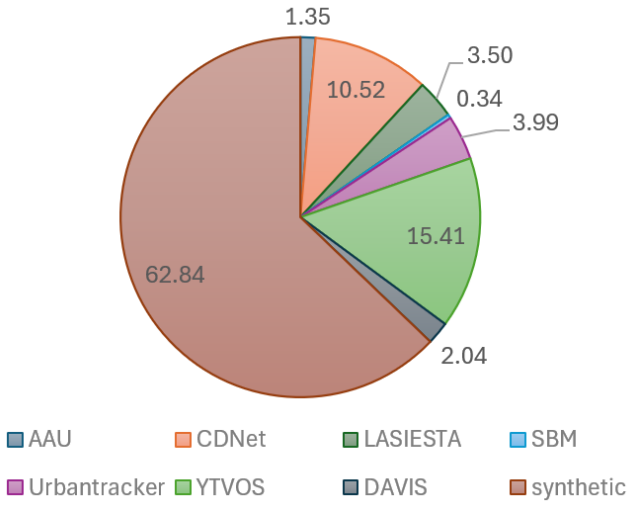

Fig. 5. The ratio of each dataset in the compiled dataset.

3.3 Implementation Details

Training is done for 120 epochs on the compiled dataset. The learning rate is $1e-5$, the batch size is 64, and the optimizer used is AdamW with weight decay of 0.01 and beta values 0.9 and 0.999. The data augmentation used are color jitter, random horizontal flips, and random affine (i.e., random rotation up to 15 degrees and random shear up to 10). The same augmentations are applied for both the current frame and background but with different values, the reason behind this is to make sure that the background frame does not look the same as the background in the current frame. Hence, preventing the model from learning to perform pixel-to-pixel subtraction. The same augmentations with the same values are applied to the ground truth mask except for the color jitter. The loss function used is a combination of focal loss [26] and IoU loss. The focal loss is calculated as follows:

$$\ell_{focal}(p) = \begin{cases} -\alpha(1-p)^{\gamma}\log(p) & \text{if } y = 1 \\ -(1-\alpha)p^{\gamma}\log(1-p) & \text{otherwise.} \end{cases} \tag{1}$$

where α is a weighting factor used to balance the importance of background and foreground pixels, p is the model estimated probability for the ground truth class, and γ is the focusing parameter which reduces the relative loss for well-classified examples, putting more focus on hard (i.e., misclassified examples). We calculated the ratio r of foreground to background pixels in all masks and set the value of α to $1-r$ to pay more attention to foreground pixels. The value of γ is typically chosen based on empirical testing resulting in $\gamma = 2$. On the other hand, the IoU (i.e., intersection over union) loss is calculated as follows:

$$\ell_{iou} = 1 - \frac{|A \bigcap B|}{|A \bigcup B|} \tag{2}$$

where A is the ground truth and B is the predicted regions. The reason why it is $1 - iou$ is that we want to maximize the iou while the loss function is usually minimized during training. The final loss function is:

$$\mathcal{L} = \ell_{focal} + \ell_{iou} \tag{3}$$

4 Experiments and Results

In this section, we evaluate the proposed BgSub using a moving object detection benchmark datasets called ChangeDetection.Net (CDNet) [46]. The results are compared against state-of-the-art moving object detection models. In this experiment, our model is trained using our compiled dataset which contains CDNet inside it. However, only 600 frames per frame sequence are in the compiled dataset. The rest of the frames are left for testing. Similarly, all other models are trained using the same 600 frames from each fame sequence in CDNet. The evaluation metrics used are recall (Re), precision (Pr), and F-score ($\mathcal{F}$).

Table 1. Results of BgSub against state-of-the-art moving object detection methods on CDNet.

Method	Re	Pr	$\mathcal{F}$
sEnDec [2]	0.903	0.798	0.842
FgSegNetV2 [24]	0.893	0.741	0.801
CascadeCNN [45]	0.821	0.771	0.786
IUTIS-5 [7]	0.789	0.808	0.771
SemanticBGS [11]	0.789	0.830	0.789
BSUV-Net [43]	0.820	0.811	0.786
BSUV-Net-SBGS [43]	0.817	0.831	0.798
DeepBS [5]	0.754	0.833	0.745
SuBSENSE [41]	0.812	0.751	0.741
WisenetMD [22]	0.817	0.766	0.753
PAWCS [42]	0.771	0.785	0.740
MODSiam [34]	0.823	0.692	0.714
TransBlast [31]	0.867	0.811	0.831
FeSh-Net [33]	0.889	0.813	0.851
MODY-Net [35]	0.903	0.822	0.865
REFNet [32]	0.961	0.827	0.881
REFNet-TBPI [32]	0.969	0.850	0.901
BgSub*	0.961	0.849	0.903
BgSub	**0.971**	**0.863**	**0.917**

Table.1 shows the results of the proposed BgSub against state-of-the-art moving object detection methods using the CDNet dataset. We reported the results of 2 versions of BgSub. The first one is BgSub trained using CDNet only, referred to as BgSub* in the table. The performance of BgSub is 0.2% better than the top-performing model REFNet-TBPI. However, the second version of the model is trained using our compiled dataset. This version of BgSub outperforms other models by 1.6% in the F-score, which is 1.4% better than training the same model on CDNet only. These results highlight the effectiveness of training the model on a large-scale MOD dataset. The detailed performance on different challenges of the CDNet is shown in Table.2. As shown in the table, the performance of BgSub is superior on all challenges except for night videos, PTZ, and turbulence. The problem with night videos is that it is hard to locate the objects even with the naked eye. However, the model is still able to detect 88.7% of the moving object pixels. For the PTZ challenge, the performance of BgSub is limited due to the fact that the background scene changes dramatically in this challenge from one frame to another. Since the model is based on background subtraction when the background of the current frame is dramatically different from the background frame the model produces false positives due to uncertainty. As the model is not sure whether the new objects in the current frame is a moving object or a static object. Finally, the moving objects in the turbulence challenge are very small in size and the model is sometimes unable to detect tiny objects.

Table 2. Detailed results of BgSub on the 11 different challenges in CDNet.

Method	# of videos	$\mathcal{F}$
Bad weather	4	0.950
Baseline	4	0.931
Camera jitter	4	0.941
Dynamic background	6	0.957
Intermittent object motion	6	0.926
Low frame rate	4	0.945
Night videos	6	0.887
PTZ	4	0.784
Shadow	6	0.945
Thermal	5	0.919
Turbulence	4	0.903

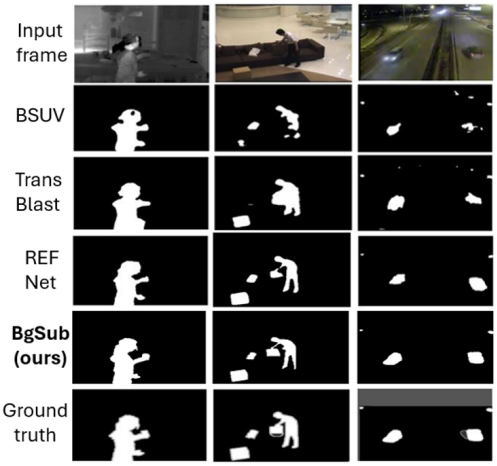

Fig. 6. Sample visual results of BgSub against other models.

4.1 Visual results

Figure 6 shows sample results from the proposed model BgSub against other MOD models on sample videos from the CDNet dataset. As shown in the figure the output of BgSub is almost the same as the ground truth, unlike other models that have some false positives and some false negatives. Two sample videos from each challenge in CDNet are shown in Figure 7. This figure highlights the effectiveness of BgSub in detecting moving objects under different challenges with superior performance. Even in the dynamic background challenge the model understands that the sea and trees are background objects and did not detect them as moving objects. The only downside of the proposed model is the PTZ challenge. This is due to the fact that the background massively changes from one frame to another. Hence, some new objects introduced to the scene may be detected as moving objects due to the model's uncertainty. Finally, to test

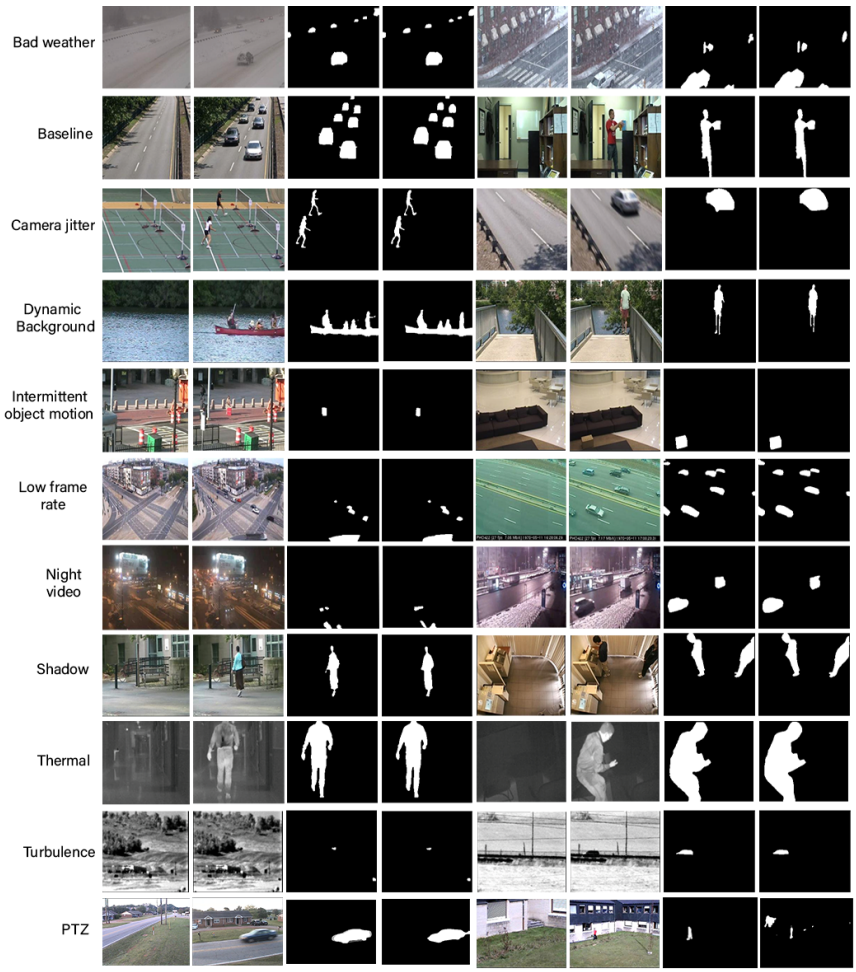

Fig. 7. Visual results of BgSub for each challenge in CDNet. The 1^{st} and 5^{th} columns are background frames, 2^{nd} and 6^{th} columns are current frames, 3^{rd} and 7^{th} columns are ground truth, 4^{th} and 8^{th} columns are BgSub output.

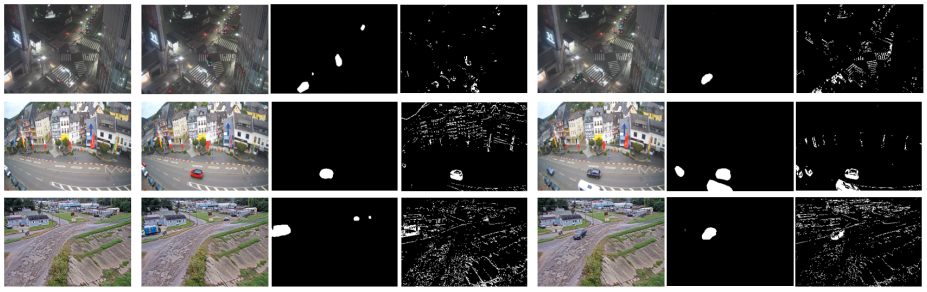

Fig. 8. Visual results of BgSub on real-world frame sequences captured from public IP cameras. The 1^{st} column is the background frame, 2^{nd} and 5^{th} columns are selected frames, 3^{rd} and 6^{th} columns are BgSub output, 4^{th} and 7^{th} columns are pixel-to-pixel subtraction between selected frame and background frame.

the proposed model generalization ability, we capture frames from live public IP cameras (unseen during training). We show the results of traditional pixel-to-pixel background subtraction to show that on the pixel-level there are differences between the background frame and every other frame. However, our proposed model learns to perform background subtraction on the pixel level. Hence, the model ignores the pixel difference and focuses on the object difference.

5 Conclusion

In this paper, we propose a moving object detection model based on background subtraction and compiled a large-scale dataset. The proposed work addresses the problem of the limited performance of other models due to the lack of a large number of labeled frames existence. Our model demonstrated superior performance on the CDNet dataset, outperforming the current state-of-the-art model by 1.6%. This performance improvement is mainly due to the compiled dataset. These results show the necessity of a large-scale labeled dataset in the field of moving object detection. As the visual results show that the model is able to detect objects from out-of-domain frame sequences, in future work, we will collect a real-world large-scale dataset from public IP cameras and use the proposed model to pseudo-label the dataset and make it publicly available.

References

1. Abdelpakey, M.H., Shehata, M.S.: Domainsiam: Domain-aware siamese network for visual object tracking. In: ISVC. pp. 45–58. Springer (2019)
2. Akilan, T., Wu, Q.J.: sendec: An improved image to image cnn for foreground localization. IEEE Transactions on Intelligent Transportation Systems **21**(10), 4435–4443 (2019)

3. Ali, S., Shah, M.: Cocoa: tracking in aerial imagery. In: Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications III. vol. 6209, p. 62090D. International Society for Optics and Photonics (2006)
4. Aslani, S., Mahdavi-Nasab, H.: Optical flow based moving object detection and tracking for traffic surveillance. International Journal of Electrical and Computer Engineering **8**(12), 840–845 (2013)
5. Babaee, M., Dinh, D., Rigoll, G.: A deep convolutional neural network for video sequence background subtraction. Pattern Recognition **76**, 635–649 (2018)
6. Bahnsen, C.H., Moeslund, T.B.: Rain removal in traffic surveillance: Does it matter? IEEE Transactions on Intelligent Transportation Systems **20**(8), 2802–2819 (2018)
7. Bianco, S., Ciocca, G., Schettini, R.: Combination of video change detection algorithms by genetic programming. IEEE Transactions on Evolutionary Computation **21**(6), 914–928 (2017)
8. Bouwmans, T.: Recent advanced statistical background modeling for foreground detection-a systematic survey. Recent Patents on Computer Science **4**(3), 147–176 (2011)
9. Bouwmans, T., Javed, S., Sultana, M., Jung, S.K.: Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. Neural Networks **117**, 8–66 (2019)
10. Braham, M., Droogenbroeck, M.V.: Deep background subtraction with scene-specific convolutional neural networks. In: 2016 International Conference on Systems, Signals and Image Processing (IWSSIP). pp. 1–4. IEEE (2016)
11. Braham, M., Piérard, S., Van Droogenbroeck, M.: Semantic background subtraction. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 4552–4556. IEEE (2017)
12. Camplani, M., Maddalena, L., Moyá Alcover, G., Petrosino, A., Salgado, L.: A benchmarking framework for background subtraction in rgbd videos. In: New Trends in Image Analysis and Processing–ICIAP 2017: ICIAP International Workshops, WBICV, SSPandBE, 3AS, RGBD, NIVAR, IWBAAS, and MADiMa 2017, Catania, Italy, September 11-15, 2017, Revised Selected Papers 19. pp. 219–229. Springer (2017)
13. Cuevas, C., Yáñez, E.M., García, N.: Labeled dataset for integral evaluation of moving object detection algorithms: Lasiesta. Computer Vision and Image Understanding **152**, 103–117 (2016)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
15. ElTantawy, A., Shehata, M.S.: Krmaro: Aerial detection of small-size ground moving objects using kinematic regularization and matrix rank optimization. IEEE Transactions on Circuits and Systems for Video Technology **29**(6), 1672–1686 (2018)
16. Eltantawy, I., Valera, M.: Accelerated krmaro: Subspace background subtraction under dynamic environments. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 100–109. IEEE (2019)
17. Giraldo, J.H., Javed, S., Bouwmans, T.: Graph moving object segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence (01), 1–1 (2020)
18. Giraldo, J.H., Javed, S., Sultana, M., Jung, S.K., Bouwmans, T.: The emerging field of graph signal processing for moving object segmentation. In: International Workshop on Frontiers of Computer Vision. pp. 31–45. Springer (2021)

19. Hu, H.N., Cai, Q.Z., Wang, D., Lin, J., Sun, M., Krahenbuhl, P., Darrell, T., Yu, F.: Joint monocular 3d vehicle detection and tracking. In: Proceedings of the IEEE international conference on computer vision. pp. 5390–5399 (2019)
20. Jodoin, J.P., Bilodeau, G.A., Saunier, N.: Urban tracker: Multiple object tracking in urban mixed traffic. In: IEEE Winter Conference on Applications of Computer Vision. pp. 885–892. IEEE (2014)
21. Johnander, J., Chatterjee, A., Felsberg, M.: Generative probabilistic modeling for background subtraction: Moving beyond gaussian mixture models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1559–1568 (2019)
22. Lee, S.h., Lee, G.c., Yoo, J., Kwon, S.: Wisenetmd: Motion detection using dynamic background region analysis. Symmetry **11**(5), 621 (2019)
23. Li, Y., Chang, H., Jiang, W., Zhang, X., Bao, L., Zhang, W.: Motion-guided attention for video salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10050–10059 (2019)
24. Lim, I.C.Y., Phung, S.L.: Fgsegnet v2: Fully automated foreground segmentation using multi-scale feature fusion. IEEE Transactions on Circuits and Systems for Video Technology **30**(12), 4682–4695 (2020)
25. Lim, L., Keles, H.: Foreground segmentation using a triplet convolutional neural network for multiscale feature encoding. arxiv 2018. arXiv preprint arXiv:1801.02225
26. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
28. Mandal, M., Vipparthi, S.K.: 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)
29. Nakaya, Y., Harashima, H.: Motion segmentation based on similarity of spatial-temporal region. In: Proceedings of 3rd International Conference on Image Processing. vol. 2, pp. 378–382. IEEE (1994)
30. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research Journal pp. 1–31 (2024)
31. Osman, I., Abdelpakey, M., Shehata, M.S.: Transblast: self-supervised learning using augmented subspace with transformer for background/foreground separation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 215–224 (2021)
32. Osman, I., Eltantawy, A., Shehata, M.S.: Task-based parameter isolation for foreground segmentation without catastrophic forgetting using multi-scale region and edges fusion network. Image and Vision Computing **113**, 104248 (2021)
33. Osman, I., Shehata, M.S.: Few-shot learning network for moving object detection using exemplar-based attention map. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 1056–1060. IEEE (2022)
34. Osman, I.I., Shehata, M.S.: Modsiam: Moving object detection using siamese networks. In: 2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE). pp. 1–6. IEEE (2020)

35. Osman, I.I., Shehata, M.S.: Mody-net: Moving object detection using multiscale output ensemble y-network. IEEE Canadian Journal of Electrical and Computer Engineering **44**(4), 491–496 (2021)
36. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
37. Pouzet, L., Garnier, L., Strugarek, J., Bascle, B.: Robust video object segmentation using region-wise classification. Signal Processing: Image Communication **29**(10), 1185–1195 (2014)
38. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: European conference on computer vision. pp. 430–443. Springer (2006)
39. Scheiner, N., Kraus, F., Wei, F., Phan, B., Mannan, F., Appenrodt, N., Ritter, W., Dickmann, J., Dietmayer, K., Sick, B., Heide, F.: Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using doppler radar. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
40. Shen, X., Yu, J., Zeng, Q.: Moving object detection in the presence of complex background. Pattern Recognition Letters **34**(9), 1035–1044 (2013)
41. St-Charles, P.L., Bilodeau, G.A., Bergevin, R.: Subsense: A universal change detection method with local adaptive sensitivity. IEEE Transactions on Image Processing **24**(1), 359–373 (2014)
42. St-Charles, P.L., Bilodeau, G.A., Bergevin, R.: A self-adjusting approach to change detection based on background word consensus. In: 2015 IEEE winter conference on applications of computer vision. pp. 990–997. IEEE (2015)
43. Tezcan, O., Ishwar, P., Konrad, J.: Bsuv-net: A fully-convolutional neural network for background subtraction of unseen videos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2774–2783 (2020)
44. Vaswani, N., Bouwmans, T., Javed, S., Narayanamurthy, P.: Robust subspace learning: Robust pca, robust subspace tracking, and robust subspace recovery. IEEE signal processing magazine **35**(4), 32–55 (2018)
45. Wang, N., Shelley, M., Heitz, G., Perona, P.: Interactive dynamic video segmentation using convolutional neural networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9. IEEE (2017)
46. Wang, Y., Jodoin, P.M., Porikli, F., Konrad, J., Benezeth, Y., Ishwar, P.: Cdnet 2014: An expanded change detection benchmark dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 387–394 (2014)
47. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
48. Zhang, X., Ma, J., Zhang, F.: Fast moving object detection with a network based on encoder-decoder structure. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 344–352 (2019)
49. Zhou, S., Li, C., Chan, K.C., Loy, C.C.: Propainter: Improving propagation and transformer for video inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10477–10486 (2023)