DSFNet: Dynamic and Static Fusion Network for Moving Object Detection in Satellite Videos

Chao Xiao, Qian Yin, Xinyi Ying, Ruojing Li, Shuanglin Wu, Miao Li, Li Liu, *Senior Member, IEEE*, Wei An, and Zhijie Chen

***Abstract*— Moving object detection (MOD) in satellite videos remains challenging due to the extremely small size of the interested targets and the highly complex background. Both the intra-frame (static) and inter-frame (dynamic) information are of great importance to MOD. In this letter, we propose a two-stream detection network named dynamic and static fusion network (DSFNet) to tackle the MOD problem in satellite videos. Specifically, the DSFNet is composed of a 2-D backbone to extract static context information from a single frame and a lightweight 3-D backbone to extract dynamic motion cues from consecutive frames. Then the extracted static and dynamic features are fused and fed into the detection head to detect the moving targets in satellite videos. We conduct extensive experiments on videos collected from Jilin-1 satellite and the results have demonstrated the effectiveness and robustness of the proposed DSFNet. Experimental results show that our DSFNet achieves the-state-of-the-art performance.**

***Index Terms*— Moving object detection (MOD), satellite videos.**

I. Introduction

SATELLITE surveillance has important applications in various scenes, such as urban planning, traffic monitoring, and military reconnaissance. Moving object detection (MOD) from satellite videos is one of the most essential tasks in automatic satellite surveillance and has received significant attention. Specifically, MOD aims at localizing and identifying objects with semantic similarity (spatial aspect) and continuous motion (temporal aspect) in a video and plays an important role in the task of object tracking. MOD from satellite videos remains unsolved due to the following challenges.

1) *Extremely small objects:* Due to the long imaging distance, the size of interested moving objects is often extremely small (e.g., most moving vehicles from Jilin-1 satellite videos are smaller than 20 pixels), leading to the lack of appearance information like texture and geometry cues.
2) *Low local contrast between objects and background:* Due to various complex scenes and imaging noise, the interested objects are sometimes submerged in cluster and thus of low local contrast to the background.
3) *Motion artifacts due to non-stationary satellite imaging platform:* The slow motion of the satellite causes local misalignment and dynamic intensity variations of some stationary background objects that are uninformative for the motion of the interested objects and can lead to motion artifacts in MOD.

Traditional methods mainly employ frame differencing [1]–[4] or background subtraction [5]–[9] to tackle the problem of MOD in satellite videos. However, these methods rely heavily on motion information and thus are easily influenced by non-stationary satellite platforms. To suppress the false alarms caused by non-stationary satellite platforms, Zhang *et al.* [10] estimated a confidence score map from optical flows across a video and employed it to promote real moving objects. However, the performance of traditional methods heavily relies on handcrafted features, and it is difficult for them to use handcrafted features and fixed hyper-parameters to handle variations of real scenes.

Recently, deep learning-based methods with its powerful feature learning capabilities have brought significant progress in object detection from common imagery [11], [12]. However, these appearance-based object detection methods may not be adequate for MOD from satellite videos due to aforementioned challenges. Despite the successful application of deep learning in common imagery, deep learning-based MOD from satellite imagery has been underexplored with very limited work [13]. LaLonde *et al.* [13] proposed a two-stage framework named ClusterNet using 2-D convolutions to learn spatio-temporal information from stacked consecutive frames to detect small objects in airborne images. However, ClusterNet performs inferior on MOD in satellite videos due to the sensitivity of motion artifacts and the lack of a semantic discriminability between real moving targets and motion artifacts introduced by non-stationary satellite platforms. For MOD in satellite videos, both the spatio-temporal and contextual information are needed to discriminate real moving targets from motion artifacts.

In this letter, motivated by the two-stream framework for action recognition [14], we propose a two-stream network named dynamic and static fusion network (DSFNet), which incorporates both the static context information and the

Manuscript received September 8, 2021; revised October 12, 2021; accepted October 25, 2021. Date of publication October 29, 2021; date of current version January 11, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62001478, Grant 61921001, Grant 61872379, and Grant 71701205. *(Corresponding authors: Miao Li; Li Liu.)*

Chao Xiao, Qian Yin, Xinyi Ying, Ruojing Li, Shuanglin Wu, Miao Li, and Wei An are with the College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: lm8866@nudt.edu.cn).

Li Liu is with the College of System Engineering, National University of Defense Technology, Changsha 410073, China (e-mail: dreamliu2010@gmail.com).

Zhijie Chen is with the National Airspace Technology Key Laboratory, Beijing 100085, China.

Digital Object Identifier 10.1109/LGRS.2021.3124222

dynamic motion cues to tackle the problem of MOD in satellite videos. In order to tailor for the studied problem, we use DLA-34 [15] as our static stream and re-customize the down-sampling scheme to adapt to the small size of moving objects in satellite videos. At the same time, we design an effective and efficient 3-D network as our dynamic stream for motion information learning, unlike the computationally expensive optical flow used in [14]. Then, features from the static and dynamic streams are fused and fed into the detection head to generate the final detection results. Note that, due to the complementary characteristics of static and dynamic streams, our DSFNet cannot only detect objects with low local contrast to the background but also suppress the false alarms caused by the non-stationary satellite platform.

The main contributions of this letter can be summarized as follows.

1) We propose a two-stream network named DSFNet for the first time to combine the static context information and the dynamic motion cues to detect small moving object in satellite videos.
2) For the static stream, we re-customize the down-sampling scheme and use the shallow layers of feature maps to maintain precise locations and details of small moving objects. For the dynamic stream, we design a lightweight 3-D convolutional network to efficiently extract dynamic motion cues from satellite videos. We further proposed a multi-scale hierarchical feature fusion scheme to fuse features from both streams, which can effectively improve the detection performance.
3) Extensive experiments on the dataset collected from Jilin-1 have demonstrated that the proposed DSFNet outperforms the state-of-the-art methods by a large margin.

II. Proposed Dynamic and Static Fusion Network

The overall architecture of the proposed DSFNet is shown in Fig. 1. As illustrated in Fig. 1, current frame and T consecutive frames are first fed to the static and dynamic streams to generate feature representations, respectively. Then we fuse the obtained features and send them to the detection head to produce the final detection results. Details of the 2-D static stream and 3-D dynamic stream are described in Sections II-A and II-B, respectively. The feature fusion and detection is presented in Section II-C.

A. Two-Dimensional Static Stream

For the static stream, each frame $I_t \in \mathbb{R}^{H\times W\times 3}$ in a video is fed to the DLA-34 [15] to generate the hierarchical features $f_{2\mathrm{d}\_i} \in \mathbb{R}^{H/2^{(i-1)}\times W/2^{(i-1)}\times C_i}$. Since features in shallow layers contain details and precise location information while features in deep layers deliver semantic information, we employ several feature fusion blocks (FFBs) to aggregate the hierarchical features. The details of the FFB module is shown in Fig. 1 and can be formulated as

$$f_{2\mathrm{d}\_i-1} = \mathrm{dc}(\mathrm{dc}(f_{2\mathrm{d}\_i-1}) + \mathrm{dc}(\mathrm{Up}(f_{2\mathrm{d}\_i}))) \tag{1}$$

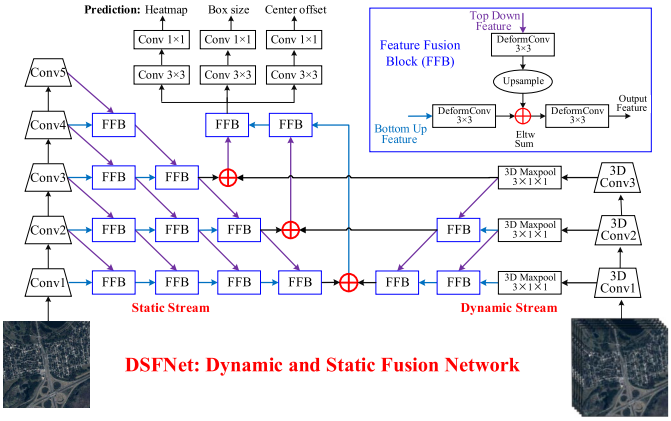

Fig. 1. Illustration of the overall framework of our proposed DSFNet. DSFnet consists of a 2-D static stream (left) and a 3-D dynamic stream (right). Features produced by these two streams are fused and fed to the detection head to generate the detection results.

where i is larger than 1. $\mathrm{dc}(\cdot)$ represents deformable convolution with a kernel size of 3×3 and $\mathrm{Up}(\cdot)$ represents the transposed convolutional layer. The deformable convolution in FFB is used to mitigate the misalignment between low-level and high-level features.

After fusion, we get three layers of enhanced features $F_{2\mathrm{d}\_i} \in \mathbb{R}^{H/2^{(i-1)}\times W/2^{(i-1)}\times C_i}$. Due to the hierarchical features fusion and the utilization of shallow layers of features, the output feature maps of the static stream can not only maintain the precise locations and details, but also enhance features of small moving objects in satellite videos.

B. Three-Dimensional Dynamic Stream

For the 3-D dynamic stream, T consecutive frames $I_\tau \in \mathbb{R}^{T\times H\times W\times 3}$ is fed to a self-developed 3-D network, which consists of three 3-D convolutional layers, to generate the hierarchical features $f_{3\mathrm{d}\_i} \in \mathbb{R}^{T\times H/2^{(i-1)}\times W/2^{(i-1)}\times C_i}$. The details of the 3-D network is shown in Fig. 1. To reduce the computational complexity, we replace each 3-D convolution by three 1-D convolution blocks, that is, a 1-D convolution, a batch normalization, and a ReLU).

To fuse the hierarchical features, we first utilize 3-D max pooling to reduce the temporal dimension and then employ several FFB modules for feature aggregation. After this, we can get three enhanced feature maps $F_{3\mathrm{d}\_i} \in \mathbb{R}^{H/2^{(i-1)}\times W/2^{(i-1)}\times C_i}$ as outputs. Note that our designed 3-D backbone consists of only three layers of 3-D convolution blocks and thus is lightweight and computationally efficient. Moreover, since 3-D convolution can get spatial and temporal information at the same time, we can extract the dynamic motion cues of the targets through the designed 3-D backbone.

C. Feature Fusion and Detection

The static stream is responsible for capturing appearance and contextual information about objects and scenes, and the dynamic stream aims to encapsulate the motion information of objects across the frames. The features from both streams

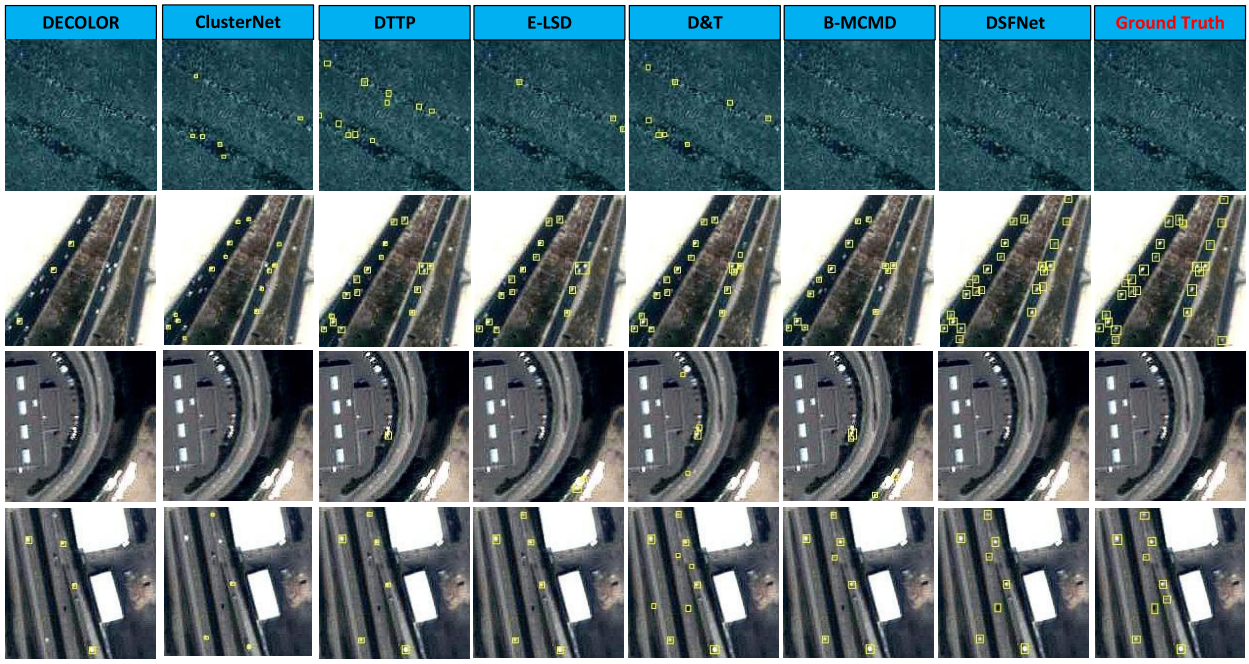

Fig. 2. Qualitative results of MOD in satellite videos. Row 1 and row 2 are from video 1. Row 3 and row 4 are from video 4. Row 1 and row 3 represent the zoom-in areas of background. Row 2 and row 4 represent the zoom-in areas of various targets.

are naturally complementary to each other, thus can be effectively fused for better representations. Therefore, we perform element-wise summation to fuse the features extracted from the two streams. The process can be formulated as

$$F_{\text{sum\_}i} = F_{\text{3d\_}i} \oplus F_{\text{2d\_}i} \tag{2}$$

where $\oplus$ represents element-wise summation.

Since the latter layers of a convolutional network capture stronger semantics and the early layers capture more detailed information of the small targets, we fuse the multi-level feature maps in a hierarchical manner. Specifically, feature maps $F_{\text{sum\_}i}$ and $F_{\text{sum\_}i+1}$ are fed into an FFB module to generate feature map $F_{\text{fused\_}i+1}$, which is then sent to another FFB module with $F_{\text{sum\_}i+2}$ to generate feature map $F_{\text{fused\_}i+2}$. The process repeat $n-1$ times to achieve progressive hierarchical feature fusion and n is set to 3 in our letter. Finally, the fused feature map $F_{\text{fused\_}n} \in \mathbb{R}^{H\times W\times C_1}$ is fed to the detection head to generate the detection results.

The detection head consists of three parallel branches to predict heatmap, object center offsets, and bounding box sizes, respectively. Each branch is implemented by a 3×3 2-D convolutional layer (with 128 channels), a ReLU, and a 1×1 2-D convolutional layer. The final detection results are obtained by decoding the outputs of these three branches.

III. Experiments

A. Experimental Setup

We collected 72 videos from Jilin-1 satellite as the training set and seven videos as the test set. The moving vehicles in the videos were selected as the targets. We used five consecutive frames as the input of our network. The batch size was set to 4. We performed random mirror and color jittering for data augmentation. We trained our network using the Adam optimizer [16] for 65 epochs with an initial learning rate of 1.25×10^{-4}. The learning rate was decreased by a factor of 10 after 45 epochs and 55 epochs. The loss function is the same as [17]. All the models were implemented on two Nvidia RTX 2080Ti GPUs.

B. Comparison to the State-of-the-Arts

We compare our method with several state-of-the-art methods, including traditional methods (VIBE [6], GoDec [7], DECOLOR [8], DTTP [9], E-LSD [5], D&T [4], and B-MCMD [10]) and a CNN-based method (ClusterNet [13]). We follow [4] to use precision, recall, and F1 score as the evaluation metrics. The quantitative and qualitative results are shown in Table I and Fig. 2, respectively.

1) Quantitative Results: Table I shows the quantitative detection results of different methods. It can be observed that our DSFNet produces the highest F1 score values on all of the seven videos and achieves the highest average scores on all three metrics. It can be observed that the improvements achieved by our DSFNet over the traditional methods is significant. That is because, our DSFNet can learn discriminative features of small moving objects and performs robust to various challenging scenarios (e.g., targets with low local contrast, local misalignments, and illumination variations). Compared with deep learning-based method ClusterNet [13], our DSFNet still has an obvious improvement on detection performance. That is because ClusterNet [13] only uses the spatio-temporal information to detect moving objects and is sensitive to non-stationary objects in the scene, resulting in many false alarms (shown in Fig. 2). On the contrary, our DSFNet not only exploits spatio-temporal dynamic information to promote

TABLE I

RECALL (RE), PRECISION (PR), AND F1 SCORE (F1) VALUES ACHIEVED BY DIFFERENT METHODS ON SEVEN SATELLITE VIDEOS. THE BEST RESULTS ARE SHOWN IN RED AND THE SECOND BEST RESULTS ARE SHOWN IN BLUE

Method	Video1			Video2			Video3			Video4			Video5			Video6			Video7			AVERAGE		
	Re	Pr	F1	Re	Pr	F1	Re	Pr	F1	Re	Pr	F1	Re	Pr	F1	Re	Pr	F1	Re	Pr	F1	Re	Pr	F1
VIBE [6]	0.61	0.34	0.44	0.82	0.61	0.70	0.68	0.59	0.63	0.65	0.52	0.58	0.72	0.65	0.69	0.60	0.42	0.49	0.45	0.44	0.44	0.65	0.51	0.57
GoDec [7]	0.92	0.51	0.65	0.73	0.81	0.77	0.93	0.53	0.68	0.72	0.38	0.50	0.72	0.74	0.73	0.81	0.42	0.55	0.93	0.25	0.39	0.82	0.52	0.61
DECOLOR [8]	0.24	0.92	0.38	0.77	0.88	0.82	0.89	0.83	0.86	0.44	0.93	0.60	0.74	0.84	0.79	0.71	0.80	0.75	0.30	0.69	0.42	0.58	0.84	0.66
ClusterNet [13]	0.75	0.67	0.71	0.66	0.81	0.72	0.90	0.72	0.80	0.50	0.70	0.58	0.76	0.82	0.79	0.77	0.71	0.74	0.85	0.66	0.75	0.74	0.73	0.73
DTTP [9]	0.74	0.67	0.70	0.67	0.84	0.74	0.71	0.84	0.77	0.64	0.86	0.73	0.62	0.77	0.69	0.55	0.73	0.62	0.25	0.49	0.33	0.60	0.74	0.65
E-LSD [5]	0.71	0.83	0.77	0.75	0.88	0.81	0.64	0.67	0.65	0.61	0.86	0.72	0.57	0.92	0.70	0.55	0.82	0.66	0.58	0.61	0.60	0.63	0.80	0.70
D&T [4]	0.71	0.91	0.80	0.69	0.86	0.76	0.84	0.84	0.84	0.75	0.85	0.80	0.63	0.82	0.71	0.64	0.76	0.70	0.83	0.43	0.56	0.73	0.78	0.74
B-MCMD [10]	0.77	0.93	0.85	0.76	0.86	0.81	0.86	0.82	0.84	0.71	0.77	0.74	0.58	0.84	0.68	0.70	0.74	0.72	0.81	0.47	0.60	0.74	0.78	0.75
DSFNet(Ours)	0.92	0.92	0.92	0.88	0.85	0.86	0.95	0.81	0.88	0.83	0.92	0.87	0.95	0.67	0.79	0.82	0.86	0.84	0.83	0.87	0.85	0.88	0.84	0.86

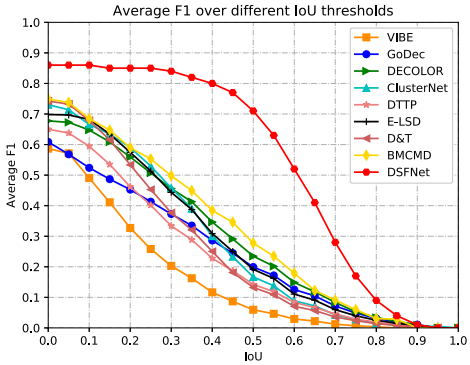

Fig. 3. Average F1 score over different intersection over union (IoU) thresholds.

the detection of moving objects, but also utilizes the static contextual information to suppress false alarms introduced by non-stationary satellite imaging platforms, thus improving the detection performance.

2) Qualitative Results: Fig. 2 shows the qualitative detection results of different methods. It can be observed that our proposed DSFNet not only reduces false alarms caused by dynamic changes (i.e., zoom-in regions of row 1 and row 3 in Fig. 2), but also improves the detection performance of moving vehicles with low local contrast to the background (i.e., zoom-in regions of row 2 and row 4 in Fig. 2). Nonetheless, the compared methods do not perform robust to various scenes, and the false alarms are increased in the region with misalignment or changing illumination.

We also evaluate the performance of different methods over varying intersection over union (IoU) thresholds. The results are shown in Fig. 3. The values of average F1 score all decrease as the IoU threshold increases. Compared with other methods, our DSFNet can maintain the best performance with the increase of IoU threshold. When the IoU threshold increases to 0.5, our DSFNet can still get an average F1 score about 0.7 while the average F1 scores of comparative methods drop over a half. That is because, all the compared methods are segmentation-based methods and they mainly focus on the locations instead of the bounding boxes of targets that can encapsulate the targets completely. Therefore, when the IoU threshold increases, the performance of the comparative methods will drop sharply. In contrast, our DSFNet not only focuses on the precise location of the targets, but also predicts the bounding boxes containing the complete targets, and thus is more robust to the varying IoU thresholds.

TABLE II

TIME COSTS (S) FOR A SINGLE FRAME OF DIFFERENT METHODS. THE BEST RESULTS ARE SHOWN IN RED AND THE SECOND BEST RESULTS ARE SHOWN IN BLUE

Methods	Time Cost	Avg Rec	Avg Pre	Avg F1
VIBE [6]	1.3	0.65	0.51	0.57
GoDec [7]	5.1	0.82	0.52	0.61
DECOLOR [8]	8.2	0.58	0.84	0.66
DTTP [9]	2.0	0.60	0.74	0.65
E-LSD [5]	34.2	0.63	0.80	0.70
D&T [4]	0.18	0.73	0.78	0.74
B-MCMD [10]	45.7	0.74	0.78	0.75
ClusterNet [13]	0.40	0.74	0.73	0.73
DSFNet	0.29	0.88	0.84	0.86

TABLE III

QUANTITATIVE ABLATION RESULTS OF DIFFERENT DESIGN CHOICES

Models	#Params(M)	FLOPs(G)	AP_{50}
CenterNet [17]	19.02	30.5	48.9
DSFNet with Static	6.62	41.61	54.3
DSFNet with Dynamic-3D-full	0.21	35.83	56.2
DSFNet with Dynamic	0.18	30.23	60.5
DSFNet-S	6.68	50.72	62.5
DSFNet-M	6.69	52.34	67.1
DSFNet-D	6.71	53.90	**70.5**

3) Time Costs: To compare the efficiency of different methods, we record the average time cost (s) of different methods on an input image with size of 1024×1024. The results are listed in Table II. It can be observed that our DSFNet outperforms the comparative methods by a large margin and is the second fastest method (slightly slower than D&T). This is because DSFNet needs to process five frames at a time to produce the detection results of a single frame, while the fastest method D&T [4] only processes three frames at a time by frame differencing. Note that compared with the background subtraction based methods, such as DECOLOR [8], ELSD [5], and B-MCMD [10], DSFNet achieves a superior balance between performance and efficiency.

C. Ablation Study

In this section, we present ablation experiments to validate the design choices of our DSFNet. All variants in the experiment are evaluated by Average Precision at IoU threshold 0.5 (AP_{50}). We use CenterNet [17] as a baseline for comparison. The results are shown in Table III.

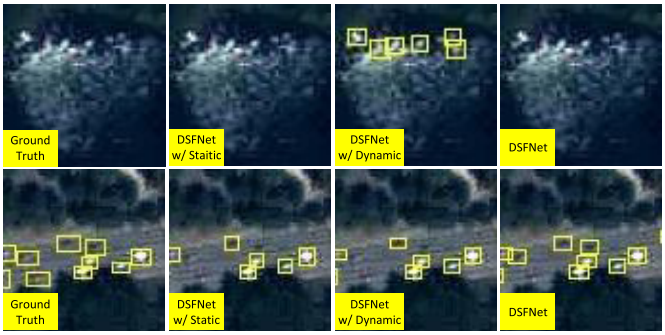

Fig. 4. Qualitative ablation results of different design choices. Row 1: zoom-in region with dynamic intensity variations. Row 2: zoom-in region with various moving targets. Our DSFNet can not only detect objects with low local contrast (shown in Row 2), but also suppress false alarms caused by illumination changes (shown in Row 1).

1) Effectiveness of Static Stream: To investigate the effectiveness of the static stream, we denote the model only with static stream as DSFNet w/ Static. It can be observed from Table III that DSFNet w/ Static outperforms CenterNet [17] with 5.4 AP_{50}. That is because our DSFNet w/ Static re-customizes the down-sampling scheme and uses shallow layers of features, thus maintaining the useful features and precise locations of small moving objects.

2) Effectiveness of Dynamic Stream: To investigate the effectiveness of the dynamic stream, we denote the model only with dynamic stream as DSFNet w/ Dynamic. It can be observed from Table III that DSFNet w/ Dynamic outperforms DSFNet w/ Static with 6.2 AP_{50}, which demonstrates the importance of the spatio-temporal dynamic information in MOD. Furthermore, to investigate the effectiveness of 3-D convolution decomposition, we denote the model only with dynamic stream consisting of 3-D convolution as DSFNet w/ Dynamic-3D-full. It can be observed from Table III that compared with DSFNet w/ Dynamic-3D-full, the AP_{50} of DSFNet w/ Dynamic is increased by 4.3 while the FLOPs is decreased by 5.6 G. That is because, the decomposition of 3-D convolution introduces extra nonlinear rectifications to enhance the representation ability of the model and reduces the computational cost simultaneously.

3) Effectiveness of Hierarchical Multi-Scale Feature Fusion: To investigate the effectiveness of the hierarchical multi-scale feature fusion, we fused the features extracted by static stream and dynamic stream at the shallow layer (denoted as DSFNet-S), shallow and middle layers (denoted as DSFNet-M), and all three layers (denoted as DSFNet-D), respectively. It can be observed from Table III that hierarchical multi-scale feature fusion can greatly boost the detection performance (AP_{50} from 62.5 of DSFNet-S to 70.5 of DSFNet-D). That is because hierarchical multi-scale feature fusion can not only enhance the feature maps with contextual information, but also maintain the representation of small targets.

The qualitative results of different model variants are shown in Fig. 4. It can be observed that by combining the static context information and the dynamic motion cues, our DSFNet can not only detect the objects with low local contrast, but also suppress the false alarms caused by non-stationary satellite platform.

IV. Conclusion

In this letter, we have proposed a two-stream detection network DSFNet for MOD in satellite videos. Our DSFNet fuses the static context information and dynamic motion cues to detect the moving targets. Extensive experimental results show that our DSFNet can not only detect moving objects with low local contrast to the background, but also suppress the false alarms caused by local misalignment and dynamic intensity variations. In addition, the experimental results show that our DSFNet surpasses previous state-of-the-arts by a large margin.

References

[1] M. Keck, L. Galup, and C. Stauffer, "Real-time tracking of low-resolution vehicles for wide-area persistent surveillance," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2013, pp. 441–448.

[2] I. Saleemi and M. Shah, "Multiframe many–many point correspondence for vehicle tracking in high density wide area aerial videos," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 198–219, Sep. 2013.

[3] L. W. Sommer, M. Teutsch, T. Schuchert, and J. Beyerer, "A survey on moving object detection for wide area motion imagery," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[4] W. Ao, Y. Fu, X. Hou, and F. Xu, "Needles in a Haystack: Tracking city-scale moving vehicles from continuously moving satellite," *IEEE Trans. Image Process.*, vol. 29, no. 7, pp. 1944–1957, Oct. 2019.

[5] J. Zhang, X. Jia, and J. Hu, "Error bounded foreground and background modeling for moving object detection in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2659–2669, Apr. 2020.

[6] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011.

[7] T. Zhou and D. Tao, "GoDec: Randomized low-rank & sparse matrix decomposition in noisy case," in *Proc. 28th Int. Conf. Mach. Learn., (ICML)*, 2011, pp. 33–40.

[8] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 597–610, Mar. 2013.

[9] S. A. Ahmadi, A. Ghorbanian, and A. Mohammadzadeh, "Moving vehicle detection, tracking and traffic parameter estimation from a satellite video: A perspective on a smarter city," *Int. J. Remote Sens.*, vol. 40, no. 22, pp. 8379–8394, Nov. 2019.

[10] J. Zhang, X. Jia, J. Hu, and K. Tan, "Moving vehicle detection for remote sensing video surveillance with nonstationary satellite platform," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 15, 2021, doi: 0.1109/TPAMI.2021.3066696.

[11] L. Liu *et al.*, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Jan. 2020.

[12] G. Chen *et al.*, "A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Jul. 17, 2020, doi: 10.1109/TSMC.2020.3005231.

[13] R. Lalonde, D. Zhang, and M. Shah, "ClusterNet: Detecting small objects in large scenes by exploiting spatio-temporal information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4003–4012.

[14] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[15] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, pp. 1–15, Dec. 2015.

[17] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.