DOMESTIC ANIMAL HEALTH MONITORING USING WIRELESS TECHNOLOGY

***Abstract*—In this paper, we consider the animal object detection and segmentation from wildlife monitoring videos captured by motion-triggered cameras, called camera-traps. For these types of videos, existing approaches often suffer from low detection rates due to low contrast between the foreground animals and the cluttered background, as well as high false positive rates due to the dynamic background. To address this issue, we first develop a new approach to generate animal object region proposals using multilevel graph cut in the spatiotemporal domain. We then develop a cross-frame temporal patch verification method to determine if these region proposals are true animals or background patches. We construct an efficient feature description for animal detection using joint deep learning and histogram of oriented gradient features encoded with Fisher vectors. Our extensive experimental results and performance comparisons over a diverse set of challenging camera-trap data demonstrate that the proposed spatiotemporal object proposal and patch verification framework outperforms the state-of-the-art methods, including the recent Faster-RCNN method, on animal object detection accuracy by up to 4.5%.**

***Index Terms*—Background modeling, camera-trap images, graph cut, object proposal, object verification.**

I. Introduction

WILDLIFE monitoring with camera-trap networks, especially with the collaborative efforts of citizen scientists, enable us to collect wildlife activity data at large space and time scales and to study the impact of climate change, habitat modification and human disturbance on species richness and biodiversity along the dimensions of scale, region, season, and species [1]. Camera-traps are stationary camera-sensor systems attached to trees in the field. Triggered by animal motion, they record short image sequences of the animal appearance and activities associated with other sensor data, such as light level, moisture, temperature, and GPS sensor data. They are an important visual sensor for wildlife that can record animal appearance without disturbance. Due to their relatively low cost, rapid deployment, and easy maintenance, camera traps are now being extensively used in wildlife monitoring, with the potential to be deployed at large scales in space and time. From camera-trap images, we can extract a rich set of information about animal appearance, biometric features, species, behaviors, their resource selection, as well as important environmental features about the surrounding habitats [2]. During the past several years, a vast amount of camera-trap data has been collected, far exceeding the capability of manual image processing and annotation by human. There is an urgent need to develop animal detection, segmentation, tracking, and biometric feature extraction tools for automated processing of these massive camera-trap datasets. In this work, we focus on accurate and reliable animal object detection and segmentation from camera-trap images.

Detecting and segmenting moving objects from the background is an important and enabling step in intelligent video analysis [3], [4]. There is a significant body of research conducted during the past two decades on background modeling and foreground object detection [5]–[7]. However, the availability of methods that are robust and generic enough to handle the complexities of natural dynamic scenes is still very limited [8]. Videos captured in natural environments represent a large class of challenging scenes that have not been sufficiently addressed in the literature [4]. These types of scenes are often highly cluttered and dynamic with swaying trees, rippling water, moving shadows, sun spots, rain, etc. It is getting more complicated when natural animal camouflage added extra complexity to the analysis of these scenes. Fig. 1 shows some examples of image sequences captured by camera-traps at days (with color images) and nights (with infrared images). Here, each column represents a camera-trap image sequence triggered by animal motion. The key challenge here is how to establish effective models to capture the complex background motion and texture dynamics while maintaining sufficient discriminative power to detect and segment the foreground animals. Traditional motion-based techniques are not suitable here since the background is highly dynamic.

Recently, approaches based on deep neural networks, such as RCNN [9] and its variations Fast-RCNN [10] and Faster-RCNN [11], are achieving the state-of-the-art performance in object detection. Typically, these methods have two major components: 1) object region proposal which scans the whole image to generate a set of candidate image regions (or bounding boxes

Manuscript received February 19, 2016; revised May 30, 2016 and July 13, 2016; accepted July 14, 2016. Date of publication July 27, 2016; date of current version September 15, 2016. This work was supported in part by the National Science Foundation under Grant CyberSEES-1539389 and Grant CPS-1544794. The work of W. Cao was supported in part by the National Science Foundation of China under Grant 61375015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alessandro Piva.

Z. Zhang and Z. He are with the Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO 65211 USA (e-mail: zzbhf@mail.missouri.edu; hezhi@missouri.edu).

G. Cao is with the School of Computer Science and Software Engineering, East China Normal University, Shanghai 200062, China (e-mail: gtcao@sei.ecnu.edu.cn).

W. Cao is with the College of Information Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: caom@shenzhenu.edu.cn).

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

Digital Object Identifier 10.1109/TMM.2016.2594138

Fig. 1. Samples of camera-trap images. Each column represents a camera-trap image sequence triggered by animal motions.

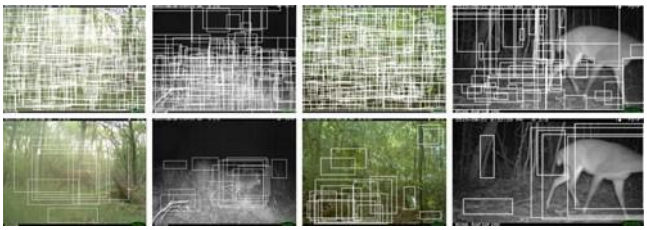

Fig. 2. Examples of bounding box proposals from spatial only and spatial-temporal methods. Each bounding box denotes a candidate object region. Top: selective search algorithm. Bottom: proposed IEC. The columns represent different scenes.

different locations and scales that could possibly contain the target objects, and 2) image classification which determines if these proposed regions are truly the objects or not. We observe that, within the context of animal detection from camera-trap images, these methods suffer from two major issues: speed and accuracy. First, the natural scenes in camera-trap images are highly cluttered. Existing object region proposal methods [12], [13] often generate a large number (thousands) of candidate object regions. We know that the deep convolutional neural network (DCNN) for region classification is computationally intensive and, more importantly, it needs to be performed thousands of times for each of these proposed object regions. Therefore, it is critical to consider the unique characteristics of camera-trap images in the spatiotemporal domain and design a new and efficient object region proposal method which can generate a small number of animal object proposals. To this end, we develop an iterative embedded graph cut (IEC) method with different foreground/background cut-off energy levels to create an embedded group of objects regions for the camera-trap image sequences. The examples in Fig. 2 show that IEC could significantly reduce the number of proposals while maintaining a sufficiently high object coverage rate. Second, we find that the direct application of DCNN to object regions in a single image is not efficient for animal-background classification. The performance can be significantly improved by extending the classification into the temporal domain using our proposed cross-frame patch verification method. Furthermore, for efficient animal object region classification, we find that a combination of DCNN and hand-crafted features achieves better classification performance. Our extensive experimental results demonstrate that the proposed method significantly improves the performance while maintaining low computational complexity.

The *major contributions* of this paper can be summarized as follows: 1) We have developed a new and efficient animal object region proposal method using IEC which jointly consider the animal motion and spatial context in the spatiotemporal domain. 2) We propose a cross-frame image verification method for accurate animal-background classification. 3) We have found that, for camera-trap images, the DCNN image features and hand-crafted histogram of oriented gradient (HOG) image features encoded with Fisher Vectors (FV) are able to enhance the classification performance for each other. 4) We have established a large dataset of camera-trap images which has been made available for the research community for developing efficient algorithms of object detection from highly cluttered natural scenes.

The remainder of the paper is organized as follows. We provide an overview of the proposed system in Section III. In Section IV, we present our animal object proposal method using iterative embedded graph cut. Section V explains the proposed cross-frame verification method. Experimental results are presented in Section VI. Section VII concludes the paper.

II. Related Work

This work is closely related to foreground-background segmentation, image verification, object region proposal, object detection and image classification. In the following, we provide a review of related work on these topics.

A. Foreground-Background Segmentation

Early work on background subtraction often operated on the assumption of stationary background. Several methods model the background explicitly, assuming a bootstrapping phase where the algorithm is presented with frames containing only the background [14], [15]. The use of multiple hypotheses to describe the behavior of an evolving scene at the pixel level significantly improves the performance of background modeling and subtraction [15]. Elgammal *et al.* [16] used a non-parametric background model to achieve better accuracy under the same constraints as the mixture of Gaussians. Sheikh and Shah incorporate the temporal and spatial consistencies into a single model [3]. Oliver *et al.* [17] focused on global statistics rather than local constraints to create a small number of eigen-backgrounds to capture the dominant variability of background. Considering spatial context and neighborhood constraints, graph cut optimization has achieved fairly good performance in image segmentation [7]. Iterated graph cut is used in [6] to search over a nonlocal parameter space. Background cut is proposed

in [18] which combines background subtraction and color or contrast-based models.

To handle background motion, various dynamic background texture models have been developed [15], [19]. Principal component analysis and autoregressive models are used in [17]. Wiener filters are used to predict the expected pixel value based on the past K samples. To reduce the computational complexity, Kahl *et al.* [20] demonstrated that using eigen-background on patches in an image is sufficient to capture the variance in dynamic scenes. In [21], for each pixel, it builds a codebook. Samples at each pixel are clustered into the set of codewords based on a color distortion metric. Gregorio and Giordano [22] use a weightless neural network to model the change in background. St-Charles and Bilodeau *et al.* [23] introduce a new strategy to tackle the problem of non-stationary background with pixel-level feedback loops to balance the local segmentation sensitivity automatically.

We recognize that, for accurate and robust video object detection and segmentation in dynamic scenes, background modeling of the dynamic pixel process at the image patch level, spatial context analysis and graph cut optimization at the region-level, and embedded foreground-background classification at the sequence level should be jointly considered. In this work, we propose to establish a new framework which tightly integrates these three important components for accurate and robust video object cut in highly dynamic scenes.

B. Region Proposals and Object Detection Using DCNN Methods

Recent studies [24], [25] have shown the extraordinary performance of DCNNs on image classification, object detection and recognition. To speed up the DCNN-based object detection process and avoid scanning of the whole image, object proposal methods have been recently developed for predicting object bounding boxes [11], [26]–[29]. Szegedy *et al.* [26] used a deep neural network as a regression model to predict the object bounding box. Sermanet *et al.* [28] developed a fully connected layer that is trained to predict the box coordinates for the localization task that assumes a single object. The fully connected layer is then turned into a convolutional layer for detecting multiple class-specific objects, which won the ILSVRC2013 localization competition. The original work on MultiBox [27] also used deep neural networks. Instead of producing bounding boxes, the MultiBox approach generates region proposals from a network whose last layer simultaneously predicts multiple class-agnostic boxes.

C. Image Verification

This work is also related to image verification. Image verification, in our particular problem, is regarded as a two-class classification problem: to verify if a proposed object image patch is an animal or belongs to the background scene. Classic learning-based image verification often involves two major steps: *feature representation* and *distance or metric learning*. Features used for image verification include colors, HOG, Haar-like descriptors, SIFT or SURF key point descriptors, maximally stable color regions, texture filters, differential local information, co-occurrence matrices, etc [30]. Statistics of low-level features, such as bag of words (BoW) descriptors, are also used for image verification to handle spatial variations. Recently, FVs [31] are developed which provides a better model to encode the local features. A number of methods built upon this FV approach [32], [33] have shown outstanding performance in image representation.

In this work, we propose to develop an effective cross-frame image verification method to determine if an image patch belongs to the background or not. This problem becomes very challenging since the background is highly dynamic and cluttered. In this work, we will demonstrate that a combination of DCNN features and hand-crafted image features specifically designed for camera-trap data is able to achieve significantly improved performance in animal image patch verification.

III. Algorithm Overview

We recognize that accurate and efficient animal detection from highly cluttered natural scenes in camera-trap images is a challenging task. To achieve accurate and fine-grain animal detection from the background, we need to perform image analysis at the pixel or small block level. However, with the low-contrast between the foreground animal and the cluttered background, it is often very difficult to determine if a pixel or a pixel block belongs to the animal or background based on local neighborhood information only, unless we resort to global image feature analysis. For example, pixels on the deer body might be very similar to the background vegetation. In this case, it is difficult for us to determine if these pixels belong to the deer based on local neighborhood information only until we see the deer head and legs, which involves global image analysis.

To address this issue, in this work, we propose a new animal-background detection framework which tightly couples object proposal using local image segmentation with global image region verification, as illustrated in Fig. 3. Specifically, it has two major components: 1) IEC for animal object proposal and 2) cross–frame patch-level object verification. The first component of IEC analyzes local image features and operates at the level of pixels or small blocks of pixels so as to maintain low computational complexity and achieve multi-level image segmentation in order to generate candidate regions for animal objects. To achieve high detection rate and ensure animals are all detected and covered in the foreground regions, we need to use a series of energy levels for the IEC so as to create an embedded set of regional proposals. Certainly, besides the target animal objects, the proposed regions will also contain regions or image patches from the background. The second component of image verification performs global comparison between foreground regions and background images across multiple frames. It extracts global features from the whole image patch, learns an image verification model to determine whether an image patch is similar to the background or not. In the following sections, we will explain these two components in more detail.

TABLE IV
PERFORMANCE COMPARISON ON CAMERA_TRAP DATASET

	EC Best	YOLO[45]	Fast-RCNN[10]	Faster-RCNN[11]	IEC+DCNN	Proposed
Train-set Finetune-set		voc07+voc12 camera-trap	voc07+voc12 camera-trap	voc07+voc12 camera-trap	camera-trap	camera-trap
Agouti	0.7382	0.7239	0.8088	0.8105	0.8218	**0.8364**
Collared Peccary	0.8436	0.8516	0.8838	0.8865	0.9049	**0.9202**
Paca	0.7799	0.7658	0.797	0.8055	0.7946	**0.8226**
Red Brocket Deer	0.7772	0.7905	0.8492	**0.8794**	0.8587	0.8723
White-nosed Coati	0.8221	0.8016	0.8739	0.8883	0.8893	**0.8993**
Spiny Rat	0.6908	0.7016	0.7729	0.7924	0.789	**0.8092**
Ocelot	0.7935	0.7893	0.8592	0.8796	0.8732	**0.8855**
Red Squirrel	0.7978	0.7761	0.8682	0.8901	0.8839	**0.8914**
Common Opossum	0.7395	0.7582	0.8187	0.8456	0.8263	**0.8623**
Bird spec	0.5505	0.4968	0.6083	0.6188	0.6515	**0.6717**
Great Tinamou	0.6964	0.7247	0.8282	0.8473	0.8546	**0.8699**
White-tailed Deer	0.7847	0.8165	0.8251	0.8549	0.8403	**0.8611**
Mouflon	0.7788	0.7743	0.8197	0.8395	0.8429	**0.8782**
Red Deer	0.8555	0.8642	0.8792	**0.9052**	0.898	0.9008
Roe Deer	0.8353	0.8548	0.8853	0.8968	0.8956	**0.9076**
Wile Boar	0.8013	0.8553	0.8732	0.9018	0.8922	**0.907**
Red Fox	0.676	0.6548	0.7538	0.7682	0.7765	**0.7933**
European Hare	0.6695	0.6561	0.7862	0.7892	0.7983	**0.8283**
Wood Mouse	0.7176	0.6815	0.7972	0.8136	0.8098	**0.8357**
Coiban Agouti	0.6678	0.6915	0.7982	0.8046	0.8121	**0.8221**
Average	0.7587	0.7674	0.8251	0.8493	0.8417	**0.8597**

Metrics showing average **Recalls**.

TABLE III
PERFORMANCE COMPARISON ON CAMERA_TRAP DATASET

	EC Best	YOLO[45]	Fast-RCNN[10]	Faster-RCNN[11]	IEC+DCNN	Proposed
Train-set Finetune-set		voc07+voc12 camera-trap	voc07+voc12 camera-trap	voc07+voc12 camera-trap	camera-trap	camera-trap
Agouti	0.7632	0.7593	0.742	0.7514	0.7875	**0.8244**
Collared Peccary	0.8209	0.8359	0.8015	0.8094	0.7682	**0.8152**
Paca	0.7969	0.8169	0.8039	0.8289	0.8122	**0.8333**
Red Brocket Deer	0.8563	0.8915	0.8517	**0.8879**	0.8658	0.8867
White-nosed Coati	0.8059	0.8314	0.7899	0.7952	0.803	**0.8221**
Spiny Rat	0.7539	0.7642	0.7193	0.7314	0.7604	**0.7756**
Ocelot	0.7918	0.8192	0.7726	0.7952	0.8011	**0.8154**
Red Squirrel	0.7345	0.7682	0.7328	0.7437	0.7638	**0.7727**
Common Opossum	0.7816	0.8164	0.7951	0.8155	0.8023	**0.8205**
Bird spec	0.6527	0.7465	0.6412	0.6619	0.6898	**0.7228**
Great Tinamou	0.789	0.8349	0.8035	0.8148	0.8313	**0.8441**
White-tailed Deer	0.8218	0.8432	0.8303	0.8792	0.8551	**0.8671**
Mouflon	0.7594	0.8448	0.7692	0.7846	0.7922	**0.8107**
Red Deer	0.7947	0.8214	0.7963	0.7991	0.8234	**0.8391**
Roe Deer	0.7969	0.8391	0.7793	0.7925	0.8022	**0.8218**
Wile Boar	0.7863	0.8417	0.7965	0.805	0.8131	**0.8282**
Red Fox	0.6471	0.7349	0.6752	0.6849	0.7056	**0.7358**
European Hare	0.7156	0.7514	0.7391	0.7485	0.753	**0.772**
Wood Mouse	0.7094	0.7539	0.7293	0.7336	0.7493	**0.7632**
Coiban Agouti	0.7316	0.7815	0.749	0.7598	0.7732	**0.7778**
Average	0.7824	0.8315	0.7801	0.7886	0.8017	**0.8209**

Agouti	0.7505	0.7436	0.7783	0.7825	0.8043	**0.8303**
Collared Peccary	0.8321	0.8246	0.8455	0.8546	0.831	**0.8646**
Paca	0.7883	0.7816	0.8004	0.8145	0.8035	**0.828**
Red Brocket Deer	0.8148	0.8241	0.8568	**0.8803**	0.8622	0.8795
White-nosed Coati	0.814	0.8348	0.8398	0.8415	0.8439	**0.859**
Spiny Rat	0.721	0.7282	0.7485	0.7503	0.7745	**0.7921**
Ocelot	0.7926	0.7844	0.8048	0.8117	0.8356	**0.849**
Red Squirrel	0.7648	0.7486	0.7892	0.7962	0.8194	**0.8278**
Common Opossum	0.76	0.7782	0.8071	0.8286	0.8142	**0.8409**
Bird spec	0.5973	0.5543	0.6367	0.6488	0.6701	**0.6963**
Great Tinamou	0.7398	0.7581	0.8143	0.8185	0.8428	**0.8568**

TABLE IV
PERFORMANCE COMPARISON ON CAMERA_TRAP DATASET

	EC Best	YOLO[45]	Fast-RCNN[10]	Faster-RCNN[11]	IEC+DCNN	Proposed
Train-set		voc07+voc12	voc07+voc12	voc07+voc12	camera-trap	camera-trap
Finetune-set		camera-trap	camera-trap	camera-trap		
~~White-tailed Deer~~	~~0.8028~~	~~0.8147~~	~~0.847~~	**~~0.8672~~**	~~0.8476~~	~~0.8641~~
Mouflon	0.769	0.7498	0.7962	0.8067	0.8168	**0.8431**
Red Deer	0.824	0.8345	0.8397	0.8416	0.8591	**0.8689**
Roe Deer	0.8157	0.8354	0.8254	0.8435	0.8463	**0.8626**
Wile Boar	0.7937	0.8491	0.8312	0.8477	0.8508	**0.8658**
Red Fox	0.6612	0.6814	0.7162	0.7211	0.7394	**0.7634**
European Hare	0.6918	0.6815	0.7573	0.7604	0.775	**0.7992**
Wood Mouse	0.7135	0.6981	0.7681	0.7692	0.7784	**0.7978**
Coiban Agouti	0.6982	0.7204	0.7582	0.7685	0.7921	**0.7993**
Average	0.7703	0.7515	0.7937	0.8043	0.8212	**0.8398**

Metrics showing average **F-scores**.

And we use 20 original classes objects as positive samples, marked as 1. For all verification models in experiments, we finetune a 2-way classification model, with a batch size 128, learning rate 0.01, momentum 0.9 and a decay of 0.0005. We continue training with a maximum iteration 40 000 on camera-trap, and 80 000 on Pascal VOC, respectively. When training accuracy is higher than 98% and stopped climbing for a while, we stop the training to prevent over-fitting.

C. Experimental Results

1) Qualitative Evaluations: Fig. 10 shows some example experimental results on four sequences from different species, which are *Red Fox*, *Great Tinamou*, *European Hare* and *Paca*, respectively. Row (a) is the original image from the camera-trap. Due to space limitations, we only include 3 out 10 images here. Row (b) shows the segmentation results after the video object graph cut. The animal body boundaries (shapes) are not very accurate and there a significant amount of incorrect segmentation results. After several iterations of cross-frame information fusion and graph cuts by utilizing existing background information, better results are achieved in row (c). We can see that false positive patches caused by background variations, such as shadows, waving leaves, and moving clouds, are still in the segmentation results. Row (d) shows the final results after animal-background verification. These false positive patches have been successfully removed. Row (e) shows the animal pixels with row (d) as the mask. We can see that the proposed method is able to achieve very accurate and reliable segmentation of the foreground animals in dynamic scenes by preserving true positives and filtering out false positives. Fig. 11 shows some examples of animal segmentation from highly cluttered and dynamic natural scenes.

2) Quantitative Results: We first compare the performance of our animal object proposal method using IEC with the following state-of-the-art object proposal methods: Spatial-Temporal Object Detection Proposals (STODP) [44], Fully Connected Object Proposals for Video Segmentation (FCOP) [43], and Learning to Segment Moving Object in Videos (MOP) [42]. For reference purposes, we also compare the performances of single frame object proposal techniques, as proposed in Geodesic Object Proposals [41] and Selective Search [12]. The latter is very popular and is used in RCNN and fast-RCNN as the enabling proposal method. Table I provides the average number of proposal bounding box required by each method in order to cover 80%, 90% and the most ground-truth animal objects, respectively. Here, coverage is the percentage of ground-truth bounding boxes which have $IoU \geq 0.5$ with any box in detection proposal list. The best coverage rate indicates the capability of detecting all objects in every frame. Improving the coverage is hard and costly, which often results in a massive amount of proposal detections. We can see that our proposed method is much more efficient than existing methods at the coverage rates of 80% and 90%, and find a good trade-off between the number of proposals (which affects the subsequent verification time) and the coverage rate. The single-frame based methods, such as the Selective search and GOP often produce a large amount of proposals. The limitation of proposed IEC method is its weakness in detecting slow moving object, which could be neglected in motion triggered dataset such as camera_trap. In return, IEC is exceptionally good at filtering out non-candidate object proposals, which is a crucial for accurate animal detection.

Tables II, III, and IV provides quantitative recall, precision and F-score comparisons on our Camera_trap dataset, respectively. We compare our proposed method.

TABLE V
AVERAGE PROCESSING TIME PER IMAGE IN SECONDS WITH VARIOUS EXPERIMENTAL CHOICES

	Experimental choices					
Use Selective Search						
Use Iterative Graph-Cut						
Run CNN feature extractor on GPU						
Run CNN feature extractor on CPU						
Run CNN using large batch						
Proposal generation	0.75	0.75	0.75	1.03	1.03	1.03
Verification	8.94	3.68	2.92	4.19	1.01	0.54
Total	9.69	4.43	3.67	5.22	2.04	1.57

[VII. CONCLUSION

In this paper, we have successfully developed an accurate method for animal object detection from highly cluttered natural scenes captured by motion-triggered cameras, called camera-traps. We developed a new approach to generate animal object region proposals using multi-level graph cut in the spatiotemporal domain. We then developed a cross-frame temporal patch verification method to determine if these region proposals are true animals or background patches. We found that the DCNN and FV-HOG features are able to enhance the performance of each other during animal object verification. Our extensive experimental results and performance comparisons over a diverse set of challenging camera-trap data demonstrated that the proposed spatiotemporal object proposal and patch verification framework is sensitive to objects in motion and confident in rejecting false alarms, thus is capable of building the basis of a robust object detection system in dynamic scenes.

REFERENCES

[1] S. Tilak *et al.*, "Monitoring wild animal communities with arrays of motion sensitive camera," *Int. J. Res. Rev. Wireless Sensor Netw.*, vol. 1, pp. 19–29, 2011.

[2] R. Kays *et al.*, "eMammal—Citizen science camera trapping as a solution for broad-scale, long-term monitoring of wildlife populations," in *Proc. North Am. Conservation Biol.*, 2014, pp. 80–86.

[3] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1778–1792, Nov. 2005.

[4] T. Ko, S. Soatto, and D. Estrin, "Background subtraction on distributions," *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 276–289.

[5] Y. Ren, C.-S. Chua, and Y.-K. Ho, "Motion detection with nonstationary background," *Mach. Vis. Appl.*, vol. 13, no. 5, pp. 332–343, Mar. 2003.

[6] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, pp. 309–314, 2004.

[7] Y. Boykov and V. Kolmogorov, "An experimental comparison of mincut max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.

[8] V. Mahadevan and N. Vasconcelos, "Background subtraction in highly dynamic scenes," *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–6.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 580–587.

[10] R. Girshick, "Fast r-CNN," in *Proc. Int. Conf. Comput. Vis.*, pp. 1440–1448, 2015.

Number of Crosss-frames used in Validation