Enhancing Anomaly Detection in Surveillance Videos with Transfer Learning from Action Recognition

Kun Liu[1], Minzhi Zhu[1], Huiyuan Fu[1], Huadong Ma[1], Tat-Seng Chua[2]
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]National University of Singapore, Singapore
{liu_kun,zmzbupt,fhy,mhd}@bupt.edu.cn,chuats@comp.nus.edu.sg

ABSTRACT

Anomaly detection in surveillance videos, as a special case of video-based action recognition, has been of increasing interest in multimedia community and public security. Action recognition in videos faces some challenges, such as cluttered background, illumination conditions. Besides these above difficulties, detecting anomaly in surveillance videos has several unique problems to be solved. For example, the lack of sufficient training samples is one of the main challenges for detecting anomalies in surveillance videos. In this paper, we propose to utilize transfer learning to leverage the good results from action recognition for anomaly detection in surveillance videos. More specially, we explore some techniques based on action recognition models from the following aspects: training samples, temporal modules for action recognition, network backbones. We draw some conclusions. First, more training samples from surveillance videos lead to higher classification accuracy. Second, stronger temporal modules designed for recognizing action and deeper networks do not achieve better results. This conclusion is reasonable since deeper networks tend to over-fitting, especially for the small-scale training set. Besides, to distinguish the hard examples from normal activities, we separately train a neural network to classify the hard category and normal events. Then we fuse the binary network and previous network to generate the final prediction for general anomaly detection. On the benchmarks of CitySCENE, our framework achieves promising performance and obtains the first prize for general anomaly detection and the second prize for specific anomaly detection.

KEYWORDS

Anomaly Detection, Transfer Learning, Action Recognition, Surveillance Videos

ACM Reference Format:
Kun Liu[1], Minzhi Zhu[1], Huiyuan Fu[1], Huadong Ma[1], Tat-Seng Chua[2]. 2020. Enhancing Anomaly Detection in Surveillance Videos with Transfer Learning from Action Recognition. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3394171.3416298

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3416298

1 INTRODUCTION

Action recognition has drawn considerable attention from the multimedia research community owing to its widespread applications, such as video classification [23, 27]. Recently, deep neural networks and large-scale datasets bring impressive performance on action recognition. For example, action recognition accuracy on UCF101 [24] reaches 98.0% with the I3D [3] pre-trained on Kinetics [11] dataset that comprise several hundred thousand videos. However, these action recognition models can not directly detect anomaly in surveillance videos because they are always trained with videos downloaded from the website.

Despite great practical importance, anomaly detection in surveillance videos is still lacking behind. This task is inherently more difficult since besides challenges action recognition faced, also problems brought by surveillance videos have to be tackled.

First, in contrast to action recognition, anomalies have no clear definition to distinguish from normal events. For example, the crowd category belongs to the anomaly in the CitySCENE while this class is defined as the normal activity in the UCFCrime [26] dataset.

Second, and even more critically, building large-scale datasets to train anomaly detection models is difficult. Different from action recognition benchmarks which are constructed by unrestrainedly downloading video from the internet, it is hard to build large-scale anomaly detection datasets since we have no access to collect enough surveillance videos through public methods. Moreover, collecting the videos which occur dangerous events or illegal activities is more difficult due to legal restrictions and privacy protection.

To demonstrate the difference between action recognition and anomaly detection datasets, we list the total video numbers, class numbers, and the average samples per category in Table 1. We can observe that action recognition benchmarks have hundreds of times more training samples than anomaly detection datasets. More importantly, the state-of-the-art approaches can obtain impressive accuracy on action recognition benchmarks. Inspired by these observations, we explore to transfer the knowledge of action recognition models to detect anomaly in surveillance videos.

In this work, we first fine-tune two types of action recognition models: 2D-CNN based methods and 3D-CNN based approaches. We observe that 2D-CNNs can achieve higher anomaly detection accuracy perhaps due to fewer parameters. Then, we explore several 2D-CNN backbones with different pre-trained datasets. Next, we progressively extend the training set by collecting some anomalies from existing datasets. Besides, to improve the poor performance in some categories, we separately train a network to classify the hard category and normal events. Finally, we fuse both networks to generate the prediction.

Table 1: Statistics of four action recognition benchmarks and two anomaly detection datasets.

Datasets	Datasets	Year	Surveillance Videos	Trimmed	Videos	Classes	Samples Per Class
Action Recognition	UCF101 [24]	2012	×	✓	13,320	101	min 101
	ActivityNet [2]	2015	×	×	28,108	200	avg 141
	Kinetics-400 [11]	2017	×	✓	306,245	400	min 400
	YouTube8M [1]	2016	×	×	8,000,000	4,800	avg 1,667
Anomaly Detection	UCFCrime [25]	2018	✓	×	1,900	14	min 50
	CitySCENE	2020	✓	✓	2,265	13	min 30

The rest of the paper is organized as follows. The related work is briefly introduced in Section 2. We report the experimental results in Section 3. Finally, Section 4 points out the direction of our work, and Section 5 concludes this work.

2 RELATED WORK

We briefly divide previous related work into two categories: (i) action recognition, (ii) anomaly detection.

Action Recognition. As one of the most related tasks to anomaly detection, action recognition offers many options to detect the anomaly. According to different network architectures, the approaches designed for action recognition fall into two categories: one-stream 3D-CNN based methods [3, 5, 14, 16, 27] and two-stream 2D-CNN architecture approaches [13, 23, 28]. The former methods utilize 3D convolutional filters and operate on a short clip. The basic architecture of two-stream is first proposed in [23] where spatial net captures single RGB image appearance and temporal net depicts the motion among a short clip with the input of ten optical flow maps.

Recently, some researchers [10, 13, 35] focus on efficient action recognition which can process the videos in real-time. Lin et al. [13] develop a novel Temporal Shift Module (TSM) which shifts part of the channels along the temporal dimension. This module is efficient since it can achieve the performance of 3D CNN with 2D CNN's complexity. Similarly, a Spatio-Temporal and Motion (STM) encoding block is designed in [10] to efficiently encode motion features in a 2D framework.

Anomaly Detection. Anomaly detection has become an increasingly popular research topic due to its extensive applications. In the last decade, a lot of research has been done on detecting violence, aggression, or accident in a specific scenario.

The authors of [20, 21, 33] introduce sparse representation to learn the dictionary of abnormal behaviors. In [6], a sort of oriented flow descriptor is designed to classify anomaly and normal events in surveillance videos. Similarly, the motion and orientation of human limbs are developed in [4] to detect abnormal behaviors in crowd scenarios. Besides, in order to distinguish violence from normal actions, a set of simple behavioral heuristics is defined in [21] to describe people's behaviors.

Inspired by the success of deep learning for video understanding [8, 18, 19, 29], researchers make extensive attempts to recognize abnormal events using deep learning in a specific scenario. More specifically, Zhao et al. [34] introduce a novel prediction loss for producing future frames to enhance the anomaly feature learning. In [22], generative adversarial nets (GANs) are trained with normal frames and corresponding optical flow images to detect anomaly. The meta-learning mechanism is introduced in [17] to eliminate background-bias in anomaly detection.

Some researchers pay attention to building datasets for anomaly detection in surveillance videos. Sultani et al. [25] construct a large-scale anomaly detection dataset and propose a novel deep multiple instance ranking framework for the weakly-labeled training set. Similarly, the authors of [15] collect trimmed surveillance videos for anomaly detection. Both benchmarks are utilized to train our model, which increases the training samples significantly and then improves the detection accuracy.

3 EXPERIMENTS

We conduct experiments on the challenging benchmark: CitySCENE. In this section, we first describe the dataset, evaluation metrics, and implementation details. Then, we present and analyze the experiment results of action recognition models. More specially, we explore the effect of different action recognition models, network backbones, and training samples. Next, we introduce several extra benchmarks that comprise hard-detected anomaly categories to mine the hard example.

3.1 Dataset

CitySCENE. CitySCENE is a new large-scale anomaly dataset consisting of 2,265 surveillance videos with 12 realistic anomalies, such as robbery, explosion, shooting, stealing, and so on. The duration of most training videos is several seconds since they are trimmed and only contain the anomaly. This benchmark can promote the research of city management and public safety.

3.2 Evaluation Metrics

To evaluate the models, we split 30% of the original training set as the validation set. We report the frame classification accuracy for general anomaly detection and video classification accuracy for specific anomaly detection.

Table 2: The accuracy of 2D-CNNs and 3D-CNNs on action recognition datasets and anomaly detection datasets.

Methods	Methods	Kinetics-400	CitySCENE
2D-CNN	TSN [28]	70.6	**90.8**
	TSM [13]	74.1	87.9
3D-CNN	SlowFast [5]	77.0	89.0
	TPN [30]	**77.7**	87.4

Table 3: The video accuracy and frame accuracy on CitySCENE using different network backbones.

Backbones	Pre-trained Dataset	Video Acc	Frame Acc
ResNet50 [7]	Kinetics	80.32	87.86
ResNet50 [7]	ImageNet	80.31	90.86
ResNet101 [7]	ImageNet	81.38	88.46
ResNeSt50 [31]	ImageNet	82.98	**91.14**
ResNeSt101 [31]	ImageNet	**83.51**	91.10

3.3 Experiment Results

First of all, we train some efficient action recognition models on the CitySCENE. Then, we fine-tune the action recognition models using different CNN backbones, different training datasets. Next, we collect more training data from the existing benchmarks. More specifically, a network is trained to classify the hard category and normal events.

The effect of different action recognition models. We adopt some strong action recognition baselines, including 2D-CNNs: TSN [28], TSM [13] and 3D-CNNs: SlowFast [5] and TPN [30]. We do not adopt two-stream architecture [23] because the heavy computational cost of optical flow can not meet the time requirement of this challenge.

On the task of action recognition, 3D-CNNs always achieve higher accuracy than 2D-CNNs. According to Table 2, we can see that the accuracy of SlowFast model [5] is 6.4% better than TSN [28] on Kinetics-400 [11]. However, 3D-CNNs present lower performance than 2D-CNNs on the task of anomaly detection in surveillance videos. Among 2D-CNNs, TSM [13] always obtain higher accuracy than TSN [28] on the action recognition, but degrade the performance on the anomaly detection. Insufficient training samples may be the main reason for this interesting phenomenon. In this case, deep networks with more parameters may be more likely to result in over-fitting.

The effect of different network backbones. On action recognition, deeper networks always achieve better performance. Based on TSN [28], we conduct experiments on CitySCENE with four ImageNet pre-trained networks: ResNet50 [7], ResNet101 [7], ResNeSt50 [31], ResNeSt101 [31] and a Kinetics pre-trained ResNet50 [7].

From Table 3, We have another two observation that is inconsistent with action recognition. First, Kinetics pre-trained model presents lower performance than ImageNet pre-trained models on anomaly detection. Second, deeper networks do not obtain higher accuracy on anomaly detection perhaps due to too many parameters. More importantly, we see that ResNeSt [31] achieve better performance than ResNet [7], perhaps because the split-attention module in ResNeSt has good generalization ability.

The effect of more training samples. In this section, we first analyze the accuracy of each category on the CitySCENE validation set. According to Table 4, the TSN model with ResNeSt101 demonstrates excellent accuracy in both explosion and graffiti categories. The unique visual information of both classes may lead to this success. In contrast, the performance of shooting, crowd, accident, and smoking is extremely poor. Insufficient training samples may be the main reason for this failure. Thus, we adopt several existing benchmarks to extend the training set.

UCFCrime [25] is a large-scale anomaly detection benchmark, which contains 1,900 untrimmed surveillance videos with 13 real-world anomalies. Recently, Liu et al. [17] make this dataset trimmed by annotating the start and end time of anomaly in the long surveillance videos. More importantly, UCFCrime [25] and CitySCENE have some same anomalies, including accident, stealing, robbery, shooting, fighting, and explosion. Thus, we extend the training set with 558 trimmed anomaly samples and 800 normal videos.

Kinetics-400 [11] is one of the largest trimmed video datasets for action recognition, which contains 306,245 videos with 400 action classes. To improve the accuracy on the smoking category, we random select 882 smoking videos from Kinetics-400 to further extend the training set. Interestingly, adding smoking samples does not bring performance improvement perhaps because the videos from Kinetics-400 [11] are collected from the web rather than surveillance cameras.

To improve the accuracy on crowd category, we choose two benchmarks to augment the training set: ShanghaiTech [32] and UCF-QNRF [9]. ShanghaiTech [32] is established for evaluating the crowd count task. This dataset is composed of two parts. Part A contains 482 images that are downloaded from the websites. Part B includes 716 images which are taken from the busy streets of metropolitan areas in Shanghai. We randomly select 30 images from Part A to generate a training sample. We choose 15 images from Part B as a training sample to construct more training samples since the images of Part B are from surveillance videos. Thus, we extend the training set with 62 crowd samples from ShanghaiTech [32].

Similarly, UCF-QNRF [9] is constructed for evaluation of human density map estimation and localization in dense crowds. This benchmark contains 1,535 images with high resolution. The authors of [9] collect this dataset from the web by selecting images captured in realistic scenarios. We randomly select 30 images from UCF-QNRF to generate a training sample. Thus, we extend the training set with 51 crowd samples from UCF-QNRF [32].

Sur5H [15] is a new benchmark for anomaly detection in surveillance videos. More importantly, Sur5H and CitySCENE have four same categories: robbery, fighting, accidents, and stealing. Therefore, we select these videos from Sur5H to increase training videos.

According to Table 5, the surveillance video datasets (e.g., Sur5H and UCFCrime) always bring performance improvement. Introducing website samples even cause the degradation of video classification accuracy. Thus, how to make full use of existing large-scale

Table 4: The category accuracy on CitySCENE.

Category	Shoot	Crowd	Carry	Accident	Smoke	Walkingdog	Sweep	Robbery	Fight	Steal	Normal	Graffiti	Explosion
Accuracy	25.0	28.6	42.9	57.1	62.5	66.7	66.7	68.8	86.9	90.0	97.3	97.5	100.0

Table 5: The video accuracy and frame accuracy on CitySCENE using different training sets.

Training Set	Video Acc	Frame Acc
CitySCENE	83.51	91.10
CitySCENE + UCFCrime [25]	84.04	91.51
CitySCENE + UCFCrime [25] + Kinetics-400 [11]	81.91	91.73
CitySCENE + UCFCrime [25] + ShanghaiTech [32] + UCF-QNRF [9]	**85.11**	91.73
CitySCENE + UCFCrime [25] + ShanghaiTech [32] + UCF-QNRF [9] + Sur5H [15]	84.04	**94.04**

Table 6: Performance of different methods on CitySCENE testing set in the general anomaly detection task.

Rank	Team Name	AUC
1	BigFish(Ours)	**89.2**
2	MonIIT	87.94
3	DeepBlueAI	86.85
4	SYSU-BAIDU	86.52
5	UHV	85.37

Table 7: Performance of different methods on CitySCENE testing set in the specific anomaly detection task.

Rank	Team Name	MF1
1	SYSU-BAIDU	**66.41**
2	BigFish(Ours)	62.11
3	GOGOGO	52.33
4	MonIIT	45.52
5	Orange-Control	40.42

website videos is an important problem for anomaly detection in surveillance videos.

The effect of mining hard category. After training on multiple extra benchmarks, we analyze the accuracy of each category. We observe that the accuracy of fighting and accident increase. For general anomaly detection, crowd videos occupy most of the failure cases that are incorrectly classified as normal events. Thus, we attempt to improve the accuracy of the hard category. A training set which only contains normal events and crowd is established for training binary networks. Next, we fuse the score of this binary network and the previous network as the final prediction. On the validation set, combining two networks achieve 96.43% on the frame accuracy from 94.04%. We report our performance on testing set in Table 6 and Table 7. We can see that our method achieves the first prize on general anomaly detection and the second prize on specific anomaly detection.

4 FUTURE WORK

In this section, we point out our future work for anomaly detection. We plan to utilize more input modalities to capture the pattern of abnormal activities. We only adopt the RGB frames as input in this work. On action recognition, two-stream networks always achieve higher accuracy than networks only adopt RGB frames. We support that fusing optical flow and RGB frames can bring performance improvement on anomaly detection. More importantly, some fast optical flow algorithms with very low time complexity, such as Dense Inverse Search-based (DIS) method [12], can be introduced to detect anomaly in real-time speed.

5 CONCLUSION

In this work, we propose to utilize transfer learning to leverage the good results from action recognition for anomaly detection in surveillance videos. More specially, we explore the training samples, temporal modules for action recognition, networks backbones. Then we draw some conclusions. First, more training data from surveillance videos leads to higher classification accuracy. Second, stronger temporal modules and deeper networks do not bring performance improvement. Besides, to distinguish this hard example from normal activities, we separately train a neural network to classify the hard category and normal events. Finally, we present our future work to inspire more works on anomaly detection.

6 ACKNOWLEDGEMENT

This work is supported in part by the Natural Science Foundation of China (NSFC) (61720106007), the Innovation Research Group Project of NSFC (61921003), the 111 Project (B18008). This research is also supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative.

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.

[3] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[4] Ankur Datta, Mubarak Shah, and N Da Vitoria Lobo. 2002. Person-on-person violence detection in video data. In *IEEE International Conference Pattern Recognition*, Vol. 1. 433–438.

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *IEEE International Conference on Computer Vision*. 6202–6211.

[6] Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, and Yi Liu. 2016. Violence detection using oriented violent flows. *Image and Vision Computing* 48 (2016), 37–41.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[8] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. 2019. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 8450–8459.

[9] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. 2018. Composition loss for counting, density map estimation and localization in dense crowds. In *European Conference on Computer Vision*. 532–546.

[10] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. 2019. Stm: Spatiotemporal and motion encoding for action recognition. In *IEEE International Conference on Computer Vision*. 2000–2009.

[11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The Kinetics Human Action Video Dataset. *IEEE Conference on Computer Vision and Pattern Recognition* (2017).

[12] Till Kroeger, Radu Timofte, Dengxin Dai, and Luc Van Gool. 2016. Fast optical flow using dense inverse search. In *European Conference on Computer Vision*. 471–488.

[13] Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *IEEE International Conference on Computer Vision*. 7083–7093.

[14] Kun Liu, Wu Liu, Chuang Gan, Mingkui Tan, and Huadong Ma. 2018. T-C3D: Temporal Convolutional 3D Network for Real-time Action Recognition. In *AAAI Conference on Artificial Intelligence*. 7138–7145.

[15] Kun Liu, Wu Liu, Huadong Ma, Wenbing Huang, and Xiongxiong Dong. 2017. Generalized Zero-Shot Learning for Action Recognition with Web-Scale Video Data. *World Wide Web Internet and Web Information Systems* 22, 2 (2017), 807–824.

[16] Kun Liu, Wu Liu, Huadong Ma, Mingkui Tan, and Chuang Gan. 2020. A Real-time Action Representation with Temporal Encoding and Deep Compression. *IEEE Transactions on Circuits and Systems for Video Technology* (2020).

[17] Kun Liu and Huadong Ma. 2019. Exploring Background-bias for Anomaly Detection in Surveillance Videos. In *ACM International Conference on Multimedia*. ACM, 1490–1499.

[18] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. 2018. PROVID: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance. *IEEE Trans. Multimedia* 20, 3 (2018), 645–658.

[19] Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3566–3574.

[20] Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal event detection at 150 fps in matlab. In *IEEE International Conference on Computer Vision*. 2720–2727.

[21] Sadegh Mohammadi, Alessandro Perina, Hamed Kiani, and Vittorio Murino. 2016. Angry crowds: Detecting violent events in videos. In *European Conference on Computer Vision*. 3–18.

[22] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. 2017. Abnormal event detection in videos using generative adversarial nets. In *International Conference Image Processing*. 1577–1581.

[23] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*. 568–576.

[24] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR abs/1212.0402* (2012).

[25] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-World Anomaly Detection in Surveillance Videos. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6479–6488.

[26] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world Anomaly Detection in Surveillance Videos. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6479–6488.

[27] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision*. 4489–4497.

[28] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*. 20–36.

[29] Qi Wang, Xinchen Liu, Wu Liu, An-An Liu, Wenyin Liu, and Tao Mei. 2020. MetaSearch: Incremental Product Search via Deep Meta-Learning. *IEEE Transactions on Image Processing* 29 (2020), 7549–7564.

[30] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. 2020. Temporal pyramid network for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 591–600.

[31] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. 2020. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955* (2020).

[32] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016. Single-image crowd counting via multi-column convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*. 589–597.

[33] Bin Zhao, Li Fei-Fei, and Eric P Xing. 2011. Online detection of unusual events in videos via dynamic sparse coding. In *Computer Vision and Pattern Recognition*. 3313–3320.

[34] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. 2017. Spatio-temporal autoencoder for video anomaly detection. In *ACM Multimedia*. 1933–1941.

[35] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. 2018. Eco: Efficient convolutional network for online video understanding. In *European Conference on Computer Vision*. 695–712.